Barreto, Humberto

Book

Intermediate microeconomics with Microsoft Excel®

Reference: Barreto, Humberto (2021). Intermediate microeconomics with Microsoft Excel®. 2nd Edition 2020 version 11 November 2021. [Greencastle, IN] : [DePauw University]. https://www.depauw.edu/learn/microexcel/MicroBook/MicroExcel.pdf.

Intermediate Microeconomics with Microsoft Excel®

Humberto Barreto
Professor of Economics and Management
DePauw University
2021

2nd Edition 2020 version 11 November 2021 CC BY SA
First published in 2009 by Cambridge University Press.

This book was typeset in LaTeX with various packages in TeXstudio. I was helped repeatedly by resources at tex.stackexchange.com. I am awed by this software and its community. I offer a deep bow to those who made these tools freely available and continue to provide support.

ACKNOWLEDGMENTS

This book would not have been possible without the help of many people, but four especially stand out: Frank Howland, Kealoha Widdows, Michele Villinski, and Tami Barreto. Thank you.

I team-taught several courses with Frank and Kay so it is not surprising that their imprint, including examples, phrasing, and pedagogical strategy are embedded in this book. They caught mistakes, gave me ideas, and profoundly influenced my thinking about the best way to teach economics.

Michele sat in on my Intermediate Micro class in the Spring of 2019. She would occasionally give me tips and make suggestions to improve the presentation. I kept a running list and included them in this edition.

Tami copy edited this manuscript, like almost everything else I have written. Her attention to detail and drive for perfection have improved the exposition immensely.

To all of my DePauw and Wabash students.

Really, it has been my pleasure.

Contents

Preface

This is the second edition of a book that was originally published in 2009 by Cambridge University Press. While the core of the book remains the same, this edition refreshes all of the screenshots based on Excel 2019 and updates the data used in real-world applications. It also fixes typos and mistakes. Finally, it includes a new chapter on rational addiction and offers several new optimization problem examples.

The preface of the first edition said:

> In the competitive world of textbooks, different is definitely bad. Authors and publishers, like politicians, stay in the safe middle. Straying too far from the herd is almost a sure way to fail. Fear is strong, but it apparently can be overcome—after all, you are reading a spectacularly unconventional textbook.
>
> The most obvious difference between this book and the usual fare is the use of Microsoft Excel to teach economic theory. This enables students to acquire a great deal of sophisticated, advanced Excel skills while learning economics. No other book does this.
>
> The use of Excel drives other differences. Excel requires concrete, numerical problems instead of the abstract functions and graphs used by other books. Excel's Solver makes possible presentation of numerical methods for solving optimization problems and equilibrium models. No other book does this.
>
> Because numerical solutions are readily available, this book is able to present and explain analytical methods that have been pushed to appendixes or completely ignored in mainstream texts. Every problem is solved twice—once with Excel and once with equations, algebra, and, when needed, calculus. No other book does this.

> Finally, this book is organized differently. It explicitly repeats a single central methodology, the economic approach, so students learn how economists think and how to think like an economist. Other books try to do this, but none brings the economic way of thinking explicitly to the surface, repeating the message in every application.
>
> I wrote this book because I learned Visual Basic and quickly realized that enhancing a spreadsheet with macros made possible a whole new way of teaching economics. When my students loved this approach, I wanted to share it with others.
>
> Because this book is so different, it will probably not challenge the top sellers. It will be the unusual professor who is willing to try something this new. It requires that the professor care enough about students and teaching to invest time and energy in mastering the material. Of course, I think the rate of return is quite high. My hope is that, though few in number, a committed, enthusiastic core of adopters will enable this book to survive.
>
> Thank you for trying this unique entry into the competitive market for micro theory textbooks. I hope you find that the reward was worth the risk.

Well, after more than ten years, I can safely say that I certainly was right that the book would not challenge the top sellers! It strayed far from the herd and went largely unnoticed. When I asked Cambridge University Press to do a second edition, they politely declined.

But, I am not giving up. I believe that teaching economics via Excel is a winner. So, I am ignoring the market, producing my own second edition, and giving it away for free.

I am well aware that this edition will not attract many adopters and that I am engaged in a quixotic fight against foes who are not even aware of my presence. I remain baffled at how badly microeconomics is taught—it is as if computers were never invented. We can and must do better. I will keep this book alive in case someone wants to try a novel, innovative approach to teaching and learning microeconomics.

This edition assumes that many will read it electronically, although you are free to print it out and I am so old school that I certainly would prefer handwriting notes and underlining on paper. Any print shop can do this and, if anyone asks, explain that this is an open access book and you have legal right to print it. You can also print it online at sites such as www.lulu.com/.

I think Adobe Acrobat Reader is a good choice if you decide to read it on screen, but you are, of course, welcome to use your favorite eReader. Here is a list of 15 pdf readers: blog.hubspot.com/marketing/best-free-pdf-reader. One advantage of digital access is that links are highlighted for easy clicking. You should use your pdf reader's commenting capabilities to highlight, search (ctrl-f), and take notes. It should also be easy to look up words you do not know or search for ideas that pique your interest so take full advantage of the electronic tools at your disposal.

I have been teaching economics for a long time now. I am positive that using Excel to learn how economists use models and see the world works for almost all students. You can learn a lot of economics, math, and Excel while working with this book. Do your best and good luck!

Humberto Barreto
hbarreto@depauw.edu
Greencastle, Indiana
November 11, 2021

The idea for the electronic spreadsheet came to me while I was a student at the Harvard Business School, working on my MBA degree, in the spring of 1978.

Dan Bricklin

User Guide

This book is essentially a manual for how to actively work with and manipulate the material in Excel. This user guide lists minimum requirements, provides instructions for downloading all of the materials and software, offers a few tips before you begin, and describes the organization of the files.

Minimum Requirements

; This book presumes that you have access to and a basic knowledge of Excel. In other words, you can open an Excel file (called a *workbook*), write a formula that adds cells together, make a chart, and save the file. As you will see, however, Excel is much more than a simple adding machine. You will learn how to use Excel in a more advanced way. In addition to analyzing data and learning many new Excel functions, you will solve optimization problems with an *add-in* (a special file that extends the functionality of Excel) called Solver.

The materials in this book will work on any Windows Excel version all the way back to 1997 (version 8). The screenshots are based on Excel 2019, but if you are using an earlier version, it should be easy to figure out what to do.

The workbooks and add-ins are optimized for use with Windows Excel. They can be accessed with a Macintosh computer, but Solver in Mac Excel is temperamental and buggy. Furthermore, Visual Basic (Excel's macro language) on a Mac is limited so not all macros work. The best solution for Mac users is to emulate Windows with software such as Parallels or Boot Camp. For students at an educational institution, accessing Excel from a server (see, for example, VMWare's Horizon software) is an easy solution for Mac users. Desktops.depauw.edu gives my students access to a Windows machine running Excel configured with necessary add-ins.

To ensure that older versions of Excel can open the files, all workbooks have been saved in "compatibility mode" (Excel 97 – 2003 Workbook) with the

.xls filename extension. If you are using Excel 2007 (version 12) or greater, you should save your completed files in the "Excel macro-enabled workbook" format, which carries the .xlsm extension. Do not save your files as an Excel workbook with the .xlsx extension, the macros will not be saved and functionality will be lost.

For non-English versions of Excel, the files will work in the sense that buttons, scroll bars, and macros will function; however, the add-ins and other content will not be translated.

Recently, Microsoft Office has moved online, offering OneDrive and Office 365 cloud access. Regrettably, as of this writing, because of security concerns, online versions of Office do not support Visual Basic, a limitation which renders these options useless for working with macro-enhanced files from within a web browser. You can save a file with macros in your favorite storage area in the cloud, but you will need to download and open it with a desktop Excel version to run the macros. Within a browser, macros cannot be executed.

Downloading and Opening Workbooks

Visit www.depauw.edu/learn/microexcel to download the files that accompany this book. You may download individual files as needed or a compressed archive with all of the files to as many different computers or devices as needed.

Figure 1 shows that, when opening a workbook with macros, Excel will alert you to their presence with a security warning under the Ribbon (and right above the formula bar).

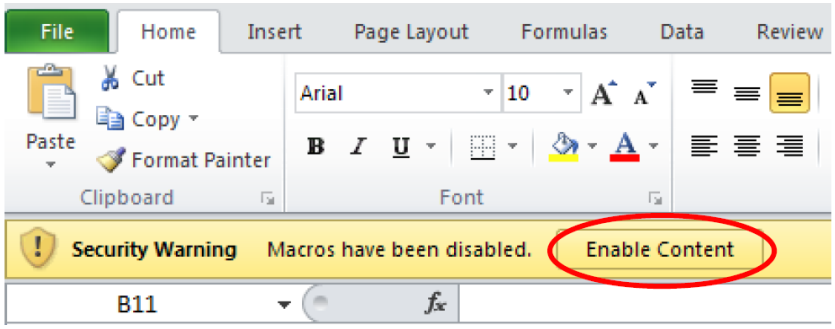


Figure 1: *Enable Content* when opening a Micro Excel workbook.

If you do not see the security warning or have no opportunity to enable content, your security level has been set to block all files with macros. Although malicious code can be harmful, you must dial down the safety measures to allow Excel to utilize fully the information in the workbook. Close the file and change the security setting to allow Excel to open files with macros.

Visit Excel's main support page at support.office.com for more help on setting security and enabling macros.

Tips and Conventions

In this book, a *figure* refers to a variety of graphics, including charts and pictures of portions of a sheet (also known as a screenshot, like Figure 1). A chart or range of cells is often displayed in this printed book as a figure, but you should look at the live version on your computer screen. Thus, in addition to a caption, many figures have a source line indicating their location in the Excel workbook.

The book follows Excel's naming convention for workbooks, sheets, and cells: [workbookname]sheetname!cell address. If the caption of a figure says, [FoodStamp.xls]BudgetConstraint, then you know the figure can be found in the *FoodStamp.xls* workbook in the *BudgetConstraint* sheet. Note that workbook and sheet names in the printed text are italicized to help you locate the proper sheet in a workbook. [RiskReturn.xls]OptimalChoice!B6 refers to cell B6 in the *OptimalChoice* sheet of the *RiskReturn.xls* workbook.

You may need to adjust your display or the objects in Excel. Use the Zoom button to magnify the display. You can also right-click objects such as buttons or scroll bars to select and move them. Once you open a workbook, you can save it to another location or name (by executing File → Save As...) and make whatever changes you wish. This is the same as underlining or writing in a conventional, printed book.

Finally, if something is not working the way you expect, there are many possible causes. It is always a good idea to close Excel completely and reopen it. Even if this does not fix the problem, slowly repeating the steps will help you debug or describe what is happening.

Organization of Files

Figure 2 shows the contents of all materials included in the MicroExcel.zip archive, after downloading it from www.depauw.edu/learn/microexcel.

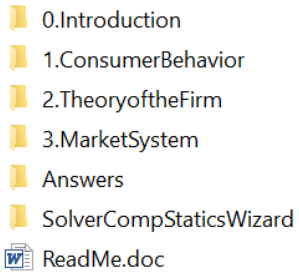


Figure 2: Organization of files.

The Answers folder contains answers to questions posed in Q&A sheets in each Excel workbook. Think of the Q&A material in the Excel workbooks as self-study questions.

There are also Exercises at the end of each chapter. Readers do not have easy access to the answers to the exercise questions. To see these answers, you must be an instructor and register online at www.depauw.edu/learn/microexcel.

The SolverCompStaticsWizard folder contains files that use the Comparative Statics Wizard Excel add-in. When used in conjunction with Excel's own Solver add-in, these files enable numerical comparative statics analysis of optimization problems and equilibrium models.

Active Learning

The most important thing you can do as you read this book is **experiment**. You might find yourself wondering, "What would happen if this cell was 10 instead of 1?" Do not just wonder, change the cell and see what happens! There is deep neuroscience at work here. When you are in control and making up your own questions, you learn best. The beauty of this approach is that everything is alive and you can make points move and lines shift. Take full advantage.

Remember that you can always download the original workbook again if needed. This means you should not worry about changing anything in a

workbook. If something goes terribly wrong, simply delete it and download it again.

There are many books devoted to microeconomics. This one is different because it is not meant to be simply read. A great deal of its value lies in the Excel workbooks and additional materials. By reading this book and working in Excel simultaneously, you will become a sophisticated user of Excel and learn a great deal of mathematics and, most importantly, economics.

Download the files from www.depauw.edu/learn/microexcel and get to work!

Spreadsheet History and Resources

For more on the history of the electronic spreadsheet, as told by one of the creators, see bricklin.com/visicalc.htm. This is the source for the epigraph.

I recommend these websites for Excel tips and tricks, workbook and add-in downloads, and Visual Basic code snippets:

- Tushar Mehta: www.tushar-mehta.com/excel/
- Chip Pearson: www.cpearson.com/excel
- Jon Peltier: peltiertech.com/Excel/
- Andy Pope: www.andypope.info

> Economics is the science which studies human behavior as a relationship between given ends and scarce means which have alternative uses.
>
> Lionel Robbins

A First Step

Economists see the world through a special pair of glasses. It takes practice and concentration to learn how to see things like an economist. The interpretation of reality that is the hallmark of modern economics has been called the economic way of thinking, the economic approach, and the method of economics. Thinking and seeing the world like an economist is the ultimate goal of this book.

You will learn the economic way of thinking by working through many examples. Here is the first one.

Optimal Allocation of Worker Hours

Suppose that you manage a tech support service for a major software company. You have two types of callers: *Regular* and *Preferred*. Your preferred customers have paid extra money for faster access, which means they expect to spend less time waiting on hold. There are equal numbers of the two types of customers and they call with equal frequency.

Management has given you a fixed number of worker hours per day to answer calls from users needing help. Daily, you have 10 workers, each working 8-hour shifts, and 5 part-time workers (4-hour shifts each); or 100 hours per day in total to support customers calling for help. These 100 hours comprise your *Total Resources*.

When customers call, an automatic message is played asking the caller to input an ID number and the caller is put on hold. The ID number is used to identify the caller as a regular or preferred customer.

Keeping callers on hold creates frustrated, unhappy customers. The callers are already angry since something has gone wrong with the software and

they need help. The faster you get support to the caller the better. You keep track of *time waiting* (the amount of time, in seconds, that the typical caller is on hold) and you know that it depends on the number of worker hours available to answer the calls.

To keep things simple, assume typical time waiting = 6000/worker hours allocated. So, say there are 80 worker hours available to answer preferred callers. Dividing 6000 by 80 yields 75, which means the typical hold time is 75 seconds. This leaves 20 worker hours for regular callers, so their hold time is 300 seconds (since 6000/20 = 300). Five minutes is a long time to wait on the phone!

The problem becomes an *economic problem* because you have two types of callers, so you must decide how to allocate your worker hours. When you have to make a decision where you trade-off one thing for another you are doing economics. In this case, the more hours you allocate to one type of caller, the lower that caller's wait time. That's the good news.

The bad news is that the fixed amount of caller-support hours means that more time devoted to one type of caller results, by definition, in fewer hours to the other type and, therefore, higher waiting times for the other type.

So the general structure of the problem is clear: You must decide how to allocate scarce support resources (worker hours) to two competing ends. Figure 3 shows a simplified picture of the problem.

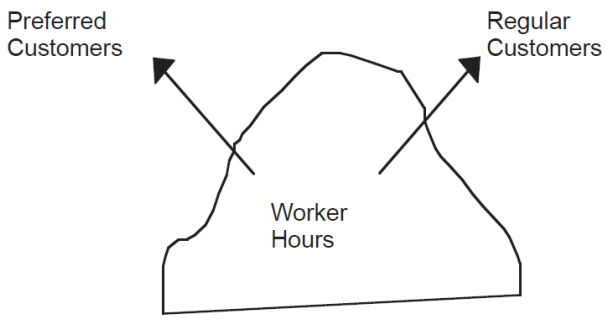


Figure 3: Allocating a scarce resource to two competing ends.

A Complication

It is unclear exactly what preferred customers expect. Do they expect to get help twice as fast or 10 times as fast as regular customers?

To incorporate the fact that the preferred customer merits greater attention, management gives you a *value weight* parameter. The value weight tells you how much more valuable the preferred caller is compared to the regular caller.

We can write the *objective function* as

$$TotalTimeWaiting = \frac{6000}{RegHours} + ValueWeight\frac{6000}{PrefHours}$$

The objective function says that time spent waiting by a preferred caller is multiplied by a factor that reflects how much more we value the preferred customer's time. If $ValueWeight = 1$, then preferred and regular callers are equally valuable. Management has decreed that preferred customers' time is worth twice that of regular customers so $ValueWeight = 2$; you (the call center manager) cannot change this parameter.

So, if you decide to allocate 50 hours each to the regular and preferred customers, then both types of customers will wait 6000/50 = 120 seconds and our objective function will be 120 + 2 x 120 = 360 seconds.

Is there a better allocation, one that yields a smaller total time waiting (adjusted with the value weight), than 50/50? This question, how to allocate 100 worker hours to answering calls from regular and preferred customers in order to minimize value weighted total time waiting, has an answer, called the optimal solution. We have to find it.

Setting Up the Problem

We will solve this problem by first organizing the information into three separate parts. All optimization problems can be set up the same way, with three components: *goal*, *endogenous variables*, and *exogenous variables*.

The goal is synonymous with the objective function. Endogenous, or choice, variables can be controlled by the decision maker. Exogenous variables are

given, fixed constants that cannot be changed by the decision maker. The exogenous variables (sometimes called parameters or independent variables) form the environment under which the decision maker acts.

In the tech support time minimization problem, we can organize the information like this:

1. Goal: minimize total time waiting (value weighted)
2. Endogenous variables: worker hours allocated to preferred and regular customers
3. Exogenous variables: total worker hours and value weight

STEP Open the Excel workbook *Introduction.xls*, read the *Intro* sheet, and then go to the *SetUp* sheet to implement the problem in Excel.

This workbook (along with all of the files that accompany this book) is available for download at www.depauw.edu/learn/microexcel. The User Guide has detailed instructions on how to properly configure Excel before downloading and opening these files.

Make sure that you enable macros when you open the file. If the buttons do not work, the most likely suspect is in the security settings.

STEP Answer the three questions in column A (below the exogenous variables). Check yourself by clicking the See Answer buttons.

Finding the Initial Solution

Now that we have set up the problem, we can turn our attention to finding the answer, the optimal solution. There are two ways to solve optimization problems:

- Analytical (algebra and calculus) methods
- Numerical (computer) methods

The analytical method uses pencil and paper to write down equations and manipulate them to find the answer. It was the only way available until computers came along and gave us algorithms for finding solutions. Numerical

methods rely on testing many trial solutions very quickly and repetitively, converging to the answer. We will ignore the analytical approach in this example and concentrate on showing how Excel's Solver works.

STEP Click the Data tab (in the Ribbon across the top of the screen), then Solver (in the Analysis group) to bring up the Solver dialog box (as in Figure 4). If Solver is not available, then use the Add-in Manager to install it. Use Excel's Help if you are having trouble or visit support.office.com.

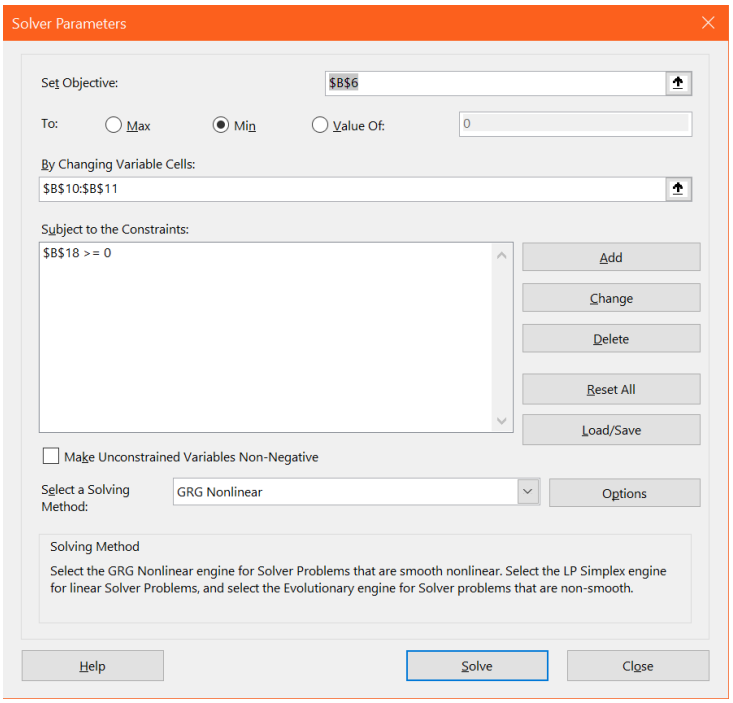


Figure 4: The Solver dialog box.

Note that necessary information is already entered. The objective cell is the (value weighted) total time waiting, the changing variable cells (the endogenous variables) are the worker hours devoted to the regular and preferred customers, and the constraint is that the sum of the worker hours not exceed the 100 hours you have been given.

STEP Click the Solve button to find the solution to the problem. Click the OK button in the Solver Results dialog box to accept Solver's solution and put the optimal solution in the *SetUp* sheet.

Congratulations! You, the call center manager, have just used Solver (a numerical methods approach to optimization) to optimally allocate your scarce resources. We can check Solver's answer for plausibility, noting that it makes sense that preferred callers have more hours allocated to them because they are more valuable. Later, we will see that we can solve this problem using analytical methods and if the two approaches give the same answer, we can be confident that we do indeed have the best solution.

Comparative Statics

We have found the initial solution, but we are usually much more interested in a follow up question: How will the optimal solution change if the environment changes?

Comparative statics is a shorthand way of describing the following procedure: Change an exogenous variable, holding the other parameters constant, and track how the optimal solution changes in response to the shock.

Like finding the initial solution, comparative statics can be done via analytical (algebra and calculus) and numerical (computer) methods. The Comparative Statics Wizard (CSWiz) add-in was used to explore how the optimal allocation of total worker hours would change if worker hours were increased by 10 hours. The CSWiz add-in will be introduced later and you will learn how to do your own comparative statics analyses. For now, we will focus on what it produces.

STEP See the results of the comparative statics analysis by going to the *CS1* sheet.

Cells A1:D15 in the *CS1* sheet were produced by the CSWiz add-in. It is easy to see that increased total worker hours are allocated to regular and preferred customers in a stable pattern. Every additional hour of total resources, holding value weight (the only other exogenous variable in this simple problem) constant, produces an increase of 0.586 hours allocated to preferred customers. The chart below the data (row 16) shows the linear relationship. Usually, economists want to determine the relationship between optimal endogenous and exogenous variables.

Summary: Introducing Optimization

This chapter used an example to show how Excel's Solver can find the optimal solution. It introduced the basics of optimization, including the three parts of every optimization problem:

1. Goal (or objective function),
2. Endogenous variables,
3. Exogenous variables.

As you work with this book, you will learn how to use analytical methods to solve optimization problems. You will also learn how to do comparative statics analysis via analytical and numerical methods.

This introductory example was completely prepared for you. All you had to do was click a few buttons. Future problems will gradually relax the Excel environment, giving you ever more freedom to make decisions and thereby learn what to do. The ultimate goal is for you to be able to set up and solve problems yourself.

Exercises

1. Suppose Management decides that preferred customers are three times as important as regular customers, so that the $ValueWeight = 3$. With 100 workers hours, what is the optimal solution? Describe your procedure and report the optimal values of *PrefHours* and *RegHours*.
2. Compared to the initial solution, when $ValueWeight = 2$, what is the change in the number of hours allocated to preferred customers?
3. The percentage change in *ValueWeight* is 50% (from 2 to 3). What is the percentage change in the number of hours allocated to preferred customers?

References

Each section ends with references and resources for further study. A citation for the epigraph (lead quotation) of the chapter is provided. References may also contain citations documenting sources used, additional information on the history of a concept or person, and suggestions for further reading.

The epigraph to this chapter is found on page 16 of the second edition of *An Essay on the Nature and Significance of Economic Science* by Lionel Robbins. This book was originally published in 1932 and the second edition is available online at www.mises.org/books/robbinsessay2.pdf. Robbins rejects old definitions of economics based on content (the study of business and work) and argues for a definition of economics based on methods used: optimization and comparative statics. Robbins made the definition of economics (in the epigraph to this chapter) famous, but he includes a footnote that cites various precursors who used a similar description of economics.

For more on Robbins, visit www.econlib.org/library/Enc/bios/Robbins.html. Econlib says that Robbins' Essay is "one of the best-written prose pieces in economics."

Nobel laureate Gary Becker's *The Economic Approach to Human Behavior* (first published in 1976) has a classic introductory chapter on the meaning of the economic approach and applies economic analysis to such non-standard topics as discrimination, crime, and marriage. Becker's statement, "what most distinguishes economics as a discipline from other disciplines in the social sciences is not its subject matter but its approach" (p. 5), greatly extends the scope of economics.

Modern economics pays little attention to its own history and how we got to be where we are today. The epigraphs in this book highlight important contributions and individuals (like Robbins and Becker) in the development of modern economic theory. Remember to experiment by clicking and searching items that catch your eye.

In Spring 2012, I videotaped my Intermediate Microeconomics classes at DePauw University. They are about an hour long and are freely available at www.depauw.edu/learn/microexcel/videos.htm. The introduction lecture covers material from this chapter.

Part I

The Theory of Consumer Behavior

Perhaps science does not develop by the accumulation of individual discoveries and inventions.

Thomas S. Kuhn

Overview

The material in this book is organized into three parts. The first part focuses on the Theory of Consumer Behavior and derives the demand curve. The second part derives the supply curve from the Theory of the Firm. Finally, these curves are combined to explain how the Market System functions as a decentralized resource allocation mechanism.

Figure I.1 expands the material in the first part, the Theory of Consumer Behavior, to give a preview of upcoming topics. The Optimal Choice chapter is key because it shows how to solve the consumer's optimization problem, but the chapter that follows is especially critical. It applies comparative statics analysis, changing the price of a good, holding everything else constant, to derive a demand curve. This is the most important concept in the Theory of Consumer Behavior.

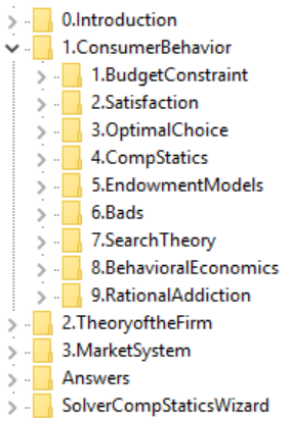


Figure I.1: Content map with focus on consumer behavior.

Focus on the repeated patterns as you work through this material. Economics has a core logic that has been referred to as "the economic way of thinking" or "the economic approach." Learning to see and think like an economist should be your ultimate goal.

References

The epigraph is from the second page of the introductory chapter to Thomas S. Kuhn's classic, *The Structure of Scientific Revolutions* (originally published in 1962). Kuhn argued that progress in science is not generated by incremental puzzle solving (what he called normal science), but that periods of calm are followed by crises that lead to paradigm shifts. The book was as revolutionary as the material it covered, causing debate and controversy in philosophical and scientific circles.

Kuhn would not have been surprised to hear that the derivation of the demand curve did not proceed in an incremental, linear fashion. In fact, the idea of demand for a product depending on the price was known well before we drew graphs of demand curves (in the second half of the 19^{th} century). It was not until economics adopted quantitative and mathematical techniques (what we now call the Marginal Revolution) that the theory of consumer behavior was developed and we could mathematically derive a demand curve.

If we hold money income constant and allow the price of X to change, the price ratio line will rotate about a pivot on the Y axis.

Milton Friedman

Chapter 1

Budget Constraint

The basic idea of the Theory of Consumer Behavior is simple: Given a budget constraint, the consumer buys a combination of goods and services that maximizes satisfaction, which is captured by a utility function. By changing the price of a particular item, ceteris paribus (everything else held constant), we derive a demand curve for that item.

Setting up and solving the consumer's utility maximization problem takes some time. We will proceed slowly and carefully. This chapter focuses on the budget constraint and how it changes when prices or income change.

What can be afforded is obviously a key factor in predicting buying behavior, but it is only part of the story. With the budget constraint alone, we cannot answer the question of how much the consumer wants to buy of each product because we are not incorporating any information about the utility gained by consumption. After we understand the budget constraint, we will model the consumer's likes and dislikes. We can then put the constraint and utility components together and solve the model.

The Budget Constraint in Equation Form

The budget constraint can be expressed mathematically like this:

$$p_1 x_1 + p_2 x_2 \leq m$$

This equation says that the sum of the amount of money spent on good x_1, which is the price of x_1 times the number of units purchased, or $p_1 x_1$, and the amount spent on good x_2, which is $p_2 x_2$, must be less than or equal to the amount of income, m (for money), the consumer has available.

Obviously, the model would be more realistic if we had many products that the consumer could buy, but the gain in realism is not worth the additional cost in computational complexity. We can easily let x_2 stand for "all other goods."

Another simplification allows us to transform the inequality in the equation to a strict equality. We will assume that no time elapses so there is no saving (not spending all of the income available) or borrowing. In other words, the consumer lives for a nanosecond – buying, consuming, and dying the same instant. Once again, this assumption is not as severe as it first looks. We can incorporate saving and borrowing in this model by defining one good as present consumption and the other as future consumption. We will use this modeling technique in a future application.

Since we know we will always spend all of our income, the budget constraint equation can be written with an equal sign, like this

$$p_1 x_1 + p_2 x_2 = m$$

Since we will want to draw a graph, we can write in the form of the equation of a line ($y = mx + b$) via a little algebraic manipulation:

$$p_1 x_1 + p_2 x_2 = m$$

$$p_2 x_2 = m - p_1 x_1$$

$$x_2 = \frac{m}{p_2} - \frac{p_1}{p_2} x_1$$

The intercept, m/p_2, is interpreted as the maximum amount of p_2 that the consumer can afford. By buying no x_1 and spending all income on x_2, the most the consumer can buy is m/p_2 units of good 2.

The slope, $-p_1/p_2$, also has a convenient interpretation: It states the rate at which the market requires the consumer to give up x_2 in order to acquire x_1. This is easy to see if you remember that the slope of a line is simply the rise (Δx_2) over the run (Δx_1). Then,

$$\frac{\Delta x_2}{\Delta x_1} = -\frac{p_1}{p_2}$$

A Numerical Example of the Budget Constraint

STEP Open the Excel workbook *BudgetConstraint.xls*, read the *Intro* sheet, and then go to the *Properties* sheet to see the budget constraint.

Figure 1.1 shows the organization of the sheet. As you can see, the consumer chooses the amounts of goods 1 and 2 to purchase, given prices and income.

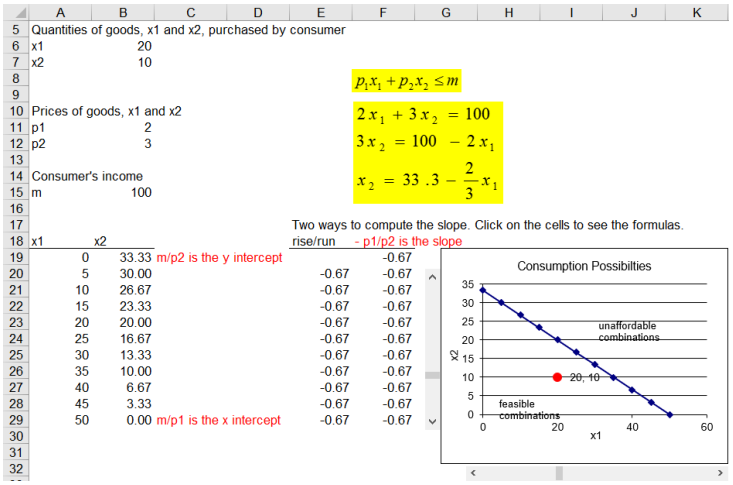


Figure 1.1: The budget line.
Source: BudgetConstraint.xls!Properties

With $p_1 = \$2$/unit, $p_2 = \$3$/unit and $m = \$100$, the equation of the budget line can be computed.

STEP Click on the scroll bars to see the red dot (which represents the consumption bundle), move around in the chart.

By rewriting the budget constraint equation as a line and then graphing it, we have a geometric representation of the consumer's consumption possibilities. All points inside or on the budget line are feasible. Points northeast of the budget line are unaffordable.

By clicking the scroll bars you can easily see that the consumer has many feasible points. The big question is, Which one of these many affordable combinations will be chosen? We cannot answer that question with the budget constraint alone. We need to know how much the consumer likes the two goods. The constraint is simply about feasible options.

Changes in the Budget Line – Pivots and Shifts

STEP Proceed to the *Changes* sheet.

The idea here is that changes in prices cause the budget line to *pivot* or *rotate*, altering the slope, but keeping one of the intercepts the same. Note that changes in income produce a different result, *shifting* the budget line in or out, leaving the slope unchanged.

STEP To see how the budget line pivots, experiment with cell K9 (the price of good 1). Change it from 2 to 5.

The chart changes to reveal a new budget line. The budget line has rotated around the y intercept because if the consumer decided to spend all income on x_2, the amount that could be purchased would remain the same.

If you lower the price of good 1, the budget line swings out. Confirm that this is true.

STEP Changing cell K10 alters the budget line by changing the price of good 2. Once again, change values in the cell to see the effect on the budget line.

STEP Next, click the Reset button to return the sheet to its initial values and work with cell K13. Cut income in half. The effect is dramatically different. Instead of rotating, the budget line has shifted in. The slope remains the same because prices have not changed. Increasing income shifts the budget line out.

This concludes the basics of budget lines. It is worth spending a little time playing with cells K9, K10, and K13 to reinforce understanding of the way budget lines move when there is a change in a price or income. These shocks will be used again when we examine how a consumer's optimal decision changes when prices or income change.

Remember the key lesson: Change in price *rotates* the budget line, but change in income *shifts* it.

Funky Budget Lines

In addition to the standard, linear budget constraint, there are many more complicated scenarios facing consumers. To give you a taste of the possibilities, let us review two examples.

STEP Proceed to the *Rationing* sheet.

In this example, in addition to the usual income constraint, the consumer is allowed a maximum amount of one of the goods. Thus, a second constraint (a vertical line) has been added. When the maximum is above the x_1 intercept (50 units), this second constraint is said to be nonbinding. As you can see from the sheet, when the maximum amount constraint is binding, it lops off a portion of the budget line.

STEP Change cell E13 to see how changing the rationed amount affects the budget constraint.

As we increase the amount of the subsidy, the horizontal line is extended. The downward sloping part has the same slope, but it is pushed outwards,

STEP Proceed to the *Subsidy* sheet.

In this example, in addition to the usual income constraint, the consumer is given a subsidy in the form of a fixed amount of the good.

Food stamps are classic example of subsidies. Suppose the consumer has \$100 of income, but is given \$20 in food stamps (which can only be spent on food), and food (x_1) is priced at \$2/unit. Then the budget constraint has a horizontal segment from 0 to 10 units of food because the most x_2 (other goods) that can be purchased remains at m/p_2 from 0 to 10 units of food (since food stamps cannot be used to buy other goods).

STEP Change cell E13 to see how changing the given amount of food (which is the dollar amount of food stamps divided by the price of food) affects the budget constraint.

Summary: Consumption Possibilities

The budget constraint is a key component of the optimization problem facing the consumer. Graphing the constraint lets us see the consumer's options. Just like a production possibilities frontier tells us what an economy can produce, the budget constraint shows what a consumer can buy. Any combination on or under the constraint is a feasible option. Points beyond the constraint are unattainable.

Changing prices has a different effect on the constraint than changing income. If prices change, the budget line pivots, swings, and rotates (pick your favorite word and remember it) around the intercept. A change in income, however, shifts the line (out or in) and leaves the slope unaffected.

The basic budget constraint is a line, but there are many other scenarios faced by consumers in which the constraint can be kinked or nonlinear. Subsidies (like food stamps) can be incorporated into the basic model. This flexibility is one of the powerful features of the Theory of Consumer Behavior.

The constraint is just one part of the consumer's optimization problem. The desirability of goods and services, also known as tastes and preferences, is another important part. The next chapter explains how we model satisfaction from consuming goods and services.

Exercises

1. Use Excel to create a chart of a budget constraint that is based on the following information: $m = \$100$ and $p_2 = \$3$/unit, but $p_1 = \$2$/unit for the first 20 units and \$1/unit thereafter. Copy your chart and paste it in a Word document.

   STEP Watch a quick, 3-minute video of how to make a chart in Excel by visiting vimeo.com/econexcel/how-to-chart-in-excel.

2. If the good on the y axis is free, what does the budget constraint look like?

3. What combination of shocks could make the new budget line be completely inside and steeper than the initial budget line?

4. What happens to the budget line if all prices and income doubles?

References

The epigraph of this chapter can be found on page 48 of Milton Friedman's revised edition of his *Price Theory* text. The book is essentially his lecture notes from the famous two-quarter price theory course that Friedman delivered for many years at the University of Chicago. It is interesting to see how Micro was taught back then, especially how little emphasis was placed on mathematics. The problems in appendix B are truly thought provoking.

Chapter 2

Satisfaction

Preferences

Utility Functions

> [Indifference] curves are negatively sloped, pass through every point in commodity space, never intersect, and are concave from above. The last-mentioned property implies that the marginal rate of substitution of X for Y diminishes as X is substituted for Y so as to maintain the same level of satisfaction.
>
> C. E. Ferguson

2.1 Preferences

The key idea is that every consumer has a set of likes and dislikes, desires, and tastes, called *preferences*. Consumer preferences enable them to compare any two combinations or bundles of goods and services in terms of better/worse or the same. The result of such a comparison has two outcomes:

- Strictly preferred: the consumer likes one bundle better than the other.
- Indifferent: the consumer is equally satisfied with the two bundles.

In terms of algebra, you can think of strictly preferred as greater than ($>$), indifferent as equal ($=$).

Since the consumer can compare any two bundles, then by repeated comparison of different bundles the consumer can rank all possible combinations from best to worst (in the consumer's opinion).

Three Axioms

Three fundamental assumptions are made about preferences to ensure internal consistency:

1. Completeness: the consumer can compare any bundles and render a preferred or indifferent judgment.
2. Reflexivity: this identity condition says that the consumer is indifferent when comparing a bundle to itself.
3. Transitivity: this condition defines an orderly relation among bundles so that if bundle A is preferred to bundle B and bundle B is preferred to bundle C then bundle A must be preferred to bundle C.

Completeness and reflexivity are easily accepted. Transitivity, on the other hand, is controversial. As a matter of pure logic, we would expect that a consumer would make consistent comparisons. In practice, however, consumers may make intransitive, or inconsistent, choices.

An example of intransitivity: You claim to like Coke better than Pepsi, Pepsi better than RC, and RC better than Coke. The last claim is inconsistent with the first two. If Coke beats Pepsi and Pepsi beats RC, then Coke must really beat RC!

In mathematics, numbers are transitive with respect to the comparison operators greater than, less than, or equal to. Because 12 is greater than 8 and 8 is greater than 3, clearly 12 is greater than 3.

Sports results, however, are not like math. Outcomes of games can easily yield intransitive results. Michigan might beat Indiana and in its next game Indiana could defeat Iowa, but few people would claim that the two outcomes would guarantee that Michigan will win when it plays Iowa.

When we assume that preferences are transitive, it means that the consumer can rank bundles without any contradictions. It also means that we are able to determine the consumer's choice between two bundles based on answers to previous comparisons.

Displaying Preferences via Indifference Curves

The consumer's preferences can be *revealed* by having her choose between bundles. We can describe a consumer's preferences with an *indifference map*, which is made up of *indifference curves*.

A single indifference curve is the set of combinations that give equal satisfaction. If two points lie on the same indifference curve, this means that the consumer sees these two bundles as tied – neither one is better nor worse than the other.

A single indifference curve and an entire indifference map can be generated by having the consumer choose between alternative bundles of goods. We can demonstrate how this works with a concrete example.

STEP Open the Excel workbook *Preferences.xls*, read the *Intro* sheet, and then go to the *Reveal* sheet to see how preferences can be mapped and the indifference curve revealed.

STEP Begin by clicking the Ask? button. For bundle B, enter 4, then a comma (,), then a 3, then click OK.

We are using the coordinate pair notation so 4,3 identifies a combination that has 4 units of the good on the x axis and 3 units of the good on the y axis.

The sheet records the bundles that are being compared in columns A and B and the outcome in column C. The choices are being made by a virtual consumer whose unknown preferences are in the computer. By asking the virtual consumer to make a series of comparisons, we can reveal the hidden preferences in the form of an indifference curve and indifference map.

Notice that Excel plots the point 4,3 on the chart. The green square means the consumer chose bundle B. This means that 3,3 and 4,3 are not on the same indifference curve.

STEP Click the Ask? button again. Offer the consumer a choice between 3,3 and 2,3.

This time the consumer chose bundle A and a red triangle was placed on the chart, meaning that the point 3,3 is strictly preferred to the point 2,3.

These two choices illustrate *insatiability*. This means that the consumer cannot be sated (or filled up) so more is always better. The combination 4,3 is preferred to 3,3, which is preferred to 2,3 because good x_2 is held constant at 3 and this consumer is insatiable, preferring more of good x_1 to less.

To reveal the indifference curve of this consumer, we must offer tougher choices, where we give more of one good and less of the other.

STEP Click the Ask? button again. This time offer the consumer a choice between 3,3 and 4,2.

The consumer decided that 3,3 is better. This reveals important information about the consumer's preferences. At 3,3, the consumer likes one more unit of x_1 less than the loss of one unit of x_2.

STEP Click the Ask? button several times more to figure out where the consumer's break-even point is in terms of how much x_2 is needed to balance the gain from the additional unit of x_1. Offer 4,2.5 and then try taking away less of good 2, such as 2.7 or 2.9. Once you find the point where the amount of x_2 taken away exactly balances the gain in x_1 of one unit (from 3 to 4), you have located two points on a single indifference curve. If it is difficult to see the points on the chart, use the Zoom control to magnify the screen (say to 200%).

You should find that this consumer is indifferent between the bundles 3,3 and 4,2.9.

STEP Now click the 100 Random ? button.

One hundred pairwise comparisons are made between 3,3 and a random set of alternatives. It is easy to see that the consumer can compare each and every point on the chart to the benchmark bundle of 3,3 and judge each and every point as better, worse, or the same.

STEP Click the Indifference button to display the indifference curve that goes through the benchmark point (3,3), as shown in Figure 2.1. Your version will be similar, but not exactly the same as Figure 2.1 since the 100 dots are chosen randomly.

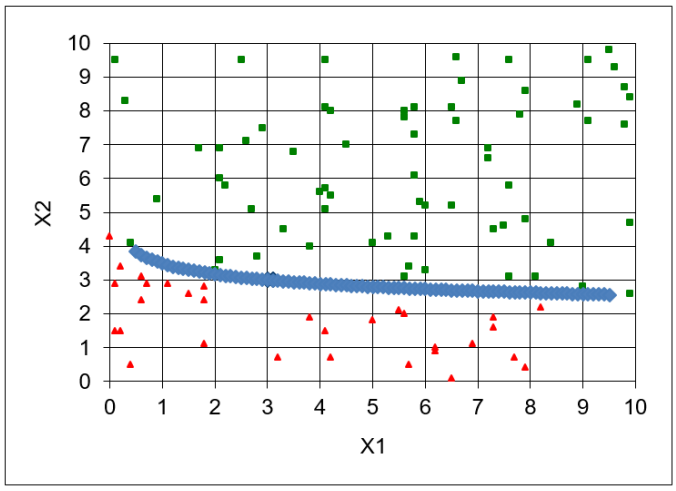


Figure 2.1: Revealing the indifference curve.
Source: Preferences.xls!Reveal

The indifference curve shows the bundles that are the same to this consumer compared to 3,3. All of the bundles for which the consumer is indifferent to the 3,3 bundle lie on the same indifference curve.

The Indifference Map

Every combination of goods has an indifference curve through it. We often display a few representative indifference curves on a chart and this is called an indifference map, as shown in Figure 2.2.

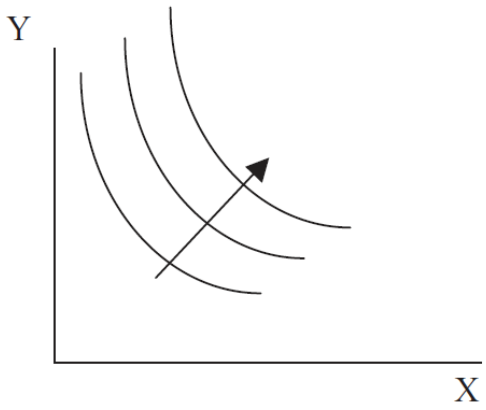


Figure 2.2: An indifference map.

Any point on the curve farthest from the origin, in Figure 2.2, is preferred to any point below it, including the ones on the two lower indifference curves. The arrow indicates that satisfaction increases as you move northeast to higher indifference curves.

There are many (in fact, an infinity) of indifference curves and they are not all depicted when we draw an indifference map. We draw just a few curves. We say that the indifference map is *dense*, which means there is a curve through every point.

STEP Build your own indifference map by copying the *Reveal* sheet and clicking the Reset button, then the Indifference button, and then the Copy Picture button.

This places a picture of the chart under the chart. This is an Excel drawing object, not a chart object, and it has no fill.

STEP Change the benchmark to 4,4 in cell B1 and click the Indifference button to get the indifference curve through the new benchmark point. Click the Copy Picture button.

This copies the chart and pastes the drawing object over the first one. Since it has no fill, it is transparent. You can separate the two pictures if you wish (click and drag), then undo the move so it is on top of the first picture.

STEP Add one more indifference curve to your map by changing the benchmark to 5,5 and clicking the Indifference button, then clicking the Copy Picture button.

You have created an indifference map with three representative indifference curves. Satisfaction increases as you move northeast to higher indifference curves.

Marginal Rate of Substitution

Having elicited a single indifference curve from the virtual consumer in the Excel workbook, we can define and work with a crucial concept in the Theory of Consumer Behavior: the *Marginal Rate of Substitution*, or MRS.

The MRS is a single number that tells us the willingness of a consumer to exchange one good for another from a given bundle. The MRS might be -18 or -0.07. Read carefully and work with Excel so that you learn what these numbers are telling you about the consumer's preferences.

STEP Return to the *Reveal* sheet (with benchmark point 3,3) and click the Copy Picture button to copy and paste an image of the current indifference curve below the graph in the *Reveal* sheet. Now click the New Preferences button to get a new virtual consumer with different preferences and then display the indifference curve for this new consumer (by clicking the Indifference button).

Notice that the indifference curve is not the same as the original one. These are two different consumers with different preferences. You can use the buttons to offer the new consumer bundles that can be compared with the 3,3 benchmark bundle, just like before.

The key idea here is that at 3,3, we can measure each consumer's willingness to trade x_2 in exchange for x_1.

Initially (as shown in Figure 2.1 and in the picture you took), we saw that the consumer was indifferent between 3,3 and 4,2.9. For one more unit of x_1 (from 3 to 4), the consumer is willing to trade 0.1 units of x_2 (from 3 to 2.9). Then the MRS of x_1 for x_2 from 3,3 to 4,2.9 is measured by $\frac{-0.1}{1}$, or -0.1.

With our new virtual consumer, the MRS at 3,3 is a different number. Let's compute it.

STEP Proceed to the MRS sheet. Click the Indifference button. Not only is the indifference curve through 3,3 displayed for this consumer, it also shows some of the bundles that lie on this indifference curve. We can use this information to compute the MRS.

You can compute the MRS at 3,3 by looking at the first bundle after 3,3. How much x_2 is the consumer willing to give up in order to get 0.1 more of x_1? This ratio, $\frac{\Delta x_2}{\Delta x_1}$, (the usual "rise over the run" definition of the slope), is the slope of the indifference curve, which is also the MRS.

The MRS also can be computed as the slope of the indifference curve *at* a point by using derivatives. Instead of computing $\frac{\Delta x_2}{\Delta x_1}$ along an indifference curve from one point to another, one can find the instantaneous rate of change at 3,3. We will do this later.

The crucial concept right now is that the MRS is a number that measures the willingness of a consumer to trade one good for another *at a specific point*. We usually think of it in terms of giving up some of the good on the y axis to get more of the good on the x axis.

Do not fall into the trap of thinking of the MRS as applying to the entire indifference curve. In fact, the MRS is different at each point on the curve. For a typical indifference curve like in Figure 2.1, the MRS gets smaller (in absolute value) as we move down the curve (as it flattens out).

The MRS is *negative* because the indifference curve is sloping downwards: a *decrease* in x_2 is compensated for by an *increase* in x_1. We often drop the minus sign because comparing negative numbers can be confusing. For example, say one consumer has an MRS of -1 at 3,3 while another has an MRS of $-\frac{1}{3}$ at that point. It is true that -1 is a smaller number than $-\frac{1}{3}$,

however, we to use the MRS to indicate the steepness of the slope. Thus, to avoid confusion, we make the comparison using the absolute value of the MRS. Figure 2.3 shows that the bigger in absolute value is the MRS, the

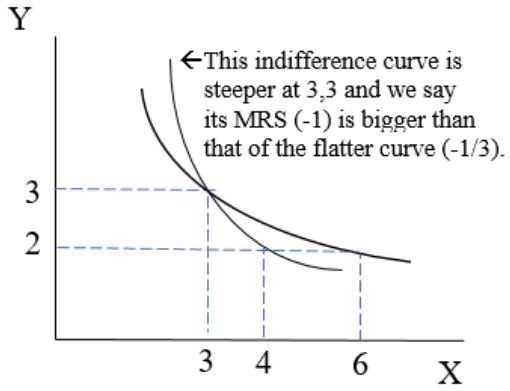


Figure 2.3: Comparing MRS.

more the consumer is willing to trade the good on the y axis for the good on the x axis. Thus, an MRS of -1 at 3,3 means the indifference curve has a steeper slope at that point than if the MRS was $-\frac{1}{3}$. We would say the MRS is bigger at -1 than $-\frac{1}{3}$ even though -1 is a smaller number than $-\frac{1}{3}$ because we look only at the absolute value of the MRS.

Funky Preferences and Their Indifference Curves

We can depict a wide variety of preferences with indifference maps. Here are some examples.

Example 1: Perfect Substitutes — constant slope (MRS)

If the consumer perceives two things as perfectly substitutable, it means they can get the same satisfaction by replacing one with the other.

Consider having one five-dollar bill and five one-dollar bills (as long as we are not talking about several hundred dollars worth of bills). If the consumer does not care about having $10 as a single ten-dollar bill, one five-dollar bill and five one-dollar bills, or ten one-dollar bills, then the indifference curve is a straight line as shown in Figure 2.4. You could argue that there is an indivisibility here and there are actually just 3 points that should not be connected by a line, but the key idea is that the indifference curve is a straight

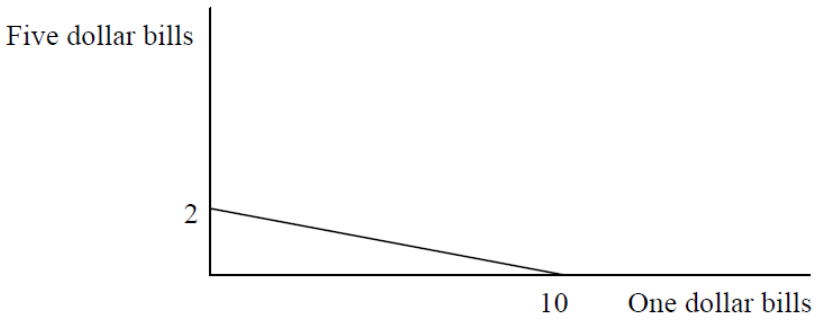


Figure 2.4: Perfect substitutes.

line in the case of perfect substitutes. It has a constant MRS (the slope of the line is $-\frac{1}{5}$), unlike a typical indifference curve where the MRS falls (in absolute value) as you move down the curve.

Example 2: Perfect Complements — L-shaped Indifference Curves

The polar opposite of perfect substitutes are perfect complements. Suppose the goods in questions have to be used in a particular way, with no room for any flexibility at all, like cars and tires. You need four tires for a car to work. With only three tires the car is worthless. Ignoring the spare, having more than four tires does not help you if you still have just one car.

Figure 2.5 illustrates the indifference map for this situation. It says that eight tires with one car gives the same satisfaction as four tires with one car. It also says that eight tires and two cars is preferred to four tires and one car (or eight tires and one car) because the middle L-shaped indifference curve (I_1) is farther from the origin than the lowest indifference curve (I_0).

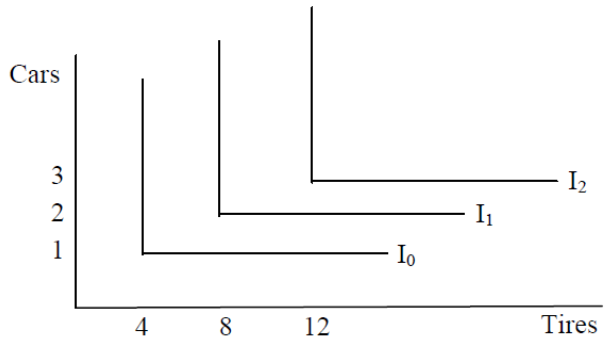


Figure 2.5: Perfect complements.

Notice how the usual indifference curve lies between the two extremes of perfect substitutes (straight lines) and perfect complements (L-shaped). Thus, the typical indifference curve reflects a level of substitutability between goods that is more than perfect complements (one good cannot replace another at all), but less than perfect substitutes (one good can take the place of another with no loss of satisfaction).

Example 3: Bads

What if one of the goods is actually a *bad*, something that lowers satisfaction as you consume more of it, like pollution? Figure 2.6 shows the indifference map in this case.

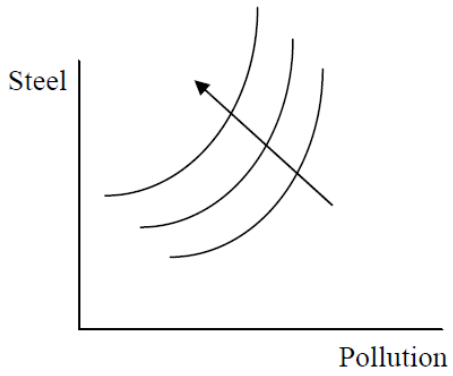


Figure 2.6: Bads.

Along any one of the indifference curves, more steel and more pollution are equally satisfying because pollution is a bad that cancels out the additional good from steel. The arrow indicates that satisfaction increases by moving northwest, to higher indifference curves.

Example 4: Neutral Goods

What if the consumer thinks something is neither good nor bad? Then it is a *neutral good* and the indifference map looks like Figure 2.7.

The horizontal indifference curves for the neutral good on the x axis in Figure 2.7 tell you that the consumer is indifferent if offered more X. The arrow indicates that satisfaction rises as you move north (because Y is a good and having more of it increasing satisfaction).

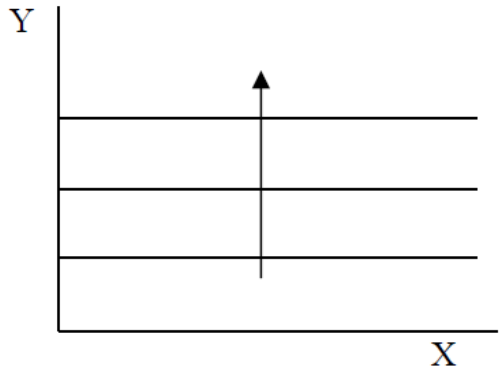


Figure 2.7: X is a neutral good.

These are just a few examples of how a variety of preferences can be depicted with an indifference map. When we want to describe generic, typical preferences that produce downward sloping indifference curves, as in Figure 2.2, economists use the phrase "well-behaved preferences."

Another technical term that is often used in economics is *convexity*, as in convex preferences. This means that midpoints are preferred to extremes. In Figure 2.8, there are two extreme points, A and B, which are connected by a dashed line. Any point on the dashed line, like C, can be described by the equation $zA + (1 - z)B$, where $0 < z < 1$ controls the position of C. This equation is called a convex combination.

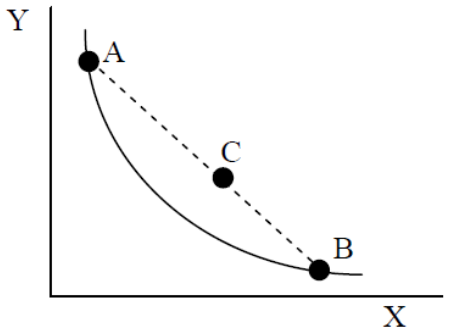


Figure 2.8: Convex preferences.

If preferences are convex, then midpoints like C are strictly preferred to extreme points like A and B. Convexity is used as another way of saying that preferences are well-behaved.

An important property that arises out of well-behaved or convex preferences is that of *diminishing MRS*. As explained earlier, the MRS varies along an indifference curve and applies to a specific point (not to the entire curve). The MRS will start large (in absolute value) at the top left corner, like point A in Figure 2.8, and get smaller as we travel down the indifference curve to point B. This makes common sense. The consumer is readily willing to trade a lot of Y for X (so the MRS is high in absolute value) when he has a lot of Y and little X. When the amounts are reversed, such as point B, a small MRS means he is willing to give up very little Y (since he has little of it) for more X (which he has a lot of already).

Indifference Curves Reflect Preferences

Preferences, a consumer's likes and dislikes, can be elicited or revealed by asking the consumer to pick between pairs of bundles. The indifference curve is that set of bundles that the consumer finds equally satisfying.

The MRS is a single number that measures the willingness of the consumer to exchange one good for another at a particular point. If the MRS is high (in absolute value), the indifference curve is steep at that point and the consumer is willing trade a lot of Y for a little more X.

Standard, well-behaved preferences yield a set of smooth arcs (like Figure 2.2), but there are many other shapes that depict preferences for different kinds of goods and the relationship between goods.

Exercises

1. What is the MRS at any point if X is a neutral good? Explain why.
2. If the good on the y axis was a neutral good and the other good was a regular good, then what would the indifference map look like. Use Word's Drawing Tools to draw a graph of this situation.
3. If preferences are well-behaved, then indifference curves cannot cross. Use Figure 2.9 to help you construct an explanation for why this claim must be true. Note that point C has more X and Y than point A, thus, by insatiability, C must be preferred to A. The key to defending the claim lies in the assumption of transitivity.

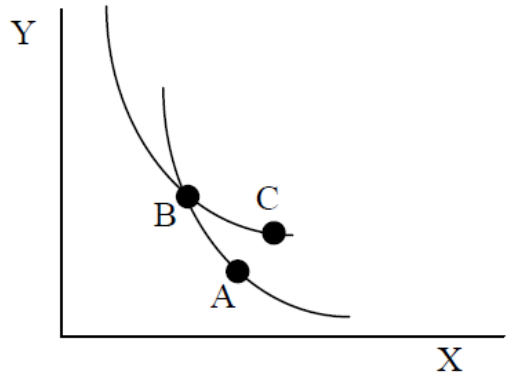


Figure 2.9: An impossible indifference map.

4. Suppose we measure consumer A's and B's MRS at the same point and find that $MRS_A = -6$ and the $MRS_B = -2$. What can we say about the preferences of A and B at this point?

References

The epigraph is from page 26 of C. E. Ferguson's *Microeconomic Theory* (revised edition, 1969), a popular micro text in the 1960s and 1970s. In the preface Ferguson wrote, "This is a textbook; its content is taken from the public domain of economic literature. Conventional topics are treated in conventional ways; and there is no real innovation." Perhaps, but Ferguson adopted a much more mathematical presentation and added content, including general equilibrium theory, that made his book different and important.

> [A] cardinal measure of utility is in any case unnecessary; only an ordinal preference, involving "more" or "less" but not "how much," is required for the analysis of consumer's behavior.
>
> Paul A. Samuelson

2.2 Utility Functions

Previously, we showed that a consumer has preferences that can be revealed and mapped. The next step is to identify a particular functional form, called a *utility function*, which faithfully represents the person's preferences. Once you understand how the utility function works, we can combine it with the budget constraint to solve the consumer's optimization problem.

Cardinal and Ordinal Rankings

Jeremy Bentham (1748-1832) was a utilitarian philosopher who believed that, in theory, the amount of utility from consuming a particular amount of a good could be measured. So, for example, as you ate an apple, we could hook you up to some device that would report the number of "utils" of satisfaction received. The word *utils* is in quotation marks because they do not actually exist, but Bentham believed they did and would one day be discovered with an advanced measuring instrument. This last part is not so crazy—an fMRI machine is exactly what he envisioned.

Bentham also believed that utils were a sort of common currency that enabled them to be compared across individuals. He thought society should maximize aggregate or total utility and utilitarianism has come to be associated with the phrase "the greatest happiness for the greatest number." Thus, if I get 12 utils from consuming an apple and you get 6, then I should get the apple. Utilitarianism also implies that if I get more utils from punching you in the face than you lose, I should punch you. This is why utilitarianism is not highly regarded today.

This view of utility treats satisfaction as if we could place it on a cardinal scale. This is the usual number line where 8 is twice as much as 4 and the difference between 33 and 30 is the same as that between 210 and 207.

Near the turn of the 20th century, Vilfredo Pareto (1848-1923, pronounced pa-RAY-toe) created the modern way of thinking about utility. He held that satisfaction could not be placed on a cardinal scale and that you could never compare the utilities of two people. Instead, he argued that utility could be measured only up to an ordinal scale, in which there is higher and lower, but no way to measure the magnitude between two items.

Notice how Pareto's approach matches exactly the way we assumed that a consumer could choose between bundles of goods as preferring one bundle or being indifferent. We never claimed to be able to measure a certain amount of satisfaction from a particular bundle.

For Pareto, and modern economics, the numerical value from a particular utility function for a given combination of goods has no meaning. These values are like the star ranking system for restaurants.

Suppose Critic A uses a 10-point scale, while Critic B uses a 1000-point scale to judge the same restaurants. We would never say that B's worst restaurant, which scored say 114, is better than A's best, a perfect 10. Instead, we compare their rankings. If A and B give the same restaurant the highest ranking (regardless of the score), it is the best restaurant.

Now suppose we are reading a magazine that uses a 5-star rating system. Restaurant X earns 4 stars and Restaurant Y 2 stars. X is better, but can we conclude that X is twice as good as Y? Absolutely not. An ordinal scale is ordered, but the differences between values are not important.

Pareto revolutionized our understanding of utility. He rejected Bentham's cardinal scale because he did not believe that satisfaction could be measured like body temperature or blood pressure. Pareto showed that we could derive demand curves with the less restrictive more-or-less ranking of bundles.

The transition from Bentham's cardinal view of utility to Pareto's ordinal view was not easy. Using the same word, utility, creates confusion (although, to be fair, Pareto tried to create a new word, *ophelimity*, but it never caught on). It bears repeating that, for a modern economist, although a utility function will show numerical values, these should not be interpreted on a cardinal scale, nor should numerical utilities of different people be compared. Since we cannot make interpersonal utility comparisons to add utilities of different people, we cannot give me the apple or let me punch you.

Monotonic Transformation

Once we reveal the consumer's indifference curve and map, we have the consumer's rankings of all possible bundles. Then, all we need to do is use a function that faithfully represents the indifference curves. The utility function is a convenient way to capture the consumer's ordering.

There are many (in fact, an infinity) of functions that could work. All the function has to do is preserve the consumer's preference ranking.

A *monotonic transformation* is a rule applied to a function that changes (transforms) it, but maintains the original order of the outputs of the function for given inputs. Monotonic is a technical term that means always moving in the same direction.

For example, star ratings can be squared and the rankings remain the same. If X is a 4-star and Y a 2-star restaurant, we can square them. X now has 16 stars and Y has 4 stars. X is still higher ranked than Y. In this case, squaring is a monotonic transformation because it has preserved the ordering and X is still higher than Y.

Can we conclude that X is now four times better? Of course not. Remember that the star ranking is an ordinal scale so the distance between items is irrelevant. We say that squaring is a monotonic transformation because it maintains the same ordering and we do not care about the distances between the numeric values. Their only meaning is "higher" and "lower," which indicate better and worse.

It is a fact that the MRS (at any point) remains constant under any monotonic transformation. This is an important property of monotonic transformations that we will illustrate with a concrete example in Excel.

Cobb-Douglas: A Ubiquitous Functional Form

STEP Open the Excel workbook *Utility.xls*, read the *Intro* sheet, and then go to the *CobbDouglas* sheet to see an example of this utility function:

$$u(x_1, x_2) = x_1^c x_2^d$$

In economics, a function created by multiplying variables that are raised to powers is called a *Cobb-Douglas functional form.*

STEP Follow the directions on the sheet (in column K) to rotate the 2D chart so you are looking down at it.

A top-down view of the utility function looks like an indifference map. The utility function itself, in 3D, is a hill or mountain (that keeps growing without ever reaching a top—illustrating the idea of insatiability).

With a utility function, the indifference curves appear as contour lines or level curves. The curves in 2D space are created by taking horizontal slices of the 3D surface. Every point on the indifference curve has the exact same height, which is utility.

STEP The exponents (c and d) in the utility function express "likes and dislikes." Try $c = 4$ then $c = 0.2$ in cell B5.

The higher the c exponent, the more the consumer likes x_1 because each unit of x_1 is raised to a higher power as c increases. Notice that when $c = 4$, the fact that the consumer likes x_1 much more than when $c = 0.2$ is reflected in the shape of the indifference curve. The steeper the indifference curve, the higher the MRS (in absolute value) and the more the consumer likes x_1.

STEP Proceed to the *CobbDouglasLN* sheet, which applies a monotonic transformation of the Cobb-Douglas function. It applies the natural log function to the utility function.

Recall that the natural logarithm of a number x is the exponent on e (the irrational number 2.7128 . . .) that makes the result equal x. You should also remember that there are special rules for working with logs. Two especially common rules are $\ln(x^y) = y \ln x$ and $\ln(xy) = \ln x + \ln y$. We can apply these rules to the Cobb-Douglas function when we take the natural log:

$$u(x_1, x_2) = x_1^c x_2^d$$

$$\ln[u(x_1, x_2)] = \ln[x_1^c x_2^d]$$

$$\ln[u(x_1, x_2)] = c \ln x_1 + d \ln x_2$$

The *CobbDouglasLN* sheet applies the natural log transformation by using Excel's LN() function.

STEP Click on any cell between B12 and Q27 to see the formula. We are computing the natural log of utility, which is x_1 raised to the c power times x_2 raised to the d power.

How does the original utility function compare to its natural log version?

STEP Go back and forth a few times between the two (click on the *CobbDouglas* sheet tab and then the *CobbDouglassLN* sheet tab). It is obvious that the numbers are different.

But did you notice something curious?

STEP Compare the cells with yellow backgrounds in the two sheets to see that these two combinations continue to lie on the same indifference curve, even though the utility values of the two functions are different.

The fact that the cells remain on the same indifference curve after undergoing the natural log transformation demonstrates the meaning of a monotonic transformation. The utility values are different, but the ranking has been preserved. The two utility functions both maintain the same relationship between 1,14 and 2,7 and every other bundle.

So now you know that a Cobb-Douglas utility function can be used to faithfully represent a consumer's preferences (including tweaking the c and d exponents to make the curves steeper or flatter) and that we can use the natural log transformation if we wish. In addition, economists often use the Cobb-Douglas functional form for utility (and production) functions because it has very nice algebraic properties where lots of terms cancel out.

The Cobb-Douglas function is especially easy to work with if you remember the following rules:

Algebra Rules: $\frac{x^a}{x^b} = x^{a-b}$ and $x^{a^b} = x^{ab}$

Calculus Rule: $\frac{dax^b}{dx} = bax^{b-1}dx$

These rules may seem irrelevant right now, but we will see that they make the Cobb-Douglas function much easier to work with than other functions. This goes a long way in explaining the repeated use of the Cobb-Douglas functional form in economics.

Expressing Other Preferences with Utility Functions

STEP Proceed to the *PerfSub* sheet and look around. Scroll down (if needed) and look at the two charts.
Notice how this functional form is producing straight line indifference curves (in the 2D chart). If the consumer treated two goods as perfect substitutes, we would use this functional form instead of Cobb-Douglas. The coefficients (a and b) can be tweaked to make the lines steeper or flatter.

STEP Proceed to the *PerfComp* sheet. This shows how the *min()* functional form produces L-shaped indifference curves.

The *min()* function outputs the smaller of the two terms, ax_1 and bx_2. This means that getting more of one good while holding the amount of the other good constant does not increase utility. This produces an L-shaped indifference curve.

Finally, the *Quasilinear* sheet displays indifference curves that are actually curved, but rather flat.

STEP Go to the *Quasilinear* sheet and click on the different functional form options. These are just a few of the many transformations that can be applied to x_1 and then added to x_2 to produce what is called quasilinear utility. Later, we will see that this functional form has different properties than Cobb-Douglas.

Note that we can represent many different kinds of preferences with utility functions. An important point is that there are many (to be more exact, an infinity) of possible utility functions available to us. We would choose one that faithfully reflects a particular consumer's preferences. We can always apply a monotonic transformation and it will not alter the consumer's preferences.

Computing the MRS for a Utility Function

Now that we have utility functions to represent a consumer's preferences, we are able to compute the MRS from one point to another (like we did in the previous chapter) or by using the instantaneous rate of change, better known as the derivative.

This is not a mathematics book, but economists use math so we need to see exactly how the derivative works. The core idea is convergence: make the change in x (the run) smaller and smaller and the ratio of the rise over the run (the slope) gets closer and closer to its ultimate value. The derivative is a shortcut that gives us the answer without the cumbersome process of making the change smaller and smaller.

But this is way too abstract. We can see it in Excel.

STEP Proceed to the *MRS* sheet to see how the MRS can be computed via a discrete-size change versus an infinitesimally-small change.

The utility function is x_1x_2. This is Cobb-Douglas with exponents (implicitly) equal to 1.

Suppose we are interested in the indifference curve that gives all combinations with a utility of 10. Certainly 5,2 works (since 5 times 2 is 10). It is the red dot in the graph on the *MRS* sheet (and in Figure 2.10).

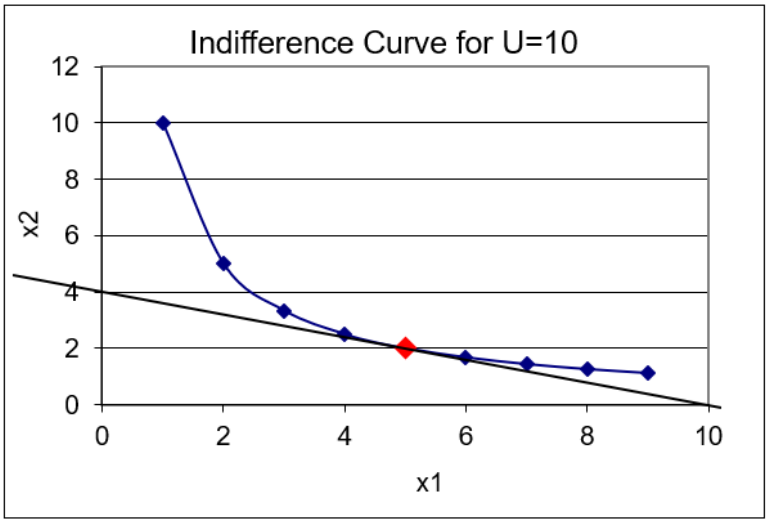


Figure 2.10: Computing the MRS.
Source: Utility.xls!MRS

From the bundle 5,2, if we gave this consumer 1 more unit of x_1, by how much would we have to decrease x_2 to stay on the $U = 10$ indifference curve? A little algebra tells us.

We know that $U = x_1x_2$ and the initial bundle 5,2 yields $U = 10$. We want to maintain U constant with $x_1 = 6$ because we added one unit to x_1, so we have:

$$U = x_1x_2$$
$$10 = 6x_2$$
$$x_2 = \frac{10}{6}$$

We have two bundles that yield $U = 10$ (5,2, and 6,$\frac{10}{6}$). We can compute the MRS as the change in x_2 divided by the change in x_1. The delta (or difference) in x_2 is $-\frac{1}{3}$ (because $\frac{10}{6}$ is $\frac{1}{3}$ less than 2) and the delta in x_1 is 1 (6 - 5), so starting *from* the point 5,2, the MRS from $x_1 = 5$ to $x_1 = 6$ is $-\frac{1}{3}$. This is what Excel shows in cell C18.

Another way to compute the MRS uses the calculus approach. Instead of a "large" or discrete-size change in x_1, we take an infinitesimally small change, computing the slope of the indifference curve not from one point to another, but as the slope of the tangent line (as shown in Figure 2.10). We use the derivative to compute the MRS *at* a particular point.

For this simple utility function, holding U constant at 10, we can rewrite the function as x_2 in terms of x_1, then take the derivative.

$$U = x_1x_2$$

$$x_2 = \frac{10}{x_1}$$

$$\frac{dx_2}{dx_1} = -\frac{10}{x_1^2}$$

At $x_1 = 5$, substitute in this value and the MRS at that point is $-\frac{10}{25}$ or -0.4. This is what Excel shows in cell D18. If you need help with derivatives, the next chapter has an appendix that reviews basic calculus.

Computing the MRS this way relies on the ability to write x_2 in terms of x_1. If we have a utility function that cannot be easily rearranged in this way, we will not be able to compute the MRS. There is, however, a more general approach. The procedure involves taking the derivative of the utility function with respect to x_1 (called the marginal utility of x_1) and dividing by the derivative of the utility function with respect to x_2 (called the marginal utility of x_2). Do not forget to include the minus sign when you use this approach. Here is how it works.

With $U = x_1x_2$, the derivatives are simple: $\frac{dU}{dx_1} = x_2$ and $\frac{dU}{dx_2} = x_1$. Thus, we can substitute these into the numerator and denominator of the MRS expression:

$$MRS = -\frac{\frac{dU}{dx_1}}{\frac{dU}{dx_2}} = -\frac{x_2}{x_1}$$

Because we are considering the point 5,2, we evaluate the MRS at that point (which means we plug in those values to our MRS expression), like this:

$$MRS = -\frac{x_2}{x_1}\Big|_{\substack{x_1=5\\x_2=2}} = -\frac{2}{5} = -0.4$$

Note that minus the ratio of the marginal utilities gives the same answer as the $\frac{dx_2}{dx_1}$ method. Both are using infinitesimally small changes to compute the instantaneous rate of change of the indifference curve at a particular point.

Also note that the ratio of the marginal utilities approach requires that you divide the marginal utility of x_1 (the good on the x axis) by the marginal utility of x_2 (the good on the y axis). Since we used $\frac{\Delta y}{\Delta x}$ in the discrete-size change approach, it is easy to confuse the numerator and denominator when computing the MRS via the derivative. Remember that $\frac{dU}{dx_1}$ goes in the numerator.

Comparing Δ and d Methods

So far, we know there are two ways to get the MRS: move from one point to another along the indifference curve (discrete change, Δ) or slope of the tangent line at a point (infinitesimally small change, d). We also know that we have two ways of doing the latter (solve for x_2 then take the derivative or compute the ratio of the marginal utilities.)

But you may have noticed a potential problem in that the two procedures to get the MRS yield different answers. In the *MRS* sheet and our work above, the discrete change approach tells us that the MRS as measured from $x_1 = 5$ to $x_1 = 6$ is $-\frac{1}{3}$, whereas the derivative method says that the MRS at $x_1 = 5$ is -0.4.

This difference in measured MRS is due to the fact that the two approaches are applying a different size change in x_1 to a curve. As the discrete-size change gets smaller, it approaches the derivative measure of the MRS. You can see this clearly with Excel.

STEP Change the step size in cell B7 to 0.5 and watch how cell C18 changes. Notice that the chart is also slightly different because the point at $x_1 = 6$ is now at 5.5.

You have made the size of the change in x_1 smaller so the point is now closer to the initial value, 5.

STEP Do it again, this time changing the step size in cell B7 to 0.1. The point with $x_1 = 5.1$ is so close to 5 that it is hard to see, but it is there. Do one last change to the step size, setting it at 0.01.

With the step size at 0.01, you cannot see the initial and new points because they are so close together, but they are still a discrete distance apart. Excel displays the point-to-point delta computation in cell C18. It is really close to the derivative measure of the MRS in cell D18 because the derivative is simply the culmination of this process of making the change in x_1 smaller and smaller.

In Figure 2.10, the discrete change approach is computing the rise over the run using two separate points on the curve, while the calculus approach is computing the slope of the tangent line.

STEP Look at the values of the cells in the yellow highlighted row.

The MRS for a given approach are exactly the same. In other words, columns C, H, and M are the same and columns D, I, and N are the same. This shows that the MRS remains unaffected when the utility function is monotonically transformed.

Utility Functions Represent Preferences

Utility functions are equations that represent a consumer's preferences. The idea is that we reveal preferences by having the consumer compare bundles, and then we select a functional form that faithfully reflects the indifference curves of the consumer.

In selecting the functional form, there are many possibilities and economists often use the Cobb-Douglas form. The values of utility produced by inputting amounts of goods are meaningless and any monotonic transformation (because it preserves the preference ordering) will work as a utility function. Monotonic transformations do not affect the MRS.

The MRS is an important concept in consumer theory. It tells us the willingness to trade one good for another and this measure the consumer's likes and dislikes. Willingness to trade a lot of y for a little x produces a high MRS (in absolute value) and this indicates that the consumer values x more than y.

The MRS computed *from* one point to another (Δ), but it can also be computed using the derivative (d) *at* a point. Both are valid and the resulting number for the MRS is interpreted the same way (willingness to trade).

Exercises

The utility function, $U = x - 0.03x^2 + y$, has a quasilinear functional form. Use this function to the answer the questions below. You can see what it looks like by choosing the Polynomial option in the *Quasilinear* sheet.

1. Compute the value of the utility function at bundle A, where $x = 10$ and $y = 1$. Show your work.
2. Working with bundle A, find the MRS as x rises from $x = 10$ to $x = 20$. Show your work.
3. Find the MRS at the point 10,1 (using derivatives). Show your work.
4. Why do the two methods of determining the MRS yield different answers?
5. Which method is better? Why?

References

The epigraph can be found on page 91 of the revised edition of *The Foundations of Economic Analysis*, by Paul Samuelson. This remarkable book, written by one of the greatest economists of the 20^{th} century, took economics to a new level of mathematical sophistication. Samuelson could not have picked a better opening quote, "Mathematics is a Language," by J. Willard Gibbs.

Chapter 3

Optimal Choice

Initial Solution

More Practice and Understanding Solver

Food Stamps

Cigarette Taxes

Joseph Louis Lagrange, the greatest mathematician of the eighteenth century, was born at Turin on January 25, 1736, and died at Paris on April 10, 1813. . . . In appearance he was of medium height, and slightly formed, with pale blue eyes and a colourless complexion. In character he was nervous and timid, he detested controversy, and to avoid it willingly allowed others to take credit for what he had himself done.

W. W. Rouse Ball

3.1 Initial Solution

What you know so far:

1. The *budget constraint* shows the consumer's possible consumption bundles. The standard, linear constraint is $p_1x_1 + p_2x_2 = m$. There are many other situations, such as subsidies and rationing, which give more complicated constraints with kinks and horizontal/vertical segments.

2. The *indifference map* shows the consumer's preferences. The standard situation is a set of convex, downward sloping indifference curves. There are many alternative preferences, such as perfect substitutes and perfect complements. Preferences are captured by utility functions, which accurately reflect the shape of the indifference curves.

Our job is to combine these two parts, one expressing what is affordable and the other what is desirable, to find the combination (or bundle) that maximizes satisfaction (as described by the indifference map or utility function) given the budget constraint. The answer will be in terms of how much the consumer will buy in units of each good.

The optimal solution is depicted by the canonical graph in Figure 3.1. The word *canonical* is used here to mean standard, conventional, or orthodox. In economics, a canonical graph is a core, essential graph that is understood by all economists, such as a supply and demand graph.

It is no exaggeration to say that Figure 3.1 is one of the most fundamental and important graphs in economics. It is the foundation of the Theory of Consumer Behavior and with it we will derive a demand curve.

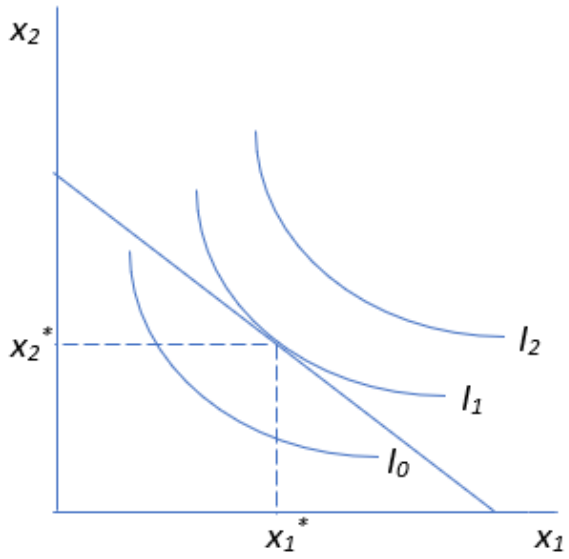


Figure 3.1: The canonical graph of the optimal solution.

One serious intellectual obstacle with Figure 3.1 is that it is highly abstract. Below we work on a concrete problem, with actual numbers, to explain what is going on in this fundamental graph.

Before we dive in, we need to discuss solution strategies. There are two ways to find the optimal solution:

1. Analytical methods using algebra and calculus—this is the conventional, paper and pencil approach that has been used for a long time.

2. Numerical methods using a computer, for example, Excel's Solver—this is a modern solution strategy that uses the computer to do most of the work.

Analytical Approach

Unfortunately, constrained optimization problems are harder to solve than unconstrained problems. The appendix to this chapter offers a short calculus review along with a few common derivative and algebra rules. If the material below makes little sense, go to the appendix and then return here.

Because this is a constrained optimization problem, the analytical approach uses the method developed by Joseph Louis Lagrange. His brilliant idea is based on transforming a constrained optimization problem into an unconstrained problem and then solving by using standard calculus techniques. In the process, a new endogenous variable is created. It can have a meaningful economic interpretation.

Lagrange gave us a recipe to follow that requires four steps:

1. Rewrite the constraint so that it is equal to zero.
2. Form the Lagrangean function.
3. Take partial derivatives with respect to x_1, x_2, and λ.
4. Set the derivatives equal to zero and solve for $x_1{}^*$, $x_2{}^*$, and λ^*.

A Concrete Example

Suppose a consumer has a Cobb-Douglas utility function with exponents both equal to 1 and a budget constraint, $2x_1 + 3x_2 = 100$ (which means the price of good 1 is \$2/unit, the price of good 2 is \$3/unit, and income is \$100).

The problem is to maximize utility subject to (s.t.) the budget constraint. It is written in equation form like this:

$$\max_{x_1,x_2} U(x_1, x_2) = x_1 x_2$$
$$\text{s.t. } 100 = 2x_1 + 3x_2$$

This problem is not solved directly. It is first transformed into an unconstrained problem, and then this unconstrained problem is solved. Here is how we apply the recipe developed by Lagrange.

1. Rewrite the constraint so that it is equal to zero.

$$0 = 100 - 2x_1 - 3x_2$$

2. Form the Lagrangean function.

$$\max_{x_1,x_2,\lambda} L = x_1 x_2 + \lambda(100 - 2x_1 - 3x_2)$$

Most math books use a fancy script L for the Lagrangean, like this $\mathcal{L}$, but this is difficult to do in Word's Equation Editor (which you will be using) so

an extra-large L will work just as well. Also, many books spell Lagrangean with an *i*, Lagrangian, but both spellings are acceptable.

Note that the Lagrangean function, L, is composed of the original objective function (in this case, the utility function) plus a new variable, the Greek letter lambda, λ, times the rewritten constraint. Called the *Lagrangean multiplier*, λ is a new endogenous variable that is introduced as part of Lagrange's solution strategy.

The next step in Lagrange's recipe can be intimidating. This is not the time to rush through and turn the page. Refer to the appendix at the end of this section if things start to get confusing.

3. Take partial derivatives with respect to x_1, x_2, and λ.

$$\frac{\partial L}{\partial x_1} = x_2 - 2\lambda$$
$$\frac{\partial L}{\partial x_2} = x_1 - 3\lambda$$
$$\frac{\partial L}{\partial \lambda} = 100 - 2x_1 - 3x_2$$

The derivative used here is a partial derivative, denoted by ∂, which is an alternative way of writing a lowercase Greek letter d (which is why the more common symbol for the letter δ is also used). The partial derivative symbol is usually read as the letter d, so the first equation read out loud would be "d L d x one equals x two minus two times lambda." It is also common to read the derivative in the first equation as "partial L partial x one."

The partial derivative is a natural extension of the regular derivative. Consider the function $y = 4x^2$. The derivative of y with respect to x is $\frac{dy}{dx} = 8x$. Suppose, however, that we had a more complicated function, like this: $y = 4zx^2$. This multivariate function says that y depends on two variables, z and x. We can explore the rate of change of this function along the x axis by treating it as a partial function, meaning that we hold the z variable constant. Then the partial derivative of y with respect to x is $\partial y/\partial x = 8zx$. If we hold x constant and vary z, then the partial derivative of y with respect to z is $\partial y/\partial z = 4x^2$.

Applying this logic to the Lagrangean in step 2, when we take the partial derivative with respect to x_1, the first term is x_2 because it is as if we had "$x4$" and took the derivative with respect to x, getting 4.

If we multiply λ through the parenthetical expression in the Lagrangean, we get:

$$\lambda(100 - 2x_1 - 3x_2)$$
$$\lambda 100 - \lambda 2x_1 - \lambda 3x_2 = 0$$

The first and third terms on the left-hand side do not have x_1 so the derivative with respect to x_1 is zero (just like the derivative of a constant is zero). The derivative with respect to x_1 of the middle term produces $-\lambda 2$ which is written by convention as -2λ.

Can you do the other two derivatives in step 3?

4. Set the derivatives equal to zero and solve for x_1*, x_2*, and λ*.

$$\frac{\partial L}{\partial x_1} = x_2 - 2\lambda = 0$$
$$\frac{\partial L}{\partial x_2} = x_1 - 3\lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = 100 - 2x_1 - 3x_2 = 0$$

There are many ways to solve this system of equations, which are known as the first-order conditions. Sometimes, this is the hardest part of the Lagrangean method. Depending on the utility function and constraint, there may not be an analytical solution.

A common strategy involves moving the λ terms in the first two equations to the right-hand side and then dividing the first equation by the second one.

$$x_2 = 2\lambda$$
$$x_1 = 3\lambda$$
$$\frac{x_2}{x_1} = \frac{2\lambda}{3\lambda}$$

The λ terms then cancel out, leaving us with two equations (the one above and the third equation from the original three first-order conditions) and two unknowns (x_1 and x_2).

$$\frac{x_2}{x_1} = \frac{2}{3}$$
$$100 - 2x_1 - 3x_2 = 0$$

The top equation has a nice economic interpretation. It says that, at the optimal solution, the MRS (slope of the indifference curve) must equal the price ratio (slope of the budget constraint).

From the top equation, we can solve for x_2.

$$x_2 = \frac{2}{3}x_1$$

We can then substitute this expression into the bottom equation (the budget constraint) to get the optimal value of x_1.

$$\begin{aligned} 100 - 2x_1 - 3\left[\frac{2}{3}x_1\right] &= 0 \\ 100 - 2x_1 - 2x_1 &= 0 \\ 100 &= 4x_1 \\ x_1^* &= 25 \end{aligned}$$

Then we substitute x_1* into the expression for x_2 to get x_2*.

$$x_2 = \frac{2}{3}[25]$$

$$x_2{}^* = 16\frac{2}{3}$$

The asterisk is used to represent the optimal solution for a choice variable. This work says that this consumer should buy 25 units of good 1 and $16\frac{2}{3}$ units of good 2 in order to maximize satisfaction given the budget constraint. We can use either equation 1 or 2 from the original first-order conditions to find the optimal value of λ. Either way, we get $\lambda^* = 8\frac{1}{3}$.

For many optimization problems, we would be interested in knowing the numerical value of the maximum by evaluating the objective function (in

this case the utility function) at the optimal solution. But recall that utility is measured only up to an ordinal scale and the actual value of utility is irrelevant. We want to maximize utility, but we do not care about its actual maximum value. The fact that utility is ordinal, not cardinal, also explains why the optimal value of lambda is not meaningful. In general, the Lagrangean multiplier tells us how the maximum value of the objective function changes as the constraint is relaxed. With utility as the objective function, this interpretation is not applicable.

Numerical Approach

Instead of calculus (via the method of Lagrange) and pencil and paper, we can use numerical methods to find the optimal solution.

To use the numerical approach, we need to do some preliminary work. We have to set up the problem in Excel, carefully organizing things into a goal, endogenous variables, exogenous variables, and constraint. Once we have everything organized, we can use Excel's Solver to get the solution.

STEP Open the Excel workbook *OptimalChoice.xls*, read the *Intro* sheet, and then go to the *OptimalChoice* sheet to see how the numerical approach can be used to solve the problem we worked on above.

Figure 3.2 reproduces the display you see when you first arrive at the *OptimalChoice* sheet.

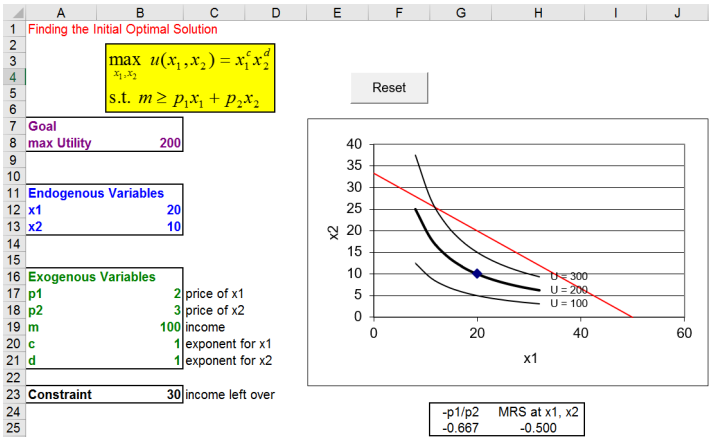


Figure 3.2: The initial display in the *OptimalChoice* sheet.
Source: OptimalChoice.xls!OptimalChoice

Notice how the sheet is organized according to the three components of the optimization problem: goal, endogenous, and exogenous variables. The constraint cell displays how much of the consumer's budget remains available for buying goods. The consumer in Figure 3.2 is not using all of the income available so we know satisfaction cannot be maximized at the point 20,10.

STEP Let's have the consumer buy x_2 with the remaining \$30. At \$3/unit, 10 additional units of x_2 can be purchased. Enter 20 in the x_2 cell (B13) and hit the Enter key. The chart refreshes to display the point 20,20, which is on the budget constraint, and draws three new indifference curves.

Although 20,20 does exhaust the available income, it is not the optimal solution. While you know the answer is $25,16\frac{2}{3}$, there is another way to tell that the consumer can do better.

STEP Look carefully at the display below the chart. It reveals the MRS does not equal the price ratio. This immediately tells us that something is amiss here.

MRS $> p_1/p_2$ tells us that the slope of the indifference curve at that point is greater than the slope of the budget constraint. The consumer cannot change the slope of the budget constraint, but the MRS can be altered by choosing a different the combination of goods. This consumer needs to lower the MRS (in absolute value) to make the two equal. This can be done by moving down the budget constraint.

If the consumer buys 10 more of good 1 (so 30 units of x_1 total), consumption of x_2 must fall by $6\frac{2}{3}$ units to $13\frac{1}{3}$.

STEP Enter 30 in cell B12 and the formula $= 13 + 1/3$ in B13. Now you are on the other side of the optimal solution. The MRS is less than the price ratio.

You could, of course, continue adjusting the cells manually, but there is a faster way.

STEP Click the Data tab in Excel's Ribbon (on the top of the screen) and click Solver (grouped under the Analyze tab) or execute Tools: Solver in older versions of Excel to bring up the Solver Parameters dialog box (displayed in Figure 3.3).

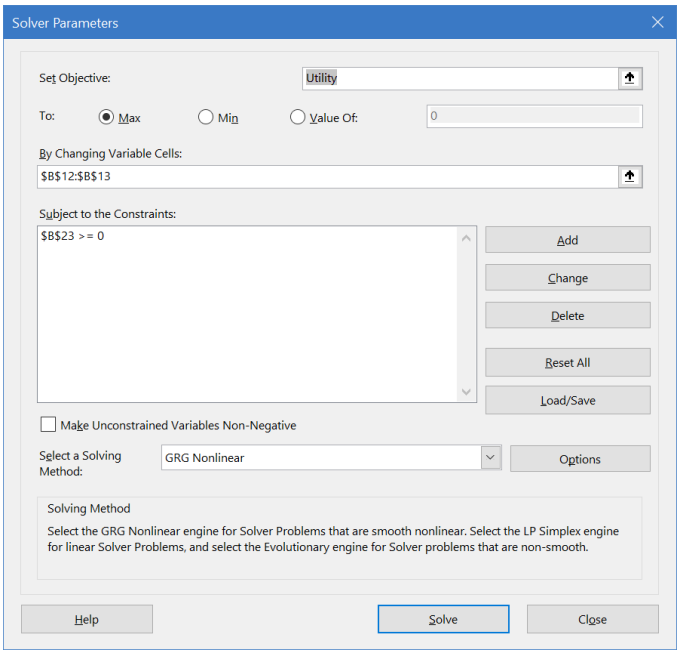


Figure 3.3: Excel's Solver interface.

If you do not have Solver available as a choice, bring up the Add-in Manager dialog box and make sure that Solver is listed and checked. If Solver is not listed, you must install it. Solver is included in a standard installation of Excel. For help, try support.office.com or www.solver.com.

Note how Excel's Solver includes information on the objective function (the target cell), the choice variables (the changing cells), and the budget constraint. These have all been filled in for you, but you will learn how to do this yourself in future work.

STEP Since all of the information has been entered into the Solver Parameters dialog box, simply click the Solve button at the bottom of the dialog box.

Excel's Solver works by trying different combinations of x_1 and x_2 and evaluating the improvement in the target cell, while trying to stay within the constraint. When it cannot improve very much more, it figures it has found the answer and displays a message as shown in Figure 3.4.

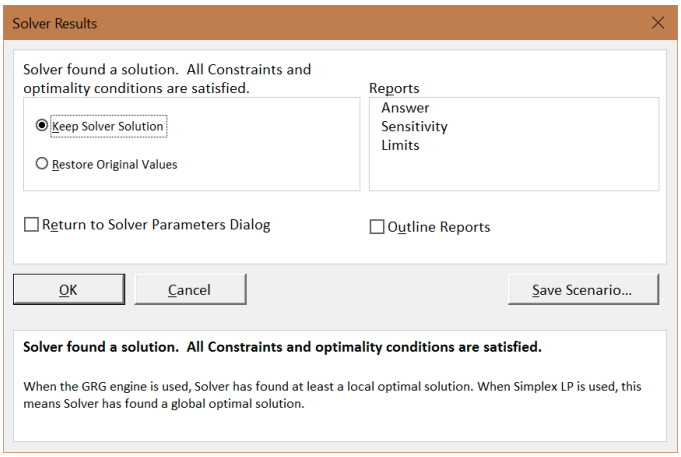


Figure 3.4: Solver reports success.

Although Solver gets the right answer in this problem, we will see in future applications that Solver is not perfect and does not deserve blind trust.

STEP Click the Sensitivity option under Reports and click OK; Excel puts the Solver solution into cells B12 and B13. It also inserts a new sheet into the workbook with the Sensitivity Report.

STEP Click on cells B12 and B13. Notice that Excel did not get exactly 25 and $16\frac{2}{3}$. It got extremely close and you can certainly interpret the result as confirming the analytical solution, but Solver's output requires interpretation and critical thinking by the user. We will focus on the issue of the exactly correct answer later.

STEP Proceed to the *Sensitivity Report* sheet (inserted by Solver) to confirm that this numerical method gives substantially the same absolute value for the Lagrangean multiplier that we found via the Lagrangean method ($8\frac{1}{3}$). We postpone explanation of this because utility's ordinal scale makes interpretation of the Lagrangean multiplier pointless. For now, we simply note that Solver can report optimal lambda and its results agreed with the Lagrangean method.

You might notice that Excel reports a Lagrangean multiplier value of -8.33 (with a few more trailing 3s) yet our analytical work did not produce a negative number. It turns out that we ignore the sign of λ^*. If we set up the Lagrangean as the objective function minus (instead of plus) lambda times the constraint or rewrite the constraint as $0 = 2x_1 + 3x_2 - 100$ (instead of

$0 = 100 - 2x_1 - 3x_2$), we would get a negative value for λ^* in our analytical work. The way we write the constraint or whether we add or subtract the constraint is arbitrary, so we ignore the sign of λ^*.

To be clear, unlike the sign, the magnitude of λ^* can be meaningful, but it is not in this application because utility is not cardinal. We will, however, see examples where the value of λ^* is useful and has an economic interpretation.

Using Analytical and Numerical Methods to Find the Optimal Solution

There are two ways to solve optimization problems:

1. The traditional way uses pencil and paper, derivatives, and algebra. The Lagrangean method is used to solve constrained optimization problems, such as the consumer's choice problem.

2. Advances in computers have led to the creation of numerical methods to solve optimization problems. Excel's Solver is an example of a numerical algorithm that can be used to find optimal solutions.

In the chapters that follow, we will continue to use both analytical and numerical approaches. You will see that neither method is perfect and both have strengths and weaknesses.

Exercises

The utility function, $U = 10x - 0.1x^2 + y$, has a quasilinear functional form. Use this utility function to answer the questions below.

1. Suppose the budget line is $100 = 2x + 3y$. Use the analytical method to find the optimal solution. Show your work.

2. Suppose the consumer considers the bundle 0,33.33, buying no x and spending all income on y. Use the MRS compared to the price ratio logic to explain what the consumer will do and why.

3. This utility function can be written in a more general form with letters instead of numbers, like this: $U = ax - bx^c + dy$. If a increases, what happens to the optimal consumption of x*? Explain how you arrived at your answer.

References

The epigraph is from page 421 of W. W. Rouse Ball's *A Short Account of the History of Mathematics* (first published in 1888). Of course, there are many books on the history of mathematics, but this classic is fun and easy to read. It mixes stories about people with real mathematical content.

This entire book (and many others) is freely available at books.google.com. You can read it online or download it as a pdf file.

Appendix: Derivatives and Optimization

A *derivative* is a mathematical expression that tells you how y in a function $y = f(x)$ changes given an infinitesimally small change in x. Graphically, it is the slope, or rate of change, of the function at that particular value of x.

Linear functions have a constant slope and, therefore, a constant value for the derivative. For the linear function $y = 6 + 3x$, the derivative of y with respect to x is written $\frac{dy}{dx}$ (pronounced "d y d x") and its value is 3. This tells you that every time the x variable goes up, the y variable goes up threefold. So, if x increases by 1 unit, y will increase by 3 units. This is easy to see in Figure 3.5.

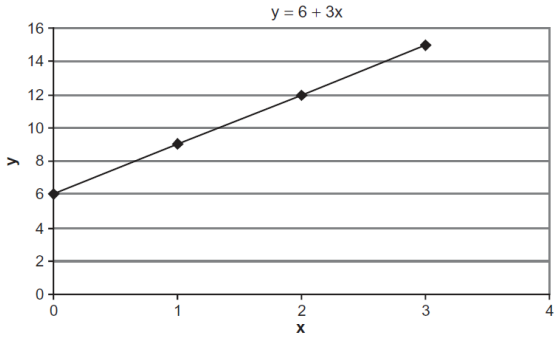


Figure 3.5: A linear function.

Nonlinear functions have a changing slope and, therefore, a derivative that takes on different values at different values of x. Consider the function $y = 4x - x^2$. Figure 3.6 graphs this function. Its derivative is $\frac{dy}{dx} = 4 - 2x$. When

evaluated at a specific point, such as $x = 1$, the derivative is the slope of the tangent line at that point.

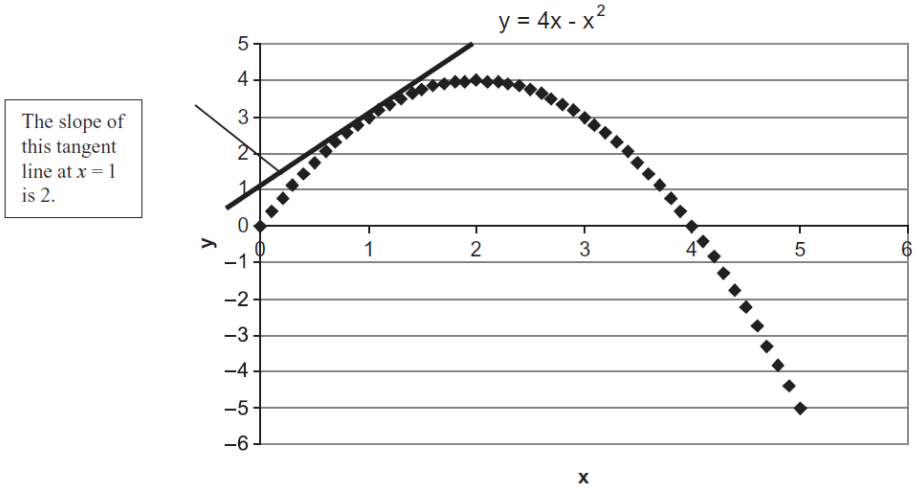


Figure 3.6: A nonlinear function with tangent line at $x = 1$.

Unlike the previous case, this derivative has x in it. This means this function is nonlinear. The slope depends on the value of x. At $x = 1$, the derivative is 2, but at $x = 2$, it is zero $(4 - 2[2])$ and at $x = 3$, it is -2 $(4 - 2[3])$.

In addition, because it is nonlinear, the size of the change in x affects the measured rate of change. For example, the change in y from $x = 1$ to $x = 2$ is 1 (because we move from $y = 3$ to $y = 4$ as we increase x by 1). If we increase x by a smaller amount, say 0.1 (from 1 to 1.1), then $\frac{\Delta y}{\Delta x} = \frac{3.19-3}{1.1-1} = 1.9$. By taking a smaller change in x, we get a different measure of the rate of change.

If we compute the rate of change via the derivative, by evaluating $4 - 2x$ at $x = 1$, we get 2. The derivative computes the rate of change for an infinitesimally small change in x. The smaller the change in x, the closer $\frac{\Delta y}{\Delta x}$ gets to $\frac{dy}{dx}$. You can see this happening as $\frac{\Delta y}{\Delta x}$ went from 1 to 1.9 as Δx fell from 1 to 0.1. If we go even smaller, making $\Delta x = 0.01$ (going from 1 to 1.01), then $\frac{\Delta y}{\Delta x} = \frac{3.0199-3}{1.01-1} = 1.99$.

Optimizing with the Derivative

An optimization problem typically requires you to find the value of an endogenous variable (or variables) that maximizes or minimizes a particular

objective function. We can use derivatives to find the optimal solution. This is called an analytical approach.

If we draw tangent lines at each value of x in Figure 3.6, only one would be horizontal (with derivative and slope of zero) and that would be the one at the top. This gives us a solution strategy: to find the maximum, find the value of x with the flat tangent line. This is equivalent to finding the value of x where the derivative is zero.

By solving for the value of x where $\frac{dy}{dx} = 0$, we find the optimal solution. For $y = 4x - x^2$, this is easy. We set the derivative equal to zero and solve for x^*.

$$\begin{aligned} \frac{dy}{dx} &= 4 - 2x^* = 0 \\ 4 &= 2x^* \\ x^* &= 2 \end{aligned}$$

The equation that you make when you set the first derivative equal to zero is called the *first-order condition*. The first-order condition is different from the derivative because the derivative by itself is not equal to anything—you can plug in any value of x and the derivative expression will pump out an answer that tells you whether and by how much the function is rising or falling at that point. The first-order condition is a special situation in which you are using the derivative to find a horizontal tangent line to figure out where the function has a flat spot.

A *reduced form* is the answer that you get when the derivative is set equal to zero and solved for the optimal solution. It may be a number or a function of exogenous variables. It cannot have any endogenous variables in the expression. Sometimes, you cannot solve explicitly for x^*. We say there is no closed form solution in these cases. The solution may exist (and numerical methods may be used to find it), but we cannot express the answer as an equation.

The second derivative is the derivative of the first derivative. It tells you the slope of the slope function. For example, if a function has a constant slope, we saw that its first derivative is a constant value (like 3 in the first example above). Then the second derivative is zero.

Second derivatives are useful in optimization for the following reason: when you find the value of the endogenous variable that makes the first derivative

equal to zero, the point that you have located could be either a maximum or a minimum. If you want to be sure which one you have found, you can check the second derivative. For $y = 4x - x^2$, the first derivative is $4 - 2x$ and the second derivative is, therefore, -2. Because the second derivative is negative, we know that our flat spot at $x = 2$ is a maximum and not a minimum.

In this book, we will not use second derivatives to check that our solutions are truly maxima or minima. Our functions will be (mostly) well behaved and we will focus on the economics of the problem, not the mathematics.

In summary, derivatives are used to measure the rate of change of a function based on a vanishingly small change in x. If we set a derivative equal to zero, we are trying to find an optimal solution by finding a value for x where the tangent line is flat. This solution strategy is based on the idea that a point where the tangent line is horizontal must mean that we are at the top of the function (or bottom, if we are minimizing).

Useful Math Facts

This appendix concludes with a short list of common rules for taking derivatives and working with exponents. The idea here is to sharpen your math skills so you can solve optimization problems analytically.

A derivative can be computed by directly applying the definition—i.e., taking the limit of the change in x as it approaches zero and determining the change in y. Fortunately, however, there is an easier way. Differentiation rules have been developed that make it much less tedious to take a derivative. Most calculus books have inside covers that are full of rules. Many students never grasp that these rules are actually shortcuts. Here is a short list, with special emphasis on those used in economics.

The derivative rules are followed by a few algebra rules relating to legal operations on exponents. We will use these rules often to find optimal solutions and reduce complicated expressions to simpler final answers.

Reading these equations is boring and tedious, but may save a lot of time and effort in the future (especially if your math is rusty). You should consider writing out the examples for a different number, say 6. So, instead of x^4, what is the derivative with respect to x for x^6?

Derivative Rules

Let x be the variable and a be a constant.

General Rule	Example of its Application
$\frac{d}{dx}(x) = 1$	
$\frac{d}{dx}(ax) = a$	$\frac{d}{dx}(4x) = 4$
$\frac{d}{dx}(a) = 0$	$\frac{d}{dx}(4) = 0$
$\frac{d}{dx}(x^a) = ax^{a-1}$	$\frac{d}{dx}(x^4) = 4x^3$
$\frac{d}{dx}(a \ln x) = \frac{a}{x}$	$\frac{d}{dx}(4 \ln x) = \frac{4}{x}$

When you take a derivative of a function with respect to a variable, you apply the rules to the different parts of the function. For example, if $y = 4x - x^2$, then you apply the $\frac{d}{dx}(ax) = a$ rule to $4x$, getting 4. You apply the $\frac{d}{dx}(x^a) = ax^{a-1}$ rule to $-x^2$ and get $-2x$. Thus, the derivative of y with respect to x is $\frac{dy}{dx} = 4 - 2x$.

There are other calculus rules, of course, such as the chain rule, but we will explain them when they are needed.

Laws of Exponents

General Rule	Example of its Application
$x^0 = 1$	
$x^{-a} = \frac{1}{x^a}$	$x^{-\frac{1}{2}} = \frac{1}{\sqrt{x}}$
$x^a x^b = x^{a+b}$	$x^2x^3 = x^5 \Rightarrow 2^2 2^3 = 2^5 = 32$
$\frac{x^a}{x^b} = x^{a-b}$	$\frac{x^5}{x^3} = x^2 \Rightarrow \frac{2^5}{2^3} = 2^2 = 4$
$(xy)^a = x^a y^a$	$(xy)^2 = x^2y^2 \Rightarrow (2 \cdot 3)^2 = 2^2 3^2 = 36$
$(x^a)^b = x^{ab}$	$(x^2)^3 = x^6 \Rightarrow (2^2)^3 = 2^6 = 64$

> The methods of mathematics apply as soon as spatial or numerical attributes are associated with our phenomena, as soon as objects can be located by points in space and events described by properties capable of indication or measurement in numbers.
>
> R. G. D. Allen

3.2 More Practice and Understanding Solver

We know there are two approaches to solving optimization problems.

1. Analytical methods using algebra and calculus (conventional, paper and pencil, using the Lagrangean method): The idea is to transform the consumer's constrained optimization problem into an unconstrained problem and then solve it using standard unconstrained calculus techniques—i.e., take derivatives, set equal to zero, and solve the system of equations.

2. Numerical methods using a computer (Excel's Solver): Set up the problem in Excel, carefully organizing things into a goal, endogenous variables, exogenous variables, and constraint; then use Excel's Solver. Use the Sensitivity Report in the Solver Results dialog box to get λ^*.

In this chapter, we apply both methods on a new problem.

Quasilinear Utility Practice Problem

A utility function that is composed of a nonlinear function of one good plus a linear function of the other good is called a quasilinear functional form. It is *quasi*, or sort of, linear because one good increases utility in a linear fashion and the other does not.

Below are a general example and a more specific example of quasilinear utility.

$$u(x_1, x_2) = v(x_1) + x_2$$
$$u(x_1, x_2) = (x_1)^c + x_2, \text{where } c < 1$$

If $c < 1$, then the quasilinear utility function says that utility increases at a decreasing rate as x_1 increases, but utility increases at a constant rate as x_2 increases.

The optimization problem is to maximize this utility function subject to the usual budget constraint. It is written in equation form like this:

$$\max_{x_1,x_2,\lambda} x_1^c + x_2$$
$$\text{s.t. } p_1x_1 + p_2x_2 = m$$

We will solve the general version of this problem, with letters representing exogenous variables instead of numbers, using the Lagrangean method.

1. Rewrite the constraint so that it is equal to zero.

$$0 = m - p_1x_1 - p_2x_2$$

2. Form the Lagrangean function.

$$\max_{x_1,x_2,\lambda} L = x_1^c + x_2 + \lambda(m - p_1x_1 - p_2x_2)$$

Note that the Lagrangean function, L, has the quasilinear utility function plus the Lagrangean multiplier, λ, times the rewritten constraint.

Unlike the concrete problem in the previous chapter, which used numerical values, this is a general problem with letters indicating exogenous variables. General problems, without numerical values for exogenous variables, are harder to solve because we have to keep track of many variables and make sure we understand which ones are endogenous versus exogenous. If the solution can be written as a function of the exogenous variables, however, it is often easy to see how an exogenous variable will affect the optimal solution.

3. Take partial derivatives with respect to x_1, x_2, and λ.

$$\frac{\partial L}{\partial x_1} = cx_1^{c-1} - p_1\lambda$$
$$\frac{\partial L}{\partial x_2} = 1 - p_2\lambda$$
$$\frac{\partial L}{\partial \lambda} = m - p_1x_1 - p_2x_2$$

Remember that the partial derivative treats other variables as constants. Thus, the partial derivative of the quasilinear utility function with respect to x_1 has no x_2 variable in it.

4. Set the derivatives equal to zero and solve for x_1*, x_2*, and λ*.

$$\frac{\partial L}{\partial x_1} = cx_1^{c-1} - p_1\lambda = 0$$
$$\frac{\partial L}{\partial x_2} = 1 - p_2\lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = m - p_1x_1 - p_2x_2 = 0$$

We use the same solution method as before, moving the lambda terms to the right-hand side and then dividing the first equation by the second, which allows us to cancel the lambda terms.

$$cx_1^{c-1} = p_1\lambda$$
$$1 = p_2\lambda$$
$$\frac{cx_1^{c-1}}{1} = \frac{p_1\lambda}{p_2\lambda}$$
$$\frac{cx_1^{c-1}}{1} = \frac{p_1}{p_2}$$

By canceling the lambda terms, we have reduced the three equation, three unknown system to two equations with two unknowns.

$$\frac{cx_1^{c-1}}{1} = \frac{p_1}{p_2}$$
$$m - p_1x_1 - p_2x_2 = 0$$

Remember that not all variables are the same. The endogenous variables, the unknowns, are x_1 and x_2. The other letters are exogenous variables.

From the first equation, we can solve for the optimal quantity of good 1 (see the appendix to the previous section if these steps are confusing).

$$\begin{aligned}
\frac{cx_1^{c-1}}{1} &= \frac{p_1}{p_2} \\
cx_1^{c-1} &= \frac{p_1}{p_2} \\
x_1^{c-1} &= \frac{p_1}{cp_2} \\
x_1^* &= \left(\frac{p_1}{cp_2}\right)^{\frac{1}{c-1}}
\end{aligned}$$

Notice that we used the rule that $(x^a)^b = x^{ab}$. Because we wanted to solve for x_1, we raised both sides to the $\frac{1}{c-1}$ power so that the $c-1$ exponent on x_1 times $\frac{1}{c-1}$ would equal 1.

Usually, when we have the MRS equal to the price ratio, we need to solve for one of the x variables in terms of the other and substitute it into the budget constraint. However, a property of the quasilinear utility function is that the MRS only depends on x_1; thus by solving for x_1, we get the reduced form solution. When solving a problem in general terms, the answer must be expressed as a function of exogenous variables alone (no endogenous variables) and this is called a reduced form.

To get x_2, we simply substitute x_1 into the budget constraint and solve for x_2.

$$\begin{aligned}
m - p_1\left[\left(\frac{p_1}{cp_2}\right)^{\frac{1}{c-1}}\right] - p_2x_2 &= 0 \\
x_2^* &= \frac{m}{p_2} - \frac{p_1}{p_2}\left(\frac{p_1}{cp_2}\right)^{\frac{1}{c-1}}
\end{aligned}$$

It is a bit messy, but it is the answer. We have an expression for the optimal amount of x_2 that is a function of exogenous variables alone.

To get the optimal value of lambda, we can use the second first-order condition, which simply says that $\lambda^* = \frac{1}{p_2}$. If you use the first condition, substituting in the value for optimal x_1, it will take a little work, but you will get the same result.

Practice with the MRS = $\frac{p_1}{p_2}$ Logic

Economists stress marginal thinking. The idea is that, from any position, you can move and see how things change. If there is improvement, continue moving. The optimal solution is on a flat spot, where improvement is impossible.

When we move the lambda terms over to the right-hand side and divide the first equation by the second equation, we get a crucial statement of the fact that improvement is impossible and we are optimizing.

The familiar MRS equals the price ratio expression, along with the third first-order condition, which says that the consumer must be on the budget line (exhausting all income), is a mathematical way of describing marginal thinking.

The MRS condition tells us that if the MRS is not equal to the price ratio, there are two possibilities, depicted in Figure 3.7.

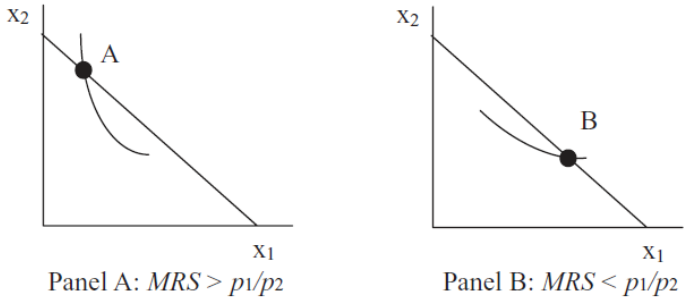


Figure 3.7: MRS does not equal the price ratio.

In Panel A, the slope of the indifference curve at point A is greater than the slope of the budget line (in absolute value). This consumer should crawl down the budget line, reaching higher indifference curves, until the MRS equals the price ratio. At this point, the slope of the indifference curve will exactly equal the slope of the budget line and the consumer's indifference curve will just touch the budget line. The consumer cannot possibly get to a higher indifference curve and stay on the budget constraint. This is the best possible solution.

In Panel B, the story is the same, but reversed. The slope of the indifference curve at point B is less than the slope of the budget line. This consumer

should crawl up the budget line, reaching higher indifference curves, until the MRS equals the price ratio. At this point, the slope of the indifference curve will exactly equal the slope of the budget line and the consumer's indifference curve will just touch the budget line.

Numerical Approach to Quasilinear Practice Problem

STEP Open the Excel workbook *OptimalChoicePractice.xls*, read the *Intro* sheet, and then go to the *QuasilinearChoice* sheet to see how the numerical approach can be used to solve this problem.

It is easy to see that the consumer cannot afford the bundle 5,20 given the prices and income on the sheet. If she buys five units of x_1, what's the maximum x_2 she can buy?

STEP Enter this amount in cell B12. Does the chart and cell B21 confirm that you got it right?

If you entered 13 in B12, then the chart updates and shows that the consumer is now on the budget line. In addition, the constraint cell, B21, is now zero.

Without running Solver or doing any calculations at all, is she maximizing at 5,13?

The answer is that she is not. It's hard to see on the chart whether the indifference curve is cutting the budget line, but the information below the chart shows that the MRS is not equal to the price ratio. That tells you that the indifference curve is, in fact, not tangent to the budget line so the consumer is not optimizing. Because the MRS is greater than the price ratio (in absolute value) we also know that the consumer should buy more x_1 and less x_2, moving down the budget line until the marginal condition is satisfied. Let's find the optimal solution.

STEP Run Solver. Select the Sensitivity Report to get λ^*.

How does Excel's answer compare to our analytical answer? Recall that we found:

$$x_1^* = \left(\frac{p_1}{cp_2}\right)^{\frac{1}{c-1}}$$
$$x_2^* = \frac{m}{p_2} - \frac{p_1}{p_2}\left(\frac{p_1}{cp_2}\right)^{\frac{1}{c-1}}$$

STEP Create formulas in Excel to compute these two solutions (using cells C11 and C12 would make sense). This requires some care with the parentheses. Here is the formula for good 1: =(p1_/(c_*p2_))^(1/(c_-1)).

You should discover that Excel's Solver is quite close to the exactly correct solution, 6.25, 12.75. We conclude that the two methods, analytical and numerical, substantially agree.

It is true, however, that Solver is ever so slightly off the computed analytical result. In general, there are two reasons for minuscule disagreement between the two methods.

1. Excel cannot display the algebraic result to an infinite number of decimal places. If the solution is a repeating decimal or irrational number, Excel cannot handle it. Even if the number can be expressed as a decimal—for example, one-half is 0.5—precision error may occur during the computation of the final answer. This is not the source of the discrepancy in this case.

2. Excel's Solver often misses the exactly correct answer by small amounts. Solver has a convergence criterion (that you can set via the Options button in the Solver Parameters dialog box) that determines when it stops hunting for a better answer. Figure 3.8 offers a graphical representation of Solver's algorithm in a one-variable case.

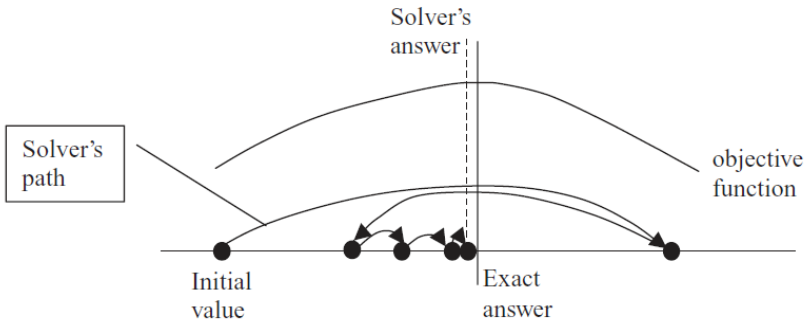


Figure 3.8: Solver in action.

The *stylized graph* (which means it represents an idea without using actual data) in Figure 3.8 shows that Solver works by trying different values and seeing how much improvement occurs. The path of the choice variable (on the x axis) is determined by Solver's internal optimization algorithm. By default, it uses Newton's method (a steepest descent algorithm), but you can choose an alternative by clicking the Options button in the Solver dialog box.

When Solver takes a step that improves the value of the objective function by very little, determined by the convergence criterion (adjustable via the Options button), it stops searching and announces success. In Figure 3.8, Solver is missing the optimal solution by a little bit because, if we zoomed in, the objective function would be almost flat at the top. Solver cannot distinguish additional improvement.

When we say that the analytical method agrees with Solver, we do not mean that the two methods exactly agree, but simply that they correspond, in a practical sense. If Solver is off the exact answer in the 15th decimal place, that is agreement, for all practical purposes.

Furthermore, it is easy to conclude that Solver must give an exact answer because it displays so many decimal places. This is incorrect. Solver's display is an example of *false precision*. It is not true that the many digits provide useful information. The exact answer is 6.25 and 12.75. What you are seeing is Solver noise. You must learn to interpret Solver's results as inexact and not report all of the decimal places.

There is another way in which Solver can fail us and it is much more serious than incorrectly interpreting the results.

Solver Behaving Badly

STEP Start from $x_1 = 1, x_2 = 20$ to see a demonstration that Solver is not perfect. After setting cells B11 and B12 to 1 and 20, respectively, run Solver. What happens?

A *miserable result* (an actual, technical term in the numerical methods literature) occurs when an algorithm reports that it cannot find the answer or displays an obviously erroneous solution. Figure 3.9 displays an example of a miserable result. Solver is clearly announcing that it cannot find an answer.

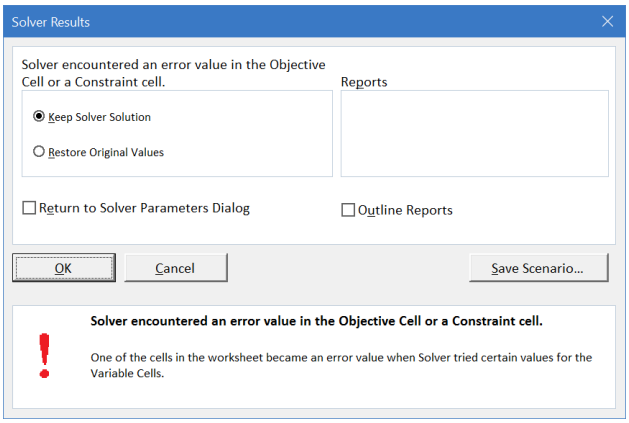


Figure 3.9: A miserable result.

If you look carefully at the spreadsheet (click cancel or OK if needed to return to the sheet), you will see that Solver blew up when it tried a negative value for x_1. The objective function cell, B7, is displaying the error #NUM! because Excel cannot take the square root of a negative number.

To be clear, when we start from 1,20, Excel tries to move left and crosses over the y axis into negative x territory. Since the utility function is $x_1^{0.5}$, it tries to take the square root of a negative number, producing an error, and crashing the algorithm.

When Solver fails, there are three basic strategies to fix the problem:

1. Try different initial values (in the changing cells). If you know roughly where the solution lies, start near it. Always avoid starting from zero or a blank cell.

2. Add more structure to the problem. Include non-negativity constraints on the endogenous variables, if appropriate. In the case of consumer theory, if you know the buyer cannot buy negative amounts, add this information.

3. Completely reorganize the problem. Instead of directly optimizing, you can put Solver to work on equations that must be met. In this problem, you know that MRS $= \frac{p_1}{p_2}$ is required. You could create a cell that is the difference between the MRS and the price ratio and have Solver find the values of the choice variable that force this cell to equal zero.

Let's try the second strategy.

STEP Reset the initial values to 1 and 20, then launch Solver (click the Data tab and click Solver) and click the Add button (at the top of the stacked buttons on the right).

Solver responds by popping up the Add Constraint dialog box.

STEP Select both of the endogenous variables in the Cell Reference field, select >=, and enter 0 in the Constraint field so that the dialog box looks like Figure 3.10. Click OK.

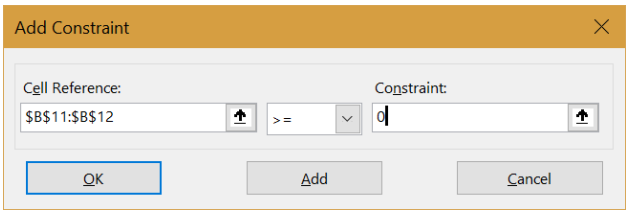


Figure 3.10: A miserable result.

You are returned to the main Solver Parameters dialog box, but you have added the constraint that cells B11 and B12 must be non-negative.

You might notice that you could have have simply clicked the *Make Unconstrained Variables Non-Negative* option, but adding the constraint shows how to work with constraints.

STEP Once back at the main Solver Parameters dialog box, click Solve.

This time, Solver succeeds. Adding the non-negativity constraint prevented Solver from trying negative x_1 values and producing an error.

Perfect Complements Practice Problem

Recall that L-shaped indifference curves represent perfect complements, which are reflected via the following mathematical function:

$$u(x_1, x_2) = minax_1, bx_2$$

Suppose $a = b = 1$ and the budget line is $50 = 2x_1 + 10x_2$.

First, We want to solve this problem analytically.

The Lagrangean method cannot be applied because the function is not differentiable at the corner of the L. The Lagrangean method, however, is not the only analytical method available. Figure 3.11 shows that when a = b = 1, the optimal solution must lie on a ray from the origin with slope +1.

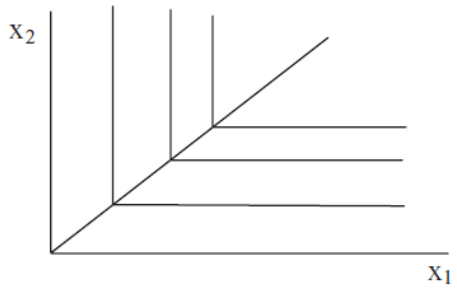


Figure 3.11: The optimal solution line with perfect complements.

The optimal solution has to be on the corner of the L-shaped indifference curves because a non-corner point (on either the vertical or horizontal part of the indifference curve) implies the consumer is spending money on more of one of the goods without getting any additional satisfaction. Thus, we know that the optimal solution must lie on the line $x_2 = x_1$.

We can combine this optimal solution equation with the budget constraint to find the optimal solution. The two equation, two unknown system can be solved easily by substitution.

$$\left.\begin{array}{l} x_2 = x_1 \\ 50 = 2x_1 + 10x_2 \end{array}\right\} \Rightarrow 50 = 2x_1 + 10\,[x_1] \Rightarrow 50 = 12x_1 \Rightarrow x_1^* = 4\frac{1}{6}.$$

Of course, we know $x_2 = x_1$ so optimal x_2 is also $4\frac{1}{6}$. Can Excel do this problem and do we get the same answer? Let's find out.

STEP Proceed to the *PerfectComplements* sheet to see how we set up the spreadsheet in Excel. Click on cell B7 to see the utility function.

STEP Run Solver and get a Sensitivity Report. Solver can be used to generate a value for the Lagrangean multiplier (via the Sensitivity Report) even though we could not use the Lagrangean method in our analytical work.

As with the previous problem (with quasilinear utility), we find that Solver and the analytical approach substantially agree. The answer is a repeating decimal, so Excel cannot get the exact answer, $4\frac{1}{6}$, but it is really close.

Previously, we saw that Solver could crash and give a miserable result. Now, let's learn that Solver can really misbehave.

STEP Starting from $x_1 = 1, x_2 = 1$, run Solver. What happens?

You are seeing an example of a *disastrous result* which occurs when an algorithm reports that it has found the answer, but it is wrong. There is no obvious error and the user may well accept the answer as true.

Solver reports a successful outcome, but the answer it gives is 1,1 and we know the right answer is $4\frac{1}{6}$ for both goods.

Disastrous results include an element of interpretation. In this case, we might notice that 1,1 is way inside the budget constraint and, therefore, the algorithm has failed. A truly disastrous result occurs when there is no way to independently test or verify the algorithm's wrong answer.

Miserable and disastrous results are well defined, technical terms in the mathematical literature on numerical methods. Disastrous results are much more dangerous than miserable results. The latter are frustrating because the computer cannot provide an answer, but disastrous results lead the user to believe an answer that is actually wrong. In the world of numerical optimization, they are a fact of life. Numerical methods are not perfect. You should never completely trust any optimization algorithm.

Understanding Solver—Be Skeptical

This chapter enabled practice solving the consumer's constrained optimization problem with two different utility functions, a quasilinear function and perfect complements. In both cases, we found that Excel's Solver agreed, practically speaking, with the analytical method.

The ability to solve optimization problems with two independent methods means we can be really sure we have found an optimal solution when they give the same answers.

In addition, we explored how Solver actually works. It evaluates the objective function for different values of the choice variables. It continues searching for a better solution until it cannot improve much (an amount determined by the convergence criterion).

Solver can fail by reporting that it cannot find a solution (called a miserable result) or—even worse—by reporting an incorrect answer with no obvious error (which is a disastrous result).

It is easy to believe that a result displayed by a computer is guaranteed to be correct. Do not be careless and trusting—numerical methods can and do fail, sometimes spectacularly.

This point deserves careful repetition. You run Solver and it happily announces that a solution has been found and offers up a 15 or 16 digit number for your inspection. The problem, however, is that the solution is *way off*. Not in the millionth or even tenth decimal place, but completely, totally wrong. How this might happen takes us too far afield into the land of numerical optimization, but suffice it to say that you should always ask yourself if the answer makes common sense.

Solver really is a powerful way to solve optimization problems, but it is not perfect. You need to always remember this. After running Solver, format the results with an eye toward ease of understanding and think about the result itself. Do not mindlessly accept a Solver result. Stay alert even if Solver claims to have hit pay dirt—it may be a disastrous result!

More explanation of Solver is available in the *SolverInstructions.doc* file in the *SolverCompStaticsWizard* folder.

Exercises

1. In the quasilinear example in this chapter, use the first equation in the first-order conditions to find λ^*. Show your work.

2. Use analytical methods to find the optimal solution for the same perfect complements problem as presented in this chapter, except that $a = 4$ and $b = 1$. Show your work.

3. Draw a graph (using Word's Drawing Tools) of the optimal solution for the previous question.

4. Use Excel's Solver to confirm that you have the correct answer. Take a picture of the cells that contain your goal, endogenous variables, and exogenous variables.

References

As economics became more mathematical, a new course was born, Math Econ. The course needed books and R. G. D. Allen's *Mathematical Analysis for Economists* (first published in 1938) became a classic textbook. As E. Schneider, a reviewer, said, "This book fills a long-felt want. At last we possess a book which presents the mathematical apparatus necessary to a serious study of economics in a form suited to the needs of the economist." See *The Economic Journal*, Vol. 48, No. 191 (September, 1938), p. 515. The epigraph is from page 2 of *Mathematical Analysis for Economists*, as Allen discusses how and why mathematics can be applied to the study of economics.

> Tastes are the unchallengeable axioms of a man's behavior; he may properly (usefully) be criticized for inefficiency in satisfying his desires, but the desires themselves are *data*.
>
> George Stigler and Gary Becker

3.3 Food Stamps

This chapter applies the consumer choice model to a real-world example. We will see that the model can be used to explain why someone would illegally sell food stamps. We also tackle an important policy question: If cash dominates food stamps, why not just help low-income people by giving them cash?

A Short History of Food Assistance in the United States

The primary responsibility for ensuring poor people (including children) in the United States have enough to eat lies with the Department of Agriculture (USDA). They run a program that enables low-income people to spend government-provided benefits on eligible food in stores.

The USDA's web page, (www.fns.usda.gov/snap/short-history-snap), is the source of the information below. The Data and Research tab on the USDA's website has usage and cost data—there are around 40 million participants and the program spends roughly $70 billion per year. This is one of the largest transfer programs in the fight against poverty. It offers critical support for low-income households.

The first Food Stamp Program, in 1939, was very different from today's version. Originally, "the program operated by permitting people on relief to buy orange stamps equal to their normal food expenditures. For every $1 worth of orange stamps purchased, 50 cents worth of blue stamps were received. Orange stamps could be used to buy any food. Blue stamps could only be used to buy food determined by the Department to be surplus."

Important changes were made in the 1960s and, in 1977, the purchase requirement was eliminated. Households below the poverty line who met other criteria (such as work or study requirements) were eligible to receive food stamps. Figure 3.12 shows that these stamps were like paper currency; they

were rectangular, but only about half the size of a dollar bill. There were different dollar denominations in a booklet. When buying food at the supermarket, the consumer tore out the stamp and paid for the food. They would pay for any non-food items with cash or a check.

Figure 3.12: Old US food stamps.
Source: Public domain file photo.

In 2008, it was renamed the Supplemental Nutrition Assistance Program (SNAP) to avoid stigma. It could be embarrassing to pay with food stamps since everyone in line immediately knew that you were receiving government assistance. Today, both names, food stamps and SNAP, are used.

SNAP has always been battered by politics, with benefits expanding and contracting depending on the rhetoric of the day. There are the usual arguments over administrative costs, but cheating on the part of recipients has been an especially contentious issue. In 2002, all states were required to use Electronic Benefits Transfer (EBT) cards. This was supposed to stop the illegal sale of food stamps (and reduce stigma), but fraud remains a focus of critics.

We can model and analyze food stamps with the Theory of Consumer Behavior. We will focus on how food stamps can be incorporated into the consumer's optimization problem and why selling food stamps is so difficult to stop.

Food Stamp Theory

Recall from the Budget Constraint chapter that food stamps are a subsidy that produces a budget constraint with a horizontal segment, as shown in Figure 3.13. We use the x_1 variable on the x axis to represent units of food. The x_2 variable on the y axis captures all other goods lumped together. We get the flat part of the constraint because food stamps can be used to buy only food.

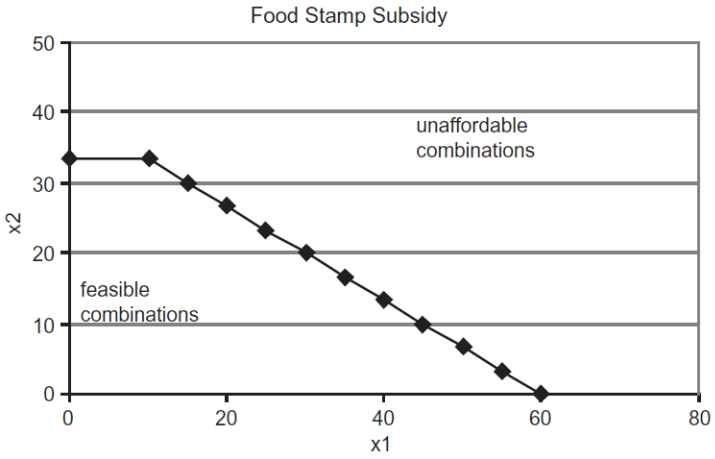


Figure 3.13: The budget constraint with food stamps.
Source: FoodStamps.xls!BudgetConstraint

STEP Open the Excel workbook *FoodStamp.xls* and read the *Intro* sheet. Proceed to the *BudgetConstraint* sheet. Change cell E13 from 10 to 20.

Notice that the horizontal segment, which is the monetary value of the food stamps divided by the price of food, gets longer. Also notice that the chart on the right, showing the budget constraint if the food stamp amount was treated as cash, has no horizontal segment. In the chart on the right, the value of the food stamp subsidy is computed (*xbar* times price of food) and then added to income as if it were cash; hence the name, cash-equivalent subsidy.

It should be quite clear that the cash-equivalent subsidy provides consumption possibilities that are unattainable above the horizontal segment of the food stamp budget constraint. The most other goods the food stamp recipient can buy is $33\frac{1}{3}$ units, while the cash-equivalent consumer can buy 40 units of x_2.

STEP Proceed to the *Inframarginal* sheet. It combines a food stamp budget constraint with a Cobb-Douglas utility function.

The word *inframarginal* (or submarginal) means below the edge or margin. The edge in this case is the kink in the budget constraint.

This consumer is *inframarginal* because his optimal solution is on the downward sloping part of the budget line, below the kink. He will use up his food stamp allotment on food and then spend some of his cash income to get additional food. The sheet reveals that he buys 35 units of food (valued at \$70, as shown in cell B15), 20 of which he obtains with food stamps and the remaining 15 he buys with cash.

We can easily see that he is optimizing because the "MRS equals the price ratio" condition is met. This is reflected in the graph where the highest attainable indifference curve is just touching the budget constraint.

STEP Click on cell B25 to see the formula for the budget constraint.

This formula is using an IF statement to implement the constraint in Excel. Expressed as an equation, the budget line looks like this:

$$\begin{aligned} &\text{if } x_1 \leq \bar{x}, \quad x_2 = m/p_2 \\ &\text{if } x_1 > \bar{x}, \quad x_2 = m/p_2 - p_1/p_2\,(x_1 - \bar{x}) \end{aligned}$$

The first equation says that if the consumer buys an amount of food that is less than or equal to *xbar*, that frees up his whole cash income to spend on good 2. This is the horizontal line component.

Things are more complicated if the consumer wants more than *xbar* of food. The second equation says that the consumer will have to use cash to buy amounts of x_1 greater than *xbar* and it computes the amount of x_2 that can be purchased as a function of x_1.

This constraint (rewritten to equal zero) has been entered in a single cell with an IF statement:

=IF(x1_<x1bar,m/p2_-x2_,m/p2_-(p1_/p2_)*(x1_-x1bar)-x2_)

The underscore (_) character is used in the variable names to distinguish them from cell addresses—e.g., p2_ is not cell P2.

From Excel's Help on the IF function:

> Returns one value if a condition you specify evaluates to TRUE and another value if it evaluates to FALSE.
>
> Use IF to conduct conditional tests on values and formulas.
>
> Syntax: IF(logical test,value if true,value if false)

Applying this information to the formula in cell B25, we can see that it has three parts, separated by commas. The first part says that if $x_1 < x1bar$ (that is the condition being evaluated), then the consumer can buy m/p_2 amount of x_2 (this second part produces the horizontal line in the budget constraint), else (the third part is what happens if x_1 is not less than $x1bar$) the consumer can buy x_2 along the downward sloping part of the budget line.

This problem shows that Excel can be used to handle complicated examples in the Theory of Consumer Behavior. This food stamp problem has a kinked budget constraint, but using Excel's IF statement allows us to implement the constraint in the workbook and use Solver to find the optimal solution.

This problem also can be solved via analytical methods, but it is cumbersome and difficult to deal with the kinked budget constraint. We will use the easier numerical approach to conduct our analysis.

STEP Proceed to the *Distorted* sheet.

This sheet is exactly the same as the *Inframarginal* sheet with one crucial exception: the preferences, in cells B21 and B22, are different. The consumer in the *Distorted* sheet prefers other goods more and food less than the consumer in the *Inframarginal* sheet.

The change in exponents in the Cobb-Douglas utility function has affected the indifference map. The curves are much flatter in the *Distorted* sheet compared with the *Inframarginal* sheet.

The *Distorted* sheet opens with the optimal values for food and other goods from the *Inframarginal* sheet. It is obvious that the MRS does not equal the price ratio and the indifference curve is cutting the budget constraint at the current bundle of x_1 and x_2. This consumer is not optimizing at this point.

Corner Solution

STEP Run Solver on the *Distorted* sheet.

Solver announces it has found the optimal solution, yet the MRS still does not equal the price ratio. Is this really the optimal solution? Yes, it is the optimal solution. We have encountered what is called a *corner solution* (or boundary optimum). In this case, the equimarginal condition, MRS $= \frac{p_1}{p_2}$, does not hold because the optimal solution is found at one of the end points (or corners) of the constraint.

STEP To see what is happening here, copy the optimal solution from the *Inframarginal* sheet (copy cells B13 and B14) and paste in the *Distorted* sheet (select cells B13 and B14 and then paste).

The graph and MRS is immediately updated and you can see that the distorted consumer would not select the inframarginal consumer's bundle. Which way should this consumer move—up or down the budget line? The graph makes clear that up is the right way to go, but you should notice that the marginal condition, MRS $< \frac{p_1}{p_2}$, tells you the same thing.

STEP Click the Crawl Up the Budget Line button. Click a few more times and pay attention to the chart and the MRS in cell H26. Also keep an eye on utility in cell B9. Each click lowers the amount of x_1 by one unit and increases the amount of x_2 by $\frac{2}{3}$.

By moving up the budget line, this consumer is improving her satisfaction and closing the gap between the MRS and the price ratio.

Do not be misled by the display – the indifference curves are not shifting. Remember that the indifference map is dense, meaning that every point has an indifference curve through it. We cannot draw in all of the indifference curves because the graph would then be solid black. The consumer is simply moving from one indifference curve to another one that was not previously displayed.

STEP Keep clicking the Crawl Up the Budget Line button. Eventually, you will hit the kink in the budget line and you will not be able to move northwest any longer. Instead, you will be on the horizontal segment and as you move strictly west, utility falls. Notice that the price ratio is now showing zero.

On the flat part of the budget line, when the amount of food purchased is less than or equal to how much food can be bought with food stamps alone, it makes sense that additional food is free, in terms of spending cash on food. The consumer simply has to use the available food stamps to acquire food and this does not reduce cash income.

Once you are on the flat part of the budget line, you should see that the graph and marginal condition point you to choosing more food.

STEP Click on the Crawl Down the Budget Line button repeatedly to move east and, eventually, down the budget line. Use the two buttons to crawl up and down until you find the bundle that maximizes utility.

You should end your travels at the kink – and MRS does not equal the price ratio there! This happens because the complicated constraint is producing a corner solution.

The distorted consumer wishes she could continue crawling up the downward sloping line, consuming less than the food stamp allotment of food and more of other goods, but she cannot do this. She cannot use food stamps to buy other goods. Thus, her best, or optimal, solution is at the kink.

In a corner solution, we accept that the "MRS equals the price ratio" condition is not met. We really are maximizing even though the MRS does not equal the price ratio. We have found the best we can do given the constraints on our choices.

Another way to explain what is happening is that we always want to minimize $|MRS - \frac{p_1}{p_2}|$. With an interior solution, we can make this difference zero, but with a corner solution, we cannot because a constraint is preventing us from reaching MRS $= \frac{p_1}{p_2}$. However, a corner solution does give us the lowest $|MRS - \frac{p_1}{p_2}|$ value and we are doing the best we can at this solution.

Corner solutions are an important concept and we will see them again in future work. They arise whenever we are prevented from continuing to improve by going in a particular direction.

Cash Instead of Food Stamps

STEP Proceed to the *Cash* sheet. Notice that cell B24 computes the cash value of the food stamps and that the chart has a linear budget constraint with no kink. Click cell B25 to see that the constraint is the familiar income minus expenditures, with income equal to the sum of income plus the cash value of the food stamps.

The idea here is that instead of giving food stamps, we provide low-income people the cash-equivalent value. They are no longer constrained to buy food alone, but can purchase any goods with the cash received. The cash subsidy shifts the budget line out, with no kink or horizontal segment like we saw with the food stamp program.

The sheet opens with the inframarginal consumer's optimal solution. It is the same as before, when she was given food stamps. Cash or food stamps are the same to this consumer.

STEP Click on the Set to Distorted button to quickly apply the preferences for the distorted consumer. Run Solver.

With cash, the distorted consumer chooses an optimal bundle that is different from the one chosen under the Food Stamp Program. She finds an interior (as opposed to a corner) solution in the far northwest corner, which means she has opted for little food and more of other goods.

Figure 3.14 summarizes our work to this point. If you compare the inframarginal consumer, by looking top left and then bottom left, in Figure 3.14, you can easily see that there is no change in his behavior: $40 in food stamps versus $40 in cash are the same to this consumer.

On the other hand, comparing the top right and bottom right panels in Figure 3,14 reveals that the distorted consumer chooses less food and more other goods when given cash. This is why we say her choices are *distorted* by the food stamp program. If she had cash, she would make different choices. The distortion results in a decrease in satisfaction for this consumer.

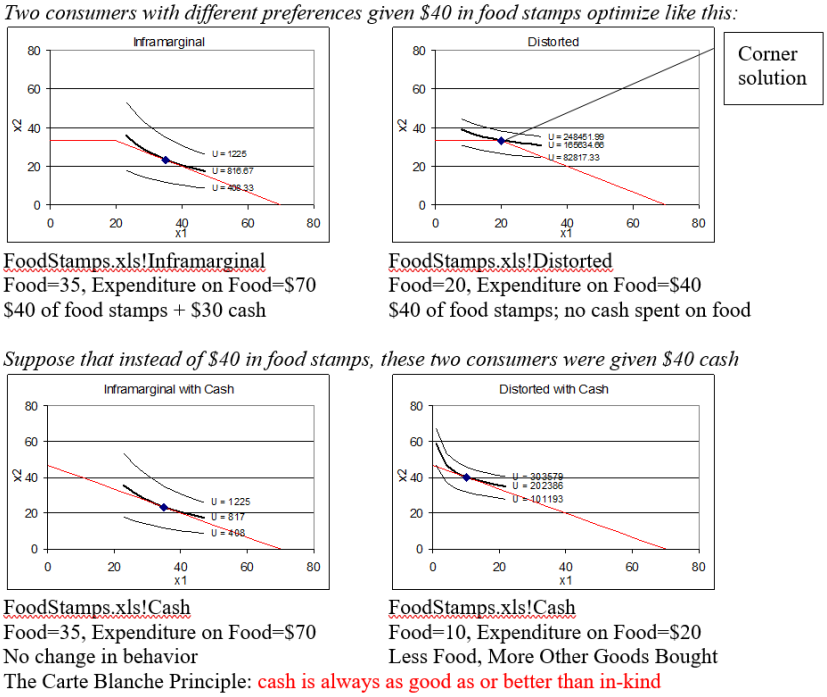


Figure 3.14: Comparing food stamps versus cash-equivalent.

The Carte Blanche Principle and Deadweight Loss

Carte blanche, a term of obvious French origin (literally, "blank document"), means unconditional authority or freedom to act in any way you wish.

In economics, the *Carte Blanche Principle* means that cash is always as good as or better than in-kind. Cash allows the consumer to buy anything, while in-kind transfers, such as food stamps, restrict the set of choices.

Figure 3.14 shows the Carte Blanche Principle in action. Cash dominates food stamps. If you are an inframarginal consumer, the cash and food stamps are the same. This consumer is going to buy more food than can be purchased with the allotment of food stamps anyway so if you gave him the cash equivalent value, he would spend the cash on food.

If you are a distorted consumer, however, you are better off if you are given cash because cash can be used to buy the other goods that you prefer over food. With food stamps, when you maximize utility and do the best you can, you end up at a lower level of utility than if you had the cash-equivalent.

In economics, *deadweight loss* is a measure of inefficiency. It is a number that tells you how much a given solution differs from the best solution. In this application, deadweight loss is the difference in utility due to using food stamps instead of cash.

We could try to compute, for each consumer, the maximum utility with cash minus the maximum utility with food stamps. For the inframarginals, this number would be zero, but it would be positive for the distorted consumers.

Unfortunately, this approach would be exceedingly difficult to actually carry out. Even if we managed to do it, remember, we cannot simply add the utility values for different people. Utility is ordinal, ranking only by higher or lower, with no meaningful information about distance or magnitude. Thus, we can never add the utilities of different people.

Theory tells us deadweight loss exists, but the inability to make interpersonal utility comparisons means we are severely limited in how we can measure the sum of deadweight losses of two or more people. As a first pass, we can try to figure out how many distorteds and inframarginals there are. After all, if there are only a few distorted consumers, then we would know that food stamps were not affecting the decisions of too many people.

A Food Stamp Experiment

The empirical work described below comes from Whitmore's "What are Food Stamps Worth?" available at arks.princeton.edu/ark:/88435/dsp01z603qx42c.

Whitmore describes two controlled experiments carried out by the USDA in the early 1990s. In the San Diego experiment, around 1,000 people who were receiving food stamps were randomly selected to participate in the experiment. Half were randomly assigned to the control group and given food stamps as usual, while the other half, the treatment group, were given cash-equivalent aid (checks).

Of the roughly 500 people given checks, about 100 were distorted—they bought less food compared to what they bought when they were given food stamps.

But what were these distorted consumers buying instead of food? This is a crucial question. Most economists are willing to let individuals choose what

to buy because the Theory of Consumer Behavior is built on rational, optimizing decision making. The fundamental world view of economic theory is that individuals know best how to spend their money.

Others, however, argue that low-income consumers make poor decisions if left free to choose what to buy. They think distortion is a good thing because they want aid recipients to buy food. Whitmore (p. 3) says this:

> To some, this distortion is the best part of the food stamp program: the government can ensure that needy families get enough to eat and that they don't spend the money on other things. To others, this distortion represents a waste of resources—it is inefficient to give in-kind transfers instead of cash.

At its most extreme, the issue can be stated this way: Taxpayers will support buying food for the poor, but not drugs, alcohol, and other wasteful consumption. But exactly how distorted consumers would spend cash is an empirical question and Whitmore has the data to answer it.

Researchers in the San Diego experiment kept careful food diaries. When Whitmore compared the purchases of the distorted treatment group to the food stamp control group, she found a marked decrease in a few specific items, like juice and soda, for distorteds. So, surprisingly,

> Even though spending on food declines for the treatment group, the food diary data from San Diego provide no firm evidence that cashing-out food stamps leads to declines in nutritional intake, and suggest that it may actually reduce extreme over-consumption of calories, an important contributing factor to obesity. (Whitmore, p. 35)

The picture that many have of the indigent as drug addicts or exceptionally poor decision makers is unsupported by Whitmore's data. It is true that if forced to spend a subsidy on food, low-income households will spend more on food, but that does not imply that this is better. By definition, low-income people are struggling with paying for, not just food, but a whole host of necessities, including shelter, clothing, transportation, and utility bills. A cash-equivalent subsidy means they can buy food if that is the greatest need or make other important purchases.

The Illegal Sale of Food Stamps

The Theory of Consumer Behavior can be used to explain what most people find puzzling when they first hear about it—there is an active, illegal market in food stamps. Whitmore (p. 4) estimated that food stamps sold for 61 cents on the dollar. The theory can also explain why it has proven incredibly difficult to stop the illegal sale of food stamps.

STEP Proceed to the *Selling* sheet.

Observe that the budget constraint has been modified yet again. The segment below the food stamp allotment (*x1bar*) is no longer horizontal. We have enabled the consumer to sell food stamps and move up the budget constraint.

The slope of this portion of the budget constraint is $ER * p_1/p_2$,where ER is the exchange rate of food stamps for cash. With ER initially set at 0.6 (in cell B24), a seller of food stamps would get 60 cents for every dollar of food stamps sold. The slope of the budget line is 60% of the p_1/p_2 ratio or 1.2.

Notice that cell B16 has been added and it reports the income generated by the sale of food stamps. It shows zero because the opening position is at the kink (20, 33.33) so this distorted consumer isn't selling any food stamps.

STEP Change cell B13 to 10 and watch how the cells and the chart change.

B16 now reports that the consumer is making $12 from the sale of food stamps. They "sold" ten units of food, valued at $20 in cash, but only 60% of that in food stamps. With $p_2 = 3$, she can buy four more units of x_2.

STEP Set cell B14 to 37.33 to move the consumer to the budget line.

But is this is the optimal solution? In fact, comparing cell G27 to H26 tells you that it is not. The consumer is selling too many food stamps at this point.

STEP Run Solver. You should get a result like Figure 3.15, which shows the consumer choosing just under 15 units of food and adding $6.29 of food stamp income (explaining how they managed to buy more than $33\frac{1}{3}$ units of x_2). Notice also that, once again, the MRS (-0.4) equals the slope of the budget constraint (-0.4) on the relevant part of the budget line.

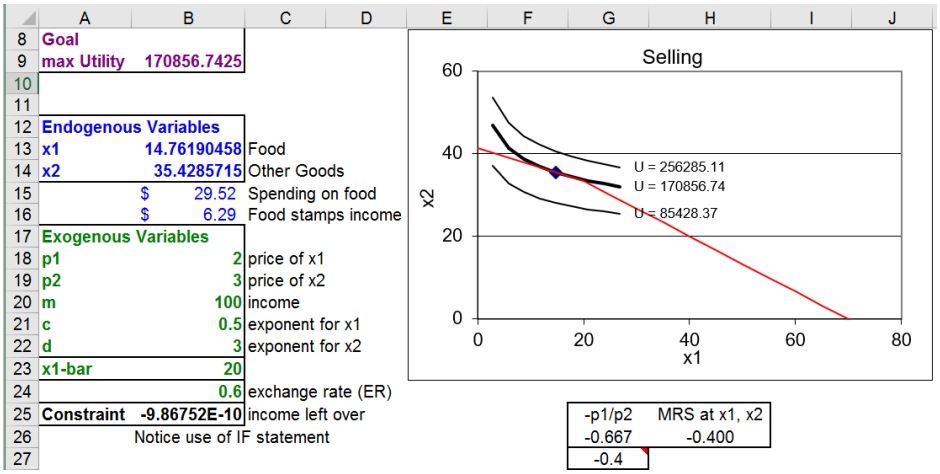


Figure 3.15: Maximizing utility by selling food stamps.
Source: FoodStamps.xls!Selling

The consumer maximizes utility and reaches a higher level of satisfaction than what is attainable by staying on the kink and not selling the food stamps. The ability to get higher satisfaction explains the unintended consequence of an active illegal trade in food stamps.

This analysis does not incorporate the costs of selling food stamps, including the risk of getting caught. There is no doubt that EBT cards make it more difficult to sell food stamps, but the inability to stop the illegal trade testifies to the forces at play—the search for higher satisfaction is powerful indeed.

One Last Question

If the Carte Blanche Principle is true, then why does the government use food stamps instead of cash to help the poor?

Whitmore devotes the conclusion of her paper (p. 38) to answering this question:

> A crucial aspect of the success of the Food Stamp Program is its political popularity. The Food Stamp Program is not an entitlement program, so its budget must be approved annually in the Farm Bill. The program's budget has always been fully funded, due largely to two factors: its popularity as a targeted welfare program among voters, and its popularity among farmers because they think it increases demand for food. (footnote omitted)

As a practical matter, it is not true that, in general, the poor will squander cash subsidies or make terrible buying decisions. Giving aid in the form of food stamps generates a deadweight loss for those distorted consumers who would have been better off with cash. As Whitmore points out, however, it is politically impossible to imagine what is today a $70 billion program being funded annually as a pure cash giveaway. Economics meets politics and the result is a flawed, but functioning anti-poverty program.

Exercises

1. Which parameter in the *Selling* sheet, with the exchange rate set to 0.9, would have to be changed to represent the case of a distorted consumer who decides not to sell food stamps for cash? What would the value of this parameter be?

2. Explain under what condition the MRS equals the price ratio rule (as a condition that the optimal solution has been found) can be violated.

3. A seller of food stamps would obviously prefer a higher price, but what would be the advantage of a higher price in terms of the Theory of Consumer Behavior?

References

The epigraph comes from the first paragraph of Stigler and Becker's "De Gustibus Non Est Disputandum," *The American Economic Review*, Vol. 67, No. 2 (March, 1977), pp. 76 - 90 (www.jstor.org/stable/1807222). The title is a Latin admonition to not quarrel over tastes—do not continue arguing once you pass the point of rational persuasion (similar to "Let's agree to disagree.").

Stigler and Becker offer, however, a second interpretation, which they prefer: "tastes neither change capriciously nor differ importantly between people." Their key point is this:

> The difference between these two viewpoints of tastes is fundamental. On the traditional view, an explanation of economic phenomena that reaches a difference in tastes between people or times is the terminus of the argument: the problem is abandoned at this point to whoever studies and explains tastes (psychologists? anthropologists? phrenologists? sociobiologists?). On

> our preferred interpretation, one never reaches this impasse: the economist continues to search for differences in prices or incomes to explain any differences or changes in behavior. (p. 76)

The idea that tastes are stable and differences in behavior are to be found in price or income shocks is a hallmark of Chicago School economics.

Diane Whitmore's working paper, "What Are Food Stamps Worth?," is available at arks.princeton.edu/ark:/88435/dsp01z603qx42c. Whitmore goes beyond simply counting the number of distorted consumers and offers estimates of deadweight loss.

For more recent work, see Hillary Hoynes and Diane Whitmore Schazenbach, "Consumption Responses to In-Kind Transfers: Evidence from the Introduction of the Food Stamp Program," *American Economic Journal: Applied Economics*, Vol. 1, No. 4 (October 2009), pp. 109 - 139, available at https://www.jstor.org/stable/25760184.

Taxes upon the necessaries of life have nearly the same effect upon the circumstances of the people as a poor soil and a bad climate.

Adam Smith

3.4 Cigarette Taxes

The Carte Blanche Principle says that cash is always as good as or better than in-kind. There is a corollary from the public finance literature: Lump sum taxes are better than quantity taxes.

Public finance is a field of economics that studies the role of government in the economy. Budgeting, collecting taxes, and government spending are some of the areas studied by public finance economists.

There are, of course, many different kinds of taxes. A *lump sum tax* is a fixed amount that must be paid, regardless of how much is purchased. A head tax, where a fee is charged to each person, is an example of a lump sum tax.

A *quantity tax* is an amount for each unit sold so it is added to the price of the product. Federal, state, and local governments levy quantity taxes on gasoline, alcoholic beverages, and tobacco. Unlike a lump sum tax, if more is bought, more quantity tax is paid.

Most people are familiar with sales tax, but this is yet another tax variant. Like a quantity tax, more is paid as more is purchased, but a sales tax is a percentage of the total purchase value. This is an *ad valorem* tax, which is Latin for "according to value."

The goals of taxation can be complicated. The primary motivation for taxes is to pay for government spending, but taxes can also be used to discourage particular activities. Both of these motivations are at play in the case of cigarettes.

Cigarette Smoking and Taxes

The average number of cigarettes sold per day in the United States and Japan since 1900 is shown in Figure 3.16. Visit ourworldindata.org/smoking to see an interactive version of this chart and add other countries. The pattern is the same around the world—rising smoking rates reach a peak, then a rapid decline.

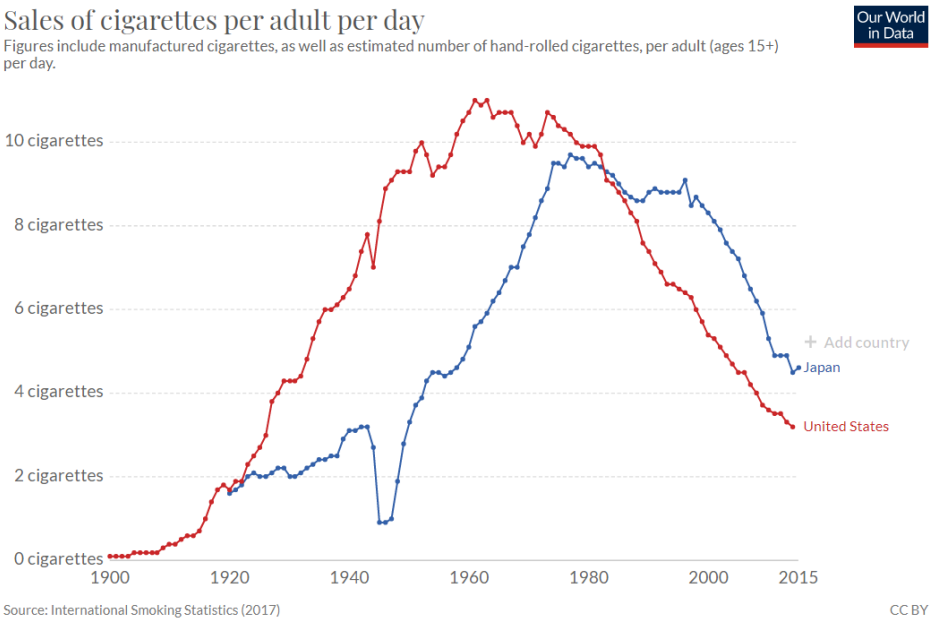


Figure 3.16: Smoking rates in Japan and the United States.
Source: ourworldindata.org/smoking

American soldiers were given cigarettes during the two world wars and this drove the sharp increase in cigarette smoking. The collapse in its smoking rate in the 1940s shows that Japan did not do this. In both countries, awareness of the damaging health effects of smoking triggered the decline.

As consumption underwent this long rise and fall, cigarette tax policies also changed dramatically. Tobacco products have always been taxed, but cigarette taxes have risen dramatically in the last few decades. Figure 3.17 shows tax rates in US states in 2019.

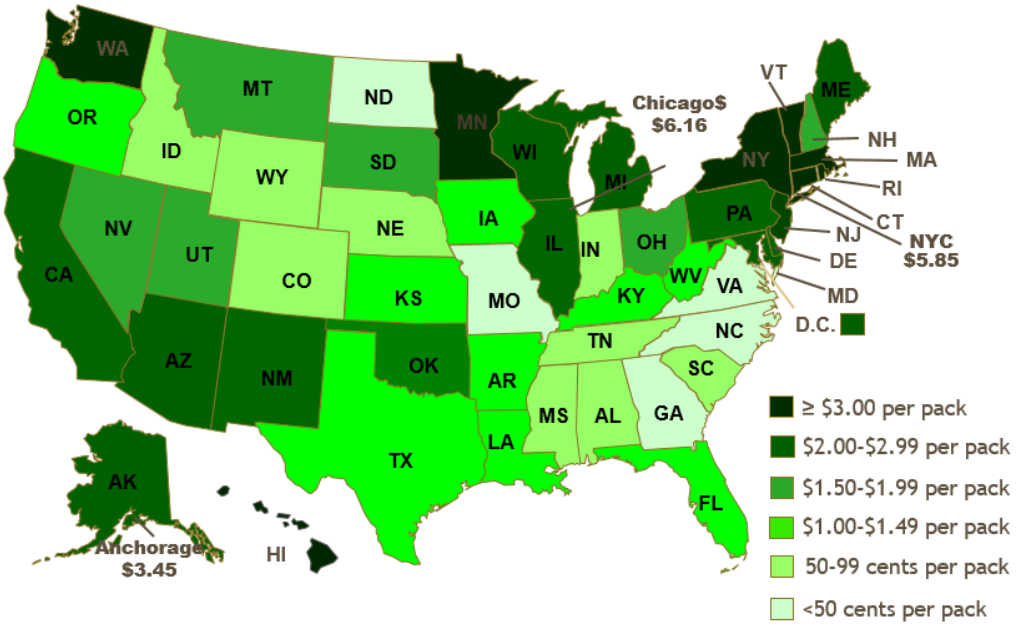


Figure 3.17: State cigarette quantity taxes in 2019.
Source: tobacconomics.org

There is wide variation in state cigarette tax rates. In 2019, New York and Connecticut had the highest state tax of $4.35 per 20-pack of cigarettes. Missouri had the lowest, $0.17 per pack.

Other governmental levels also tax cigarettes. New York City, for example, adds a $1.50 per pack tax, bringing state and local taxes to $5.85 per pack. To this we add the federal tax rate of $1.0066 per pack. Finally, smokers pay a sales tax on the total price paid (including the quantity taxes). In New York City, a pack of cigarettes cost over $10 in 2019.

We will analyze the quantity tax by using the Theory of Consumer Behavior. We will also compare it to a lump sum tax—an option that is not currently being used by the government. To make a good comparison, we have to make sure that the taxes are *revenue neutral*. This means that the tax revenues generated by the tax proposals are the same. It would not be fair to compare a quantity tax that generated $50 in revenues to a $100 lump sum tax.

Quantity Tax

STEP Open the Excel workbook *CigaretteTaxes.xls*, read the *Intro* sheet, and proceed to the *QuantityTax* sheet.

Cell B21 enables us to levy a quantity tax. The sheet opens with cell B21 = 0, which means there is no tax.

The sheet also opens with the consumer considering the bundle 20,60. The MRS is greater than the price ratio (in absolute value) and the consumer can move down the budget constraint so we know utility is not being maximized.

STEP Utility is maximized at 1250 by consuming 25 units of cigarettes (x_1) and 50 units of other goods (x_2). Run Solver to confirm this result.

Suppose we impose a \$1/unit quantity tax on cigarettes. What effect does this have on the consumer?

STEP You can find the consumer's optimal solution after levying the tax by changing cell B21 to 1 and running Solver.

Notice how the chart updated when B21 was set to one. The red budget constraint shows how the line rotated and swung in when the tax was imposed. This is the same as increasing the price of good 1. After running Solver, you can see that the consumer responds by buying fewer cigarettes.

We can also find the optimal solution using analytical methods by solving the following constrained optimization problem:

$$\begin{aligned} \max_{x_1,x_2,\lambda} \; & U(x_1,x_2) = x_1 x_2 \\ \text{s.t. } & 100 = 2(x_1 + Q\_Tax) + x_2 \end{aligned}$$

The consumer wishes to maximize utility (which is Cobb-Douglas with both exponents equal to 1), subject to the budget constraint, with parameter values for income and prices plugged in.

We leave Q_Tax as an exogenous variable so we can find the optimal solution as a function of Q_Tax. We have worked on this problem before, except $p_2 = 1$ (instead of 3) and we have added the quantity tax.

The Lagrangean procedure remains the same and we walk through the four steps to find the answer.

1. Rewrite the constraint so that it is equal to zero.

$$0 = 100 - 2(x_1 + Q\_Tax) - x_2$$

2. Form the Lagrangean function.

$$\max_{x_1, x_2, \lambda} L = x_1 x_2 + \lambda(100 - (2 + Q\_Tax)x_1 - x_2)$$

Notice that we are working with a mixed concrete and general problem. We have numerical values for prices, income, and the utility function exponents, but we have the amount of the quantity tax as a variable. We use this strategy whenever we want to find the optimal solution as a function of a particular exogenous variable.

3. Take partial derivatives with respect to x_1, x_2, and λ.

$$\begin{aligned}
\frac{\partial L}{\partial x_1} &= x_2 - (2 + Q_Tax)\lambda \\
\frac{\partial L}{\partial x_2} &= x_1 - \lambda \\
\frac{\partial L}{\partial \lambda} &= 100 - (2 + Q_Tax)x_1 - x_2
\end{aligned}$$

4. Set the derivatives equal to zero and solve for x_1*, x_2*, and λ*.

$$\begin{aligned}
\frac{\partial L}{\partial x_1} &= x_2 - (2 + Q_Tax)\lambda = 0 \\
\frac{\partial L}{\partial x_2} &= x_1 - \lambda = 0 \\
\frac{\partial L}{\partial \lambda} &= 100 - (2 + Q_Tax)x_1 - x_2 = 0
\end{aligned}$$

We use the usual solution method, moving the lambda terms to the right-hand side and then dividing the first equation by the second, which allows us to cancel the lambda terms.

$$x_2 = (2 + Q\_Tax)\lambda$$
$$x_1 = \lambda$$
$$\frac{x_2}{x_1} = \frac{(2 + Q\_Tax)\lambda}{\lambda}$$
$$x_2 = (2 + Q\_Tax)x_1$$

Finding an expression for x_2 seems like an answer, but it is not because it is a function of x_1. To be a solution (which is called a reduced form), we must solve for x_1 asa function of exogenous variables alone. We must keep working. Canceling the lambda terms has moved us closer to an answer—we have reduced the three equation, three unknown system to two equations in two unknowns.

$$x_2 = (2 + Q\_Tax)x_1$$
$$100 - (2 + Q\_Tax)x_1 - x_2 = 0$$

We substitute the first equation into the second and solve for the optimal amount of good 1.

$$100 - (2 + Q\_Tax)x_1 - [(2 + Q\_Tax)x_1] = 0$$
$$100 = 2(2 + Q\_Tax)x_1$$
$$x_1^* = \frac{50}{(2 + Q\_Tax)}$$

Then, we substitute this into our expression for x_2 to get the optimal amount of good 2.

$$x_2^* = (2 + Q\_Tax)\left[\frac{50}{(2 + Q\_Tax)}\right] = 50$$

We can check this solution with Solver's result by substituting $Q\_Tax = 1$ into the reduced form solution for the two goods. Optimal cigarette consumption is $\frac{50}{3}$ or $16\frac{2}{3}$. Because Q_Tax does not appear in the optimal solution for good 2, its value is simply 50 for any value of Q_Tax.

Lump Sum Tax

Let's see how the consumer would optimize with a lump sum tax that raised the same tax revenue for the government.

STEP Making sure that you have run Solver in the Quantity sheet with B21 = 1 so that B11 is approximately $16\frac{2}{3}$, proceed to the *LumpSumTax* sheet.

The quantity tax imposed in the *QuantityTax* sheet has been replaced with a revenue-neutral lump sum tax. With a \$1/unit quantity tax, the consumer purchases $16\frac{2}{3}$ units of x_1, which means the state generates \$16.67 of revenue from the quantity tax. It could have generated the same revenue by taxing the consumer \$16.67, regardless of how much x_1 or x_2 the consumer bought. This is called a lump sum tax because you pay a fixed amount (that's the "lump sum" part) no matter what you decide to buy.

The difference in the way the lump sum tax operates is reflected in the budget constraint equation. Instead of being part of the price of good 1 like a quantity tax, the lump sum tax is subtracted from income.

$$100 = 2(x_1 + Q\_Tax) + x_2$$
$$100 - Lump\_Tax = 2x_1 + x_2$$

The two charts show how the lump sum tax works differently than the quantity tax. Instead of rotating, the new budget line (in red) in the *LumpSum* sheet has shifted inwards. How would the consumer respond to this tax?

STEP Run Solver to find the optimal solution with the lump sum tax.

Before we compare the quantity and lump sum tax solutions, we confirm Solver's answer in the *LumpSum* sheet by solving the problem analytically.

STEP Try your hand at this problem. Check your work (or peek if you get stuck) by clicking the Show Math button.

Remember, Solver gave you an answer so can be quite sure you are correct if your analytical work gives the same result.

Comparing Quantity and Lump Sum Taxes

We now have the data needed to compare the two tax schemes, as shown in Figure 3.18.

Tax	Revenue	x_1*	x_2*	Utility*
No tax	$0	25	50	1250
Q_tax = $1/unit	$16.67	16 2/3	50	833 1/3
Lump_tax = $16.67	$16.67	20 5/6	41 2/3	868

Figure 3.18: Comparing the tax schemes.

The first row shows that the consumer will buy the bundle 25,50 when there is no tax, generating an optimal utility of 1250. Obviously, there is no revenue because there is no tax.

The second row shows that utility falls to $833\frac{1}{3}$ with an optimal solution of $16\frac{2}{3}$,50 with a \$1/unit of x_1 quantity tax. The tax produces \$16.67 of revenue for the government.

The last row shows that a revenue-neutral lump sum tax of \$16.67 would result in purchases of $21\frac{5}{6}$ and $41\frac{2}{3}$, which would give a level of utility of 868.

The primary lesson is that, for this consumer, if the government needed to raise \$16.67 of tax revenue, the lump sum tax is better than the quantity tax because the consumer's maximum utility is higher under the lump sum tax.

Notice that we are not violating the rule against interpreting utility values as being meaningful. We are not comparing two consumers. We are not treating utility as if it were on a cardinal scale by saying, for example, that there is a gain of 868 minus $833\frac{1}{3}$ equals $34\frac{2}{3}$ utils of increased satisfaction. We are merely saying that satisfaction is higher under the lump sum tax scheme than the revenue-neutral quantity tax.

A graph can be used to explain this rather curious result that lump sum taxes enable higher utility than equivalent revenue quantity taxes. It is a complicated graph, so we will build up to it in stages.

The first layer is simply the initial solution, before any tax is applied. It is shown in Figure 3.19.

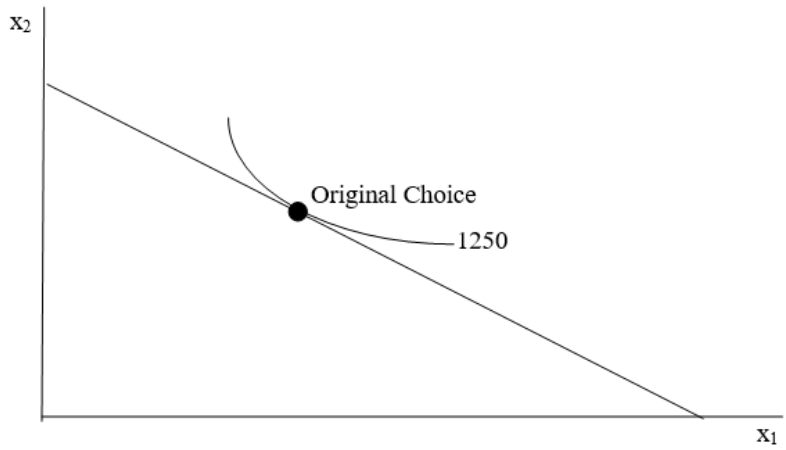


Figure 3.19: The initial optimal solution.

Figure 3.20 shows what happens with a quantity tax. The budget constraint rotates in because the price paid by the consumer (composed of the price of the product plus the tax) has increased. The consumer is forced to re-optimize and find a new optimal solution, labeled *Quantity Tax*. Utility has clearly fallen since we are on a lower indifference curve.

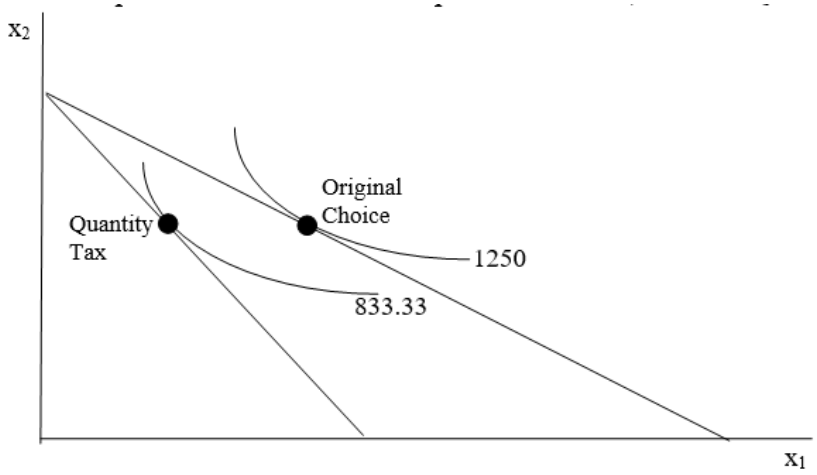


Figure 3.20: Applying a quantity tax.

Then we add a final layer to show the lump sum tax, as shown in Figure 3.21. This enables comparison of the two tax schemes.

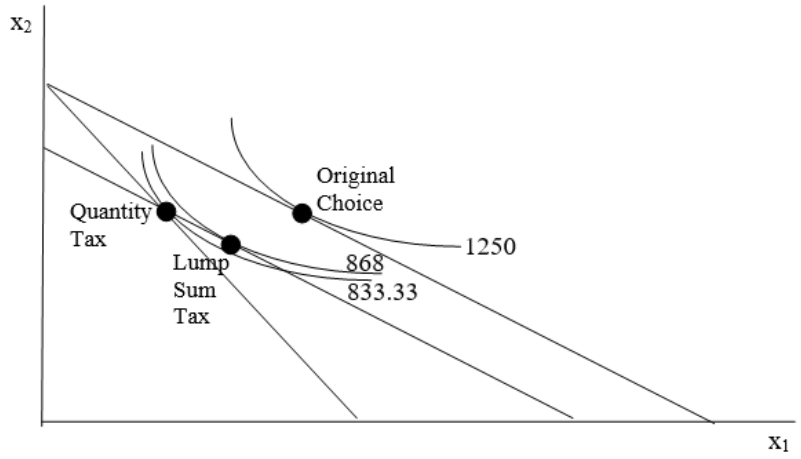


Figure 3.21: Adding a lump sum tax.

The lump sum tax budget constraint has to go through the optimal choice bundle with the quantity tax so that the lump sum tax raises the same revenue as the quantity tax. It also has to be parallel to the original budget constraint. Because it cuts the indifference curve at the quantity tax's optimal solution, we know we can move down the budget line and reach a higher indifference curve than the quantity tax solution.

Figure 3.21 shows that, starting from the *Original Choice* point, we can compare a quantity tax and a revenue-neutral lump sum tax. Figure 3.21 makes clear that the lump sum tax enables attainment of a higher level of utility than the quantity tax because the indifference curve attainable under the lump sum tax is higher than the indifference curve that maximizes utility with the quantity tax.

The reason why the lump sum tax is better is due to the fact that it is *non-distorting*. It leaves the relative prices of the two goods unchanged.

The Lesson and a Follow-up Question

The lesson is that the Theory of Consumer Behavior has been used to show that lump sum taxes are better than quantity taxes. Generating the same amount of revenue, lump sum taxes enable the consumer to reach a higher level of satisfaction than quantity taxes.

This begs a question: Why do we see quantity taxes instead of lump sum taxes? Why are cigarettes (and alcohol and gasoline) so heavily quantity taxed?

The answer lies in the diversity of consumers. The lesson holds only for each individual consumer. It is a fact that there is a revenue-neutral lump sum tax that leaves each individual consumer better off. The amount, however, of the preferable lump sum tax is different, in general, for each consumer. It depends on how many cigarettes (or alcohol or gasoline) each consumer buys. In other words, the lesson does not hold for all consumers taken as a whole. Thus, a single lump tax for all consumers will not necessarily yield higher utility than a quantity tax for each consumer.

This point is obvious if you consider a consumer who does not buy the taxed product at all. This consumer would prefer any size quantity tax to a lump sum tax. After all, if you do not smoke, you do not have to pay any quantity tax on tobacco. The collapse in smoking (see Figure 3.16) goes a long way to explaining why cigarette taxes have soared.

Lump Sum Corollary to the Carte Blanche Principle

We used the Theory of Consumer Behavior to demonstrate a corollary to the *Carte Blanche Principle*: for consumers of a particular product, a lump sum tax is better than a revenue-neutral quantity tax.

If given the option between a quantity and a revenue-neutral lump sum tax, a consumer who buys the taxed good would prefer the lump sum tax because it will leave the consumer with a higher level of utility. Unlike the quantity tax, the lump sum tax will not distort the relative prices faced by the consumer.

Although the *Lump Sum Corollary* is true, we see quantity taxes for various products because the *Lump Sum Corollary* does not apply to all consumers taken as a group. It is not true that there is a single lump sum tax that is preferred to a quantity tax by all consumers.

Exercises

1. Return to the *CigaretteTaxes.xls* workbook and apply a \$2/unit quantity tax. Run Solver. Find the solution by evaluating the reduced form. Show your work. Do the two methods agree?

2. Repeat this for the lump sum tax. Find the revenue-neutral solution via Solver, evaluate the reduced form expression at the new *Lump_Tax*, and compare the two methods. Do the two methods agree?

3. Would the percentage change in the consumer's consumption of x_1 be more affected by a quantity tax if her indifference curves were flatter, assuming a Cobb-Douglas utility function? Describe your procedure in answering this question.

References

The epigraph is from the online version of *The Wealth of Nations* by Adam Smith, who is well known as the father of economics. You can access *The Wealth of Nations* (and many other texts) online at www.econlib.org/. If you want a physical copy of the book, used copies abound or you can get a new, inexpensive copy at www.libertyfund.org/.

Cigarette sales data were obtained from Hannah Ritchie and Max Roser (2020) "Smoking," published online at ourworldindata.org/smoking. Our World in Data is a website with striking, thought-provoking data visualizations on a range of topics.

The state tax map is from Frank J. Chaloupka's PowerPoint presentation, "The Truth about Tobacco Economics," available at tobacconomics.org/. Tobacco data, research, and news from around the world makes this a good source for information on smoking and tobacco policy.

The Centers for Disease Control and Prevention publishes data from 1970 on state-level cigarette prices and taxes from the Tax Burden on Tobacco.

In addition to these data sources, the economics literature on cigarette smoking is vast. Frank A. Sloan, V. Kerry Smith, and Donald H. Taylor, "Information, Addiction, and Bad 'Choices': Lessons from a Century of Cigarettes," *Economics Letters*, Vol. 77 (2002), pp. 147-155, is an accessible, informative starting point.

For a broader, historical review, see Allan M. Brandt, *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America* (2007).

Finally, e-cigarettes are also taxed. Cotti, Chad D., Charles J. Courtemanche, Johanna Catherine Maclean, Erik T. Nesson, Michael F. Pesko, and Nathan Tefft (January 2020), "The Effects of E-Cigarette Taxes on E-Cigarette Prices and Tobacco Product Sales: Evidence from Retail Panel Data," *NBER*, www.nber.org/papers/w26724. From the abstract, "We then calculate an e-cigarette own-price elasticity of -2.6 and a positive cross-price elasticity of demand between e-cigarettes and traditional cigarettes of 1.1, suggesting that e-cigarettes and traditional cigarettes are economic substitutes."

Chapter 4

Comparative Statics

Engel Curves

More Practice with Engel Curves

Deriving a Demand Curve

More Practice with Deriving Demand

Giffen Goods

Income and Substitution Effects

More Practice with IE and SE

A Tax-Rebate Proposal

4.1 Engel Curves

The Theory of Consumer Behavior is built on an optimization problem: maximize utility subject to a budget constraint. It is written in equation form like this:

$$\max_{x_1,x_2} U(x_1, x_2)$$
$$\text{s.t. } p_1x_1 + p_2x_2 = m$$

This problem can be solved analytically or with numerical methods and the solution can be displayed by a canonical graph, as in Figure 4.1. But it turns out that this is just a first step in how economists think.

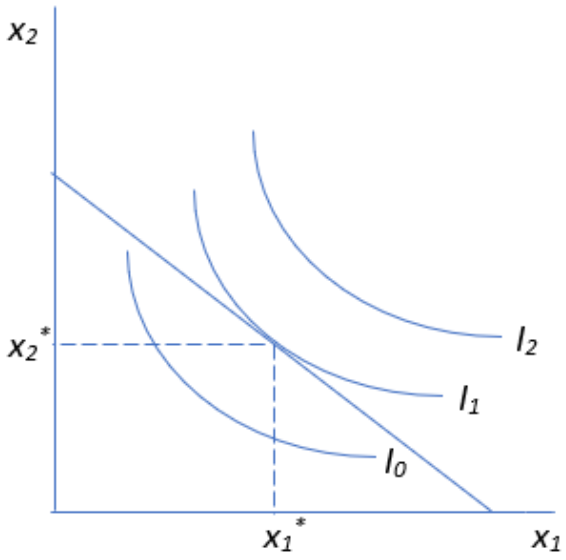


Figure 4.1: Displaying the optimal solution.

The material in this chapter gets to the heart of the economic approach: we explore how the optimal solution responds to a shock, a change in an exogenous variable, holding everything else constant. This is called comparative statics.

The most important comparative statics exercise is based on changing a price, enabling us to derive a demand curve. We start, however, by shocking income and tracking the response. This produces an Engel curve. Starting here gives you a chance to absorb and master the logic of comparative statics before diving into the demand curve.

Initial, Shock, New, Compare

To do comparative statics analysis, we follow a four-step procedure.

1. We find the ***initial*** solution.

2. We change a single exogenous variable, called the ***shock***, holding all other exogenous variables constant. Economists use a Latin phrase, ceteris paribus, as shorthand. This literally means *with other things held equal* and economists use the phrase to mean *everything else held constant*.

3. We find the ***new*** optimal solution.

4. Finally, we ***compare*** the new to the initial solution to see how the optimal solution responded to the shock.

Comparative statics is the fundamental methodology of economics. It gives a framework for interpreting observed behavior. This framework has been given many names, including: the method of economics, the economic approach, the economic way of thinking, and economic reasoning.

While *comparative* clearly points to the comparison between the new and initial solution, the meaning of *statics* (not be confused with statistics) is less obvious. It means that we are going to focus on positions of rest and not worry about the path of the solution as it moves from the initial to the new point.

There are a few complications and additional issues to be aware of when doing comparative statics analysis. Analytical and numerical methods can

be used, but they do not always exactly agree. In addition, we have several ways of comparing the new and initial solutions. A qualitative comparison focuses only on direction (up or down), while quantitative comparisons compute magnitudes of the change in response (either as a difference or a percentage change). Finally, we can display the comparative statics analysis in the canonical graph itself or a separate chart. These three issues will be demonstrated via example.

Elasticity Basics

Elasticity is a pure number (it has no units) that measures the sensitivity or responsiveness of one variable when another changes. Elasticity, responsiveness, and sensitivity are synonyms. An elasticity number expresses the impact one variable has on another. The closer the elasticity is to zero, the more insensitive or inelastic the relationship.

Elasticity is often expressed as "the something elasticity of something," like the price elasticity of demand. The first something, the price, is always the exogenous variable; the second something, in this case demand (the amount purchased), is the response or optimal value being tracked.

A less common, but perhaps easier, way is to say, "the elasticity of something with respect to something." The elasticity of demand with respect to price clearly shows that demand depends on and responds to the price.

Unlike the difference between the new and initial values, elasticity is computed as the ratio of percentage changes in the values. The endogenous or response variable always goes in the numerator and the exogenous or shock variable is always in the denominator.

The percentage change, $\frac{new-initial}{initial}$, is the change (or difference), $new-initial$, divided by the initial value. This affects the units in the computation. The units in the numerator and denominator of the percentage change cancel and we are left with a percent as the units. If we compute the percentage change in apples from 2 to 3 apples, we get 50%. The change, however, is +1 apple.

If we divide one percentage change by another, the percents cancel and we get a unitless number. Thus, elasticity is a pure number with no units. So if the price elasticity of demand for apples is -1.2, there are no apples, dollars, percents, or any other units. It's just -1.2.

The lack of units in an elasticity measure means we can compare wildly different things. No matter the underlying units of the variables, we can put the dimensionless elasticity number on a common yardstick and interpret it. Figure 4.2 shows the possible values that an elasticity can take, along with the names we give particular values.

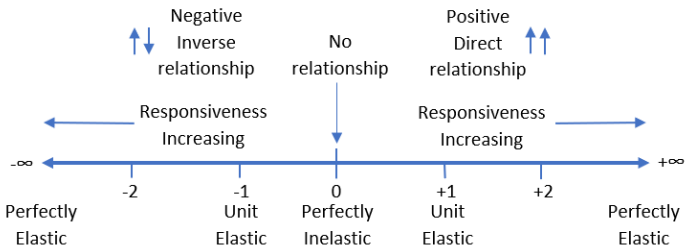


Figure 4.2: Elasticity on the number line.

Empirically, elasticities are usually low numbers around one (in absolute value). An elasticity of $+2$ is extremely responsive or elastic. It means that a 1% increase in the exogenous variable generates a 2% increase in the endogenous variable.

The sign of the elasticity indicates direction (a qualitative statement about the relationship between the two variables). Zero means that there is no relationship—i.e., that the exogenous variable does not influence the response variable at all. Thus, -2 is extremely responsive like $+2$, but the variables are inversely related so a 1% *increase* in the exogenous variable leads to a 2% *decrease* in the endogenous variable.

One (both positive and negative) is an important marker on the elasticity number line because it tells you if the given percentage change in an exogenous variable results in a smaller percentage change (when the elasticity is less than one), an equal percentage change (elasticity equal to one), or greater percentage change (elasticity greater than one) in the endogenous variable.

Elasticities are a confusing part of economics. Below are six common misconceptions and issues surrounding elasticity. Reading these typical mistakes will help you better understand this fundamental, but easily misinterpreted concept.

1. Elasticity is about the *relationship* between two variables, not just the change in one variable. Thus, do not confuse a negative elasticity as meaning that the response variable must decrease. The negative means that the two variables move in opposite directions. So, if the age elasticity of time playing sports is negative, that means both that time playing sports falls as age increases and time playing sports rises as age decreases.

2. Elasticity is a *local phenomenon*. The elasticity will usually change if we analyze a different initial value of the exogenous variable. Thus, any one measure of elasticity is a local or point value that applies only to the change in the exogenous variable under consideration from that starting point. You should not think of a price elasticity of demand of -0.6 as applying to an entire demand curve. Instead, it is a statement about the movement in price from one value to another value close by, say \$3.00/unit to \$3.01/unit. The price elasticity of demand from \$4.00/unit to \$4.01/unit may be different. There are constant elasticity functions, where the elasticity is the same all along the function, but they are a special case.

3. Elasticity can be calculated for *different size changes*. To compute the x elasticity of y, we can go from one point to another, $\frac{\%\Delta y}{\%\Delta x}$, or use the derivative's infinitesimally small change at a point, $\frac{dy}{dx}\frac{x}{y}$. These formulas will be explained below, but the point now is that economists are sloppy in their language and do not bother to distinguish elasticity calculated at a point via calculus (for an infinitesimal change) and elasticity calculated for a finite distance from one point to another. If the function is nonlinear, these two methods give different results. If an economist mentions a point elasticity, it is probably calculated via calculus as an infinitesimally small change.

4. Elasticity always puts the *response variable in the numerator*. Do not confuse the numerator and denominator in the computation. In the x elasticity of y, x is the exogenous or shock variable and y is the endogenous or response variable. Students will often compute the reciprocal of the correct elasticity. Avoid this common mistake by always checking to make sure that the variable in the numerator responds or is driven by the variable in the denominator.

5. You already know this, but remember that elasticity is *unitless*. The x elasticity of y of 0.2 is not 20%. It is 0.2. It means that a 1% increase in x leads to a 0.2% increase in y.

6. Perhaps the single most important thing to remember about elasticity is: *Do not confuse elasticity with slope.* This may be the most common confusion of all and deserves careful consideration.

Economists, unlike chemists or physicists, often gloss over the units of variables and results. If we carefully consider the units involved, we can ensure that the difference between the slope and elasticity is crystal clear.

The slope is a quantitative measure in the units of the two variables being compared. If $Q^* = \frac{P}{2}$, then the slope, $\frac{dQ^*}{dP} = \frac{1}{2}$. This says that an increase in P of \$1/unit will lead to an increase in Q^* of $\frac{1}{2}$ a unit. Thus, the slope would be measured in units squared per dollar (so that when multiplied by the price, we end up with just units of Q).

Elasticity, on the other hand, is a quantitative measure based on percentage changes and is, therefore, unitless. The P elasticity of $Q^* = 1$ says that a 1% increase in P leads to a 1% increase in Q^*. It does not say anything about the actual, numerical \$/unit increase in P, but speaks of the percentage increase in P. Similarly, elasticity focuses on the percentage change in Q^*, not the change in terms of number of units.

Thus, elasticity and slope are two different ways to measure the responsiveness of a variable as another variable changes. Elasticity uses percentage changes, $\frac{\%\Delta y}{\%\Delta x}$, while the slope does not, $\frac{\Delta y}{\Delta x}$. They are two different ways to measure the effect of a shock and mixing them up is a common mistake.

Comparative Statics Analysis of Changing Income

STEP Open the Excel workbook *EngelCurves.xls*, read the *Intro* sheet, and proceed to the *OptimalChoice* sheet.

We have run Solver and the initial solution, ${x_1}^* \approx 25$ and ${x_2}^* \approx 16\frac{2}{3}$, is displayed.

Our first attempt at comparative statics analysis is straightforward: change income, ceteris paribus, and compute the response in ${x_1}^*$ and ${x_2}^*$.

STEP Change cell B18 to 150 (this is the shock) and then run Solver to find the new optimal solution.

The budget line shifts out and the consumer takes advantage by re-optimizing and moving to a new, highest attainable indifference curve.

STEP Compare the initial and new values of x_1^* and x_2^* given the \$50 increase in income.

In qualitative terms, we would say that the increase in income has led to an increase in optimal consumption of the two goods.

In quantitative terms, we can compute the response as the change in the own units of the two variables.

The *own units* statement of comparative statics for x_1^* is $\frac{\Delta x_1^*}{\Delta m}$.

Income rose by \$50 and optimal consumption of each good went up by 12.5 units. We compute $\frac{37.5-25}{150-100}$ so we say that we get an increase of $\frac{1}{4}$ unit for every \$1 increase in income.

Elasticity is another a way to present a quantitative comparative statics result. We use a formula that multiplies the slope by the initial values.

Income elasticity of $x_1^* = \frac{\Delta x_1^*}{\Delta m}\frac{m}{x_1^*} = [\frac{37.5-25}{150-100}][\frac{100}{25}] = 1$. This elasticity is unit elastic. This means that a 1% change in income leads to a 1% change in the optimal purchase of good 1. We had a 50% increase to income and that produced a 50% increase in x_1^*.

The elasticity formula seems mysterious, but it is easily derived from the definition of the ratio of percentage changes.

$$\frac{\%\Delta x_1^*}{\%\Delta m} = \frac{\frac{\Delta x_1^*}{x_1^*}}{\frac{\Delta m}{m}} = \frac{\Delta x_1^*}{x_1^*}\frac{m}{\Delta m} = \frac{\Delta x_1^*}{\Delta m}\frac{m}{x_1^*}$$

The algebra above shows how slope and elasticity are connected. Multiplying the slope by an initial position is the same as computing a percentage change.

While it is certainly possible to do comparative statics analysis by running Solver to find the initial solution, changing a parameter on the sheet, running Solver again to find the new solution, and then comparing the initial and new solutions, the tediousness of this manual approach is obvious.

Fortunately, there is a better way. It involves using the Comparative Statics Wizard Excel add-in.

STEP Click the Reset button to make sure you start from the initial parameter values.

STEP Install the Comparative Statics Wizard add-in, *Cswiz.xla*, from the *MicroExcel* archive.

Instructions and documentation are available in the *CompStatics.doc* file in the *SolverCompStaticsWizard* folder. You can see which add-ins are installed by accessing the Add-ins Manager dialog (In Excel 2019, File: Excel Options: Add-ins: Go).

STEP Once the Comparative Statics Wizard add-in is installed, from the *OptimalChoice* sheet, click the Add-ins tab on the Ribbon, then click Wizard and Comp Statics (in earlier versions, execute Tools: Wizard: Comp Statics) to bring up the main dialog box of the CSWiz add-in, shown in Figure 4.3.

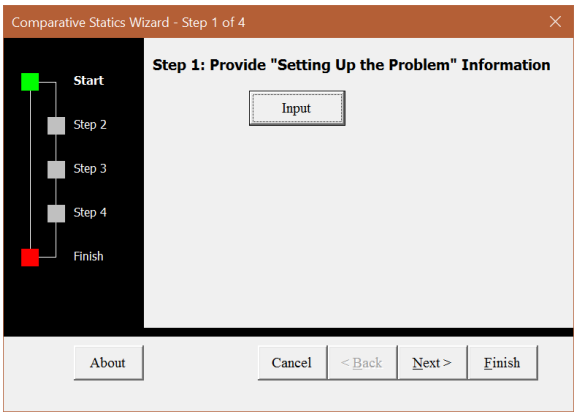


Figure 4.3: First step in the Comparative Statics Wizard.

STEP Click on the Input button and answer the three questions posed.

You are providing Excel with the information it needs to organize the results. Clearly, the goal is cell B7 so you will click on cell B7 when prompted by the first question. Excel enters the absolute reference to that cell ($B$7) in the dialog box and you click OK. Follow the same procedure for the next two questions. The endogenous variables are in cells B11:B12 and the exogenous variables are in cells B16:B20 so can click and drag to select those cells.

Notice how the Comparative Statics Wizard add-in presumes that you have properly organized and set up the problem on the spreadsheet.

STEP Once you have provided the goal, endogenous and exogenous variable cells, click the Next button.

Step 2 uses Excel's Solver to find the initial solution. It temporarily hides the Comparative Statics Wizard and brings up Solver so you can use it to find the optimal solution.

STEP At the Step 2 screen, click the Run Solver button to bring up the Solver dialog box. Click Solve to have Solver find the initial solution.

Read the message in the box after you have run Solver. It explains what you have done so far.

Having found the initial solution, we are ready to input the shock.

STEP At the Step 3 screen, click the Input button.

As in the first screen, you are asked three questions. The first question asks for the shock variable itself. In this case, click on cell B18 (the income variable value, not the label). The second question is the amount of change. Enter 50. The third question is the number of shocks. The default value is 5. Accept this value by clicking the OK button.

You have asked Excel to change income, holding the other variables constant, from 100 to 150 to 200 to 250 to 300 to 350—five jumps of 50 each from the 100 initial value.

STEP After verifying that you have entered the shock information correctly, click the Next button to continue.

The Step 4 screen is the heart of the add-in. You have provided the goal, endogenous and exogenous variable information, Solver found the initial solution, and you have told Excel which variable to shock and how. Excel is ready to run the problem over and over again for each of the shock variable values you provided. It is essentially the manual approach, but Excel does all of the tedious work.

STEP Click the Run Comparative Statics button. The bar displays Excel's progress through the repeated optimization problems. It runs Solver at each value of income, but it is very fast.

STEP Click the Next button, read the information in the box, and click the Finish button.

Excel takes you to a sheet it has inserted into the workbook with all of the comparative statics results. This sheet is similar to the *CS1* sheet. Notice how the results are arranged. It begins with the initial parameter values (widen column A if needed), then displays a table with income in column A, followed by maximum utility and the optimal values of the two goods.

The results produced by the Comparative Statics Wizard can be further processed as shown in the *CS1* sheet.

STEP Proceed to the *CS1* sheet. Columns F and G contains slope and elasticity calculations. Click on the cells to see the formulas.

Notice that you have to be careful with parentheses when doing percentage change calculations in Excel. Simply entering "= C14 – C13/C13" will not do what you want because Excel's order of operations rule will divide C13 by C13 (which is 1) and subtract that from C14.

Income Consumption and Engel Curves

There are two graphs on the *CS1* sheet. They appear to be the same, but they are not. One graph is an income consumption curve and the other is an Engel curve. They are related and understanding their connection is important.

Ernst Engel (not to be confused with Karl Marx's benefactor and friend, Friedrich Engels) was a 19^{th} century German statistician who analyzed consumer expenditure data. He found that food purchases increased as income rose, but at a decreasing rate. This became known as *Engel's Law*. A graph of quantity demanded for a good as a function of income, ceteris paribus, is called an *Engel curve*.

The *income consumption curve* (ICC) shows the effect of the increase in income in the canonical indifference-curves-and-budget-constraint graph. In

other words, the ICC shows the comparative statics analysis on the underlying, canonical graph. Panel A in Figure 4.4 shows the income consumption curve.

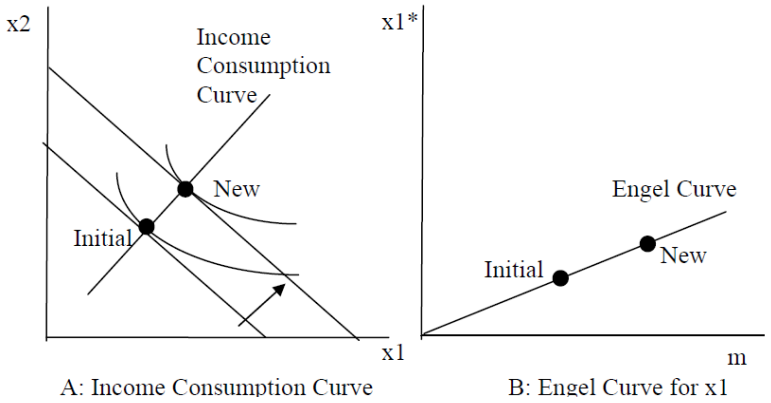


Figure 4.4: Displaying the results of a shock in income.

Panel B shows that the Engel curve for x_1 plots the relationship between income and optimal x_1. This presentation graph shows only the optimal value of the endogenous variable (x_1) as a function of the shock variable (m) and hides everything else. There is an Engel curve graph for x_2, but it is not displayed.

STEP Use your comparative statics results to make Engel and income consumption curves. This will help you understand the relationship between the two curves.

For the Engel curve, select data in m (in column A) and x_1 (in column C). For the ICC, you need to select x_1 and x_2 (in columns C and D). After selecting the data, click the Insert tab in the Ribbon and choose the Scatter chart type in the Charts group.

The slope of the Engel curve reveals if the good is normal or inferior. A *normal good*, as in Figure 4.4, has a positively sloped Engel curve: when income rises, so does optimal consumption. An inferior good has a negatively sloped Engel curve, increases in income lead to decreases in optimal consumption of the good. Figure 4.5 shows this case.

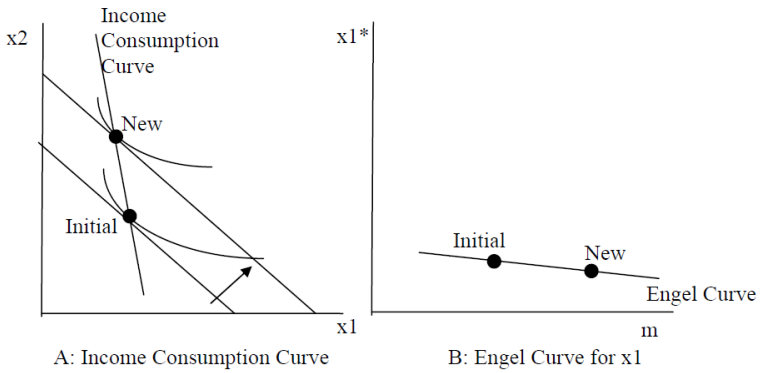


Figure 4.5: x_1 as an inferior good.

Hamburger is the classic inferior good example. As income rises, the idea is that you eat less hamburger meat and more of better cuts of beef. The example also serves to point out that goods are not either normal or inferior due to some innate characteristic, but that the relationship is a local phenomenon. Figure 4.6 shows how a consumer might react across the full range of income. Do you understand the story this graph is telling?

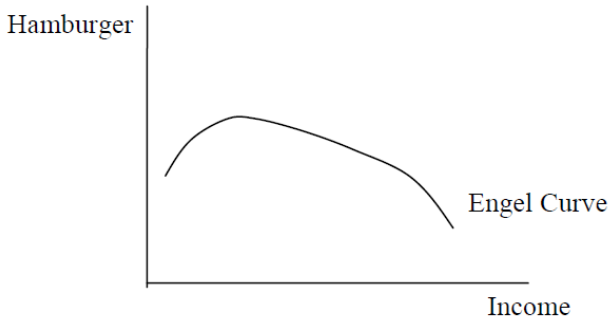


Figure 4.6: A hypothetical Engel curve for hamburger.

Figure 4.6 shows that hamburger is normal at low levels of income (with increasing consumption as income rises), but inferior at higher levels of income. Our Cobb-Douglas utility function cannot generate this complicated Engel curve.

Analytical Comparative Statics Analysis of Changing Income

We can derive the Engel curve for the problem in the *EngelCurves.xls* workbook via analytical methods.

As usual, we rewrite the constraint and form the Lagrangean, then take derivatives, and solve the system of equations. The novelty this time is that we leave m as a letter so that our final answer is a function of income. This enables us to derive an Engel curve.

1. Rewrite the constraint so that it is equal to zero.

$$0 = m - 2x_1 - 3x_2$$

2. Form the Lagrangean function.

$$\max_{x_1, x_2, \lambda} L = x_1 x_2 + \lambda(m - 2x_1 - 3x_2)$$

We take derivatives and set them equal to zero.

$$\frac{\partial L}{\partial x_1} = x_2^* - 2\lambda^* = 0$$
$$\frac{\partial L}{\partial x_2} = x_1^* - 3\lambda^* = 0$$
$$\frac{\partial L}{\partial \lambda} = m - 2x_1^* - 3x_2^* = 0$$

To solve for the optimal values of x_1 and x_2, move the lambda terms in the top two equations to the right-hand side and divide the first equation by the second to eliminate lambda (and give the familiar MRS $= \frac{p_1}{p_2}$ condition. Then solve for optimal x_2 in terms of x_1.

$$\frac{x_2^*}{x_1^*} = \frac{2}{3}$$
$$x_2^* = \frac{2}{3} x_1^*$$

Substitute this expression for x_2 into the third first-order condition and solve for optimal x_1.

$$m - 2x_1^* - 3\left[\frac{2}{3}x_1^*\right] = 0$$
$$4x_1^* = m$$
$$x_1^* = \frac{1}{4}m$$

We can evaluate this expression at any value for m. If we substitute in $m = 100$, we get $x_1{}^* = 25$ which is what we got when we solved this problem with an income of \$100.

Our reduced form expression for $x_1{}^*$ agrees with the values in columns A and C of the *CS1* sheet that we produced via the numerical approach using the Comparative Statics Wizard. The numerical method picks individual points off the Engel curve function that we derived here.

There is also an Engel curve for $x_2{}^*$. It is $x_2{}^* = \frac{1}{6}m$.

Of course, these Engel curves are for this particular consumer, with this particular utility function and set of exogenous variables. Different preferences will give different Engel curves.

If we make the problem more general, in the sense of substituting letters for numbers in the Lagrangean, then these exogenous variables will appear in the reduced form expression. In other words, the one-quarter and one-sixth constants in the Engel curves will be changed into an expression with the exogenous variables. Evaluating that expression at the current values of the exogenous variables will give one-quarter and one-sixth.

If you change an exogenous variable other than income, you will no longer move along the Engel curve. Instead, you will shift the entire Engel curve.

To compute an own units response in $x_1{}^*$ given a change in income, we can simply take the derivative with respect to m, which is simply $\frac{1}{4}$. This means the slope of the reduced form is constant at any value of m.

The elasticity at a given value of m can be computed via the following formula:

$$\frac{dx_1^*}{dm}\frac{m}{x_1^*}$$

Because it is calculated *at* a particular point, this is called *point elasticity*, as opposed to an elasticity measured from one point to another. Economists usually compute and report point elasticities, but they often omit the adjective and simply call the result an elasticity.

Notice how the point elasticity formula is similar to the elasticity formula from one point to another, $\frac{\Delta x_1^*}{\Delta m}\frac{m}{x_1^*}$. We have simply replaced the delta with a d—this shows that the two formulas are the same except for the size of the change in m. Instead of a discrete-size change, the point elasticity formula is based on an infinitesimally small change in m.

At $m = 100$, the point income elasticity of $x_1^* = (\frac{1}{4})(\frac{100}{25}) = 1$. Good x_2 also has a constant unit income elasticity. Rays from the origin always have constant unit elasticities.

The utility function plays a crucial role in comparative statics outcomes. Cobb-Douglas utility functions always yield linear Engel curves with constant unit income elasticities. We do not believe that, in the real world, Engel curves are always linear and unit income elastic. While there are other utility functions with less restrictive results, they are more difficult to work with mathematically. Ease of algebraic manipulation helps explain the popularity of the Cobb-Douglas functional form.

An Engel Curve is Comparative Statics Analysis

This chapter introduced comparative statics analysis. It focused on tracking the optimal solution as income changes. This is called an Engel curve.

Comparative statics analysis, including elasticities, can be done via numerical and analytical methods. The Comparative Statics Wizard handles much of the tedious work in the numerical approach.

We can compute an elasticity in two ways: *at* a point and *from* one point to another. The former uses the derivative and latter is based on a discrete-size change in the exogenous variable. A point elasticity is one based on

the derivative. Both elasticities are based on percentage changes, but the derivative uses infinitesimally small changes in the exogenous variable.

We will often compare the two methods. In this case, the two methods agreed perfectly. This will not always be true.

Exercises

1. Change the price of good 1 from 2 to 3 in the *OptimalChoice* sheet of the *EngelCurves.xls* workbook. From $m = 100$, use the Comparative Statics Wizard to create a graph of the Engel curve for good 1. Title the graph and label the axes. Take a picture of your graph and paste it in your Word document.

2. Why is the slope of your graph different than the one in the *CS1* sheet?

3. Compute the income elasticity of demand for good 1 from $m = 100$ to 200. Show your work.

4. Compute the income elasticity of demand for good 1 at $m = 100$. Show your work.

5. Why are your answers in question 3 and 4 the same?

References

The epigraph is from H. S. Houthakker, "Engel's Law," in J.Eatwell, M.Milgate and P.Newman (eds.) *The New Palgrave Dictionary of Economics*, (London: McMillan, 1987), pp. 143-144.

The *Palgrave* is much more than a simple dictionary. It is a reference resource with articles on specific terms or phrases. The 2008 version of the Palgrave Dictionary is edited by Stephen N. Durlauf and Lawrence E. Blume. It is available online at www.dictionaryofeconomics.com.

> I shall also argue that the most secure propositions and the most reliable predictions, even though they are conditional predictions, arise out of comparative statics, and that when we are asked the awkward question "what good is economics to anyone," apart from its usefulness in providing a gainful occupation for economists, the defense rests mainly on the achievements of rather old-fashioned comparative statics.
>
> Kenneth E. Boulding

4.2 More Practice with Engel Curves

This section derives Engel curves via numerical and analytical methods for different utility functions. It applies the same logic as the previous chapter. This is mastery by repetition. Recognizing how the same steps are used is essential to thinking like an economist.

Quasilinear Preferences

This example uses a quasilinear utility function, $U = x_1^{\frac{1}{2}} + x_2$. The budget constraint is $140 = 2x_1 + 10x_2$.

We begin with the analytical approach. We rewrite the constraint and form the Lagrangean, leaving m as a letter (since we want to derive an Engel curve).

$$\max_{x_1,x_2,\lambda} L = x_1^{1/2} + x_2 + \lambda(m - 2x_1 - 10x_2)$$

We take derivatives and set them equal to zero.

$$\frac{\partial L}{\partial x_1} = \frac{1}{2}x_1^{-1/2} - 2\lambda = 0$$
$$\frac{\partial L}{\partial x_2} = 1 - 10\lambda = 0$$
$$\frac{\partial L}{\partial \lambda} = m - 2x_1 - 10x_2 = 0$$

To solve for the optimal values of x_1 and x_2, we follow our usual approach, moving the λ terms over to the right-hand side and dividing the two equations to cancel the λs.

$$\frac{1}{2}x_1^{-1/2} = 2\lambda$$
$$1 = 10\lambda$$
$$\frac{\frac{1}{2}x_1^{-1/2}}{1} = \frac{2\lambda}{10\lambda}$$
$$\frac{\frac{1}{2}x_1^{-1/2}}{1} = \frac{2}{10}$$

Notice that the MRS is a function of x_1 alone. This is a property of the quasilinear utility function. We can solve for x_1* from the MRS equal to the price ratio equation.

$$\frac{\frac{1}{2}x_1^{-1/2}}{1} = \frac{2}{10}$$
$$\left[x_1^{-1/2}\right]^{-2} = \left[\frac{4}{10}\right]^{-2}$$
$$x_1^* = 6.25$$

Next, we plug this value into the third first-order condition and solve for x_2*.

$$m - 2[6.25] - 10x_2 = 0$$
$$10x_2 = m - 12.5$$
$$x_2^* = \frac{1}{10}m - 1.25$$

To compute an own units response in x_1* given a change in m, we can simply take the derivative with respect to m, which is zero (because m does not appear in the x_1* reduced form). Thus, increases in income leave x_1* unchanged. In other words, the Engel curve for good 1 is horizontal at 6.25.

The own units response for x_2* is $\frac{dx_2^*}{dm}\frac{m}{x_2^*} = \frac{1}{10}$. This means that an additional dollar in income leads to a $\frac{1}{10}$ increase in good 2.

We can use the income elasticity formula, $\frac{dx_1^*}{dm}\frac{m}{x_1^*}$, to compute the income elasticity. At $m = 140$, the income elasticity of $x_1^* = (0)(140/6.25) = 0$, which is perfectly inelastic. This means that changes in m have no effect at all on x_1^*.

These results seem a little strange. Perhaps the numerical approach and Excel can shed some light on what's going on here.

STEP Open the Excel workbook *EngelCurvesPractice.xls*, read the *Intro* sheet, then go to the *QuasilinearChoice* sheet. It shows the optimal solution, 6.25, 12.75, for $m = 140$. Change income to 160.

As expected the budget line shifts out.

STEP Run Solver to find the new initial solution. The resulting chart looks like Figure 4.7.

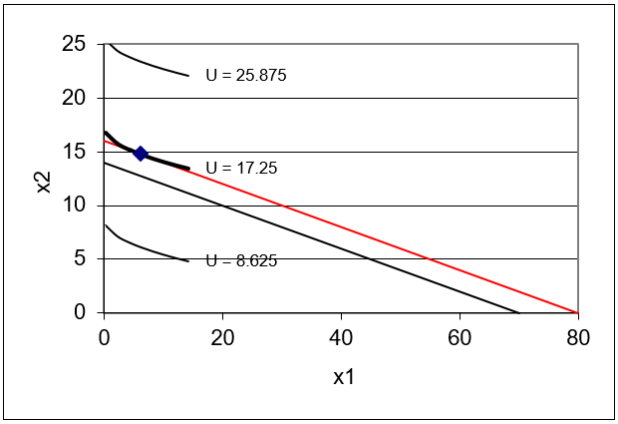


Figure 4.7: Income shock with quasilinear preferences.

Figure 4.7 and your screen show that the value of x_1^* remained unchanged as income rose from \$140 to \$160. This consumer maximizes utility by using all of the extra \$20 in income on good 2.

Figure 4.7 also displays a key property of the quasilinear functional form: the indifference curves are vertically shifted and actually parallel to each other. Thus, when we increase income, the new point of tangency is found directly, vertically up from the original solution.

STEP Return income to its initial value of $140. Run the Comparative Statics Wizard, applying 5 shocks to income in $10 dollar increments.

Your results should look like the *CS1* sheet.

STEP Create Engel and income consumption curves. For the Engel curves, this requires making a chart of x_1* as a function of m and another chart of x_2* as a function of m. For the income consumption curve, the chart is x_2* as a function of x_1*. Each point on this chart is a point of tangency between the budget line and maximum attainable indifference curve.

Your first attempt at making a chart of x_1* as a function of m will not yield a horizontal line at 6.25. Look closely, however, at the y axis scale. The problem is that Solver is reporting numbers very close to, but not exactly, 6.25 as income changes.

But these slight differences in optimal x_1 are not meaningful. They are Solver noise. In fact, for all of these values of m, optimal x_1 really is exactly 25. We need to clean up Solver's results.

Simply changing the display to fewer decimals will not work. This will change the display of the y axis, but Excel will still have the same number in its memory. Instead, we have to use Excel's ROUND function to change the numbers produced by Solver.

The ROUND function has two arguments, the cell you want to round and the number of decimal places. So, ROUND(123.456,1) evaluates to 123.5.

STEP Enter this formula in a blank cell, "=ROUND(123.456,-2)" to see what a negative argument does.

We can use the ROUND function to round Solver's results to the hundredths place. Cell F12 shows how this strategy is implemented.

STEP Apply Excel's Round function to your comparative statics results and then make a chart of the Engel curve for good 1 using the rounded data. Your final chart should look like the one in the *CS1* sheet.

Finally, we can use the CSWiz results to examine the responsiveness of the endogenous variables to the changes in income we applied.

STEP Compute the response to the income changes in own units and income elasticities for x_1* and x_1*. Check your work with the results in the *CS1* sheet.

Notice that the responsiveness results from the numerical method are the same as that via the analytical approach.

Perfect Complements

STEP Proceed to the *PerfCompChoice* sheet to practice on another utility function. This function reflects preferences in which the two goods are perfect complements. This gives L-shaped indifference curves, but our analysis proceeds as usual.

The problem is to maximize the perfect complements utility function subject to the budget constraint. The *PerfCompChoice* sheet shows that $p_1 = 2, p_2 = 10, a = b = 1$.

We do the problem first via the analytical method, leaving m as a letter so we can find $x_1^* = f(m)$ and $x_2^* = f(m)$—these are Engel curves for goods 1 and 2.

In section 3.2, we showed how to solve this problem by finding the intersection of two lines on which the solution must lie. Since $a = b = 1$, the optimal solution must be where $x_1 = x_2$ (a ray from the origin with slope +1). Of course, the solution must also lie on the budget line, so we can solve this system of two equations and two unknowns by substituting in x_1 for x_2 in the budget constraint equation.

$$\left.\begin{array}{l} x_2 = x_1 \\ m = 2x_1 + 10x_2 \end{array}\right\} \Rightarrow m = 2x_1 + 10[x_1] \Rightarrow m = 12x_1 \Rightarrow x_1^* = \frac{m}{12}$$

Since x_2 must equal x_1 at the optimal solution, we know $x_2^* = \frac{m}{12}$.

To compute an own units response in x_1* given a change in income, we can simply take the derivative with respect to m, which is simply $\frac{1}{12}$. This slope is constant and the Engel curve is linear.

The income elasticity at a given value of m can be computed via the point elasticity formula, $\frac{dx_1^*}{dm}\frac{m}{x_1^*}$. At $m = 50$, the income elasticity of $x_1^* = \frac{1}{12}\frac{50}{4.167} = 1$. This means that a 1% change in m will result in a 1% change in x_1^*.

STEP Run the Comparative Statics Wizard on the *PerfCompChoice* sheet (you can make the change in income \$10) and create Engel and income consumption curves.

STEP Compute the response to the income changes in own units and income elasticities for x_1^* and x_2^*.

Check your work with the results in the *CS2* sheet. Notice that the results in Excel are the same as the analytical approach.

The Utility Function Determines the Shape of the Engel Curve

This section ran a comparative statics analysis of a change in income on quasilinear and perfect complement utility functions. This enabled practice in deriving Engel curves and income consumption curves, along with computing responsiveness in own units and elasticities.

The quasilinear function has the peculiar result that the income elasticity of x_1^* is zero. This happens because the indifference map of a quasilinear utility function is a series of vertically parallel curves. Thus, when the budget line shifts out, the new optimal solution is found directly above the initial solution and x_1^* remains unchanged.

With the perfect complements utility function, we were able to find an analytical solution even though we could not use the Lagrangean method. The Engel curve for x_1^* has a constant slope and a unit income elasticity. These are the same properties for the Engel curve we found in the previous chapter using the Cobb-Douglas functional form.

The shape of the Engel curve, its slope and income elasticity are all influenced by the consumer's utility function. The relationship is complicated, so there is no rule or simple statement about how the functional form of utility determines the Engel curve.

Ernst Engel wanted to know how spending on food changed as income rose. He believed food purchases would increase at a decreasing rate as income increased, as shown in Figure 4.8. This makes common sense. As you get richer and richer, you can buy a much nicer house and cars, but it is difficult to spend a lot more on food. This is known as *Engel's Law.*

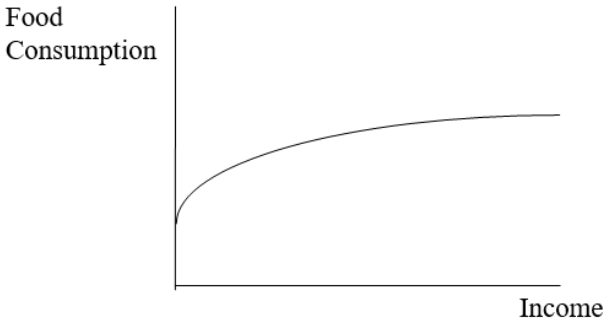


Figure 4.8: Engel's Law.

None of three utility functions we have encountered thus far (Cobb-Douglas, quasilinear, and perfect complements) are capable of generating an Engel curve that conforms to Engel's Law for food purchases. If we were interested in food, we would have to find and use a utility function with an Engel curve that conformed to Engel's Law. Such functions exist, but as you can imagine, they are more complicated than the computationally simple functions we have used thus far.

Exercises

1. In the *QuasilinearChoice* sheet, copy cell B11 and paste it in cell C11. Set income to $200 and run Solver to find the new optimal solution. In cell D11, enter a formula to find the difference between cell C11 and B11. Is this tiny difference meaningful? Explain.

2. Having changed income and run Solver in question 1, if you connected the initial and new solutions on the chart, you would get a vertical line. Why is this happening? Will this happen with every consumer?

3. Having changed income and run Solver in question 1, is good 1 a normal or an inferior good? Explain.

4. Use Word's Equation Editor to solve the general version of the perfect complements problem. In other words, find x_1* and x_2* for

$$\max_{x_1,x_2} U = \min\{ax_1, bx_2\}$$
$$\text{s.t. } m = p_1x_1 + p_2x_2$$

References

The epigraph is from pages 487 and 488 of Kenneth E. Boulding, "In Defense of Statics," *The Quarterly Journal of Economics*, Vol. 69, No. 4 (November, 1955), pp. 485–502 (www.jstor.org/stable/1881991). As you can tell from the quotation, Boulding had a well-deserved reputation for witty, biting comments. His defense of comparative statics in the article just cited notwithstanding, he once quipped, "Mathematics brought rigor to Economics. Unfortunately, it also brought mortis."

The first "empirical" demand schedule was published in 1699 by Charles Davenant.

George Stigler

4.3 Deriving a Demand Curve

We know how to find the initial optimal solution in the Theory of Consumer Behavior and we have explored the comparative statics properties of a change in income.

We are well prepared to embark on the most important comparative statics analysis in the Theory of Consumer Behavior: deriving a demand curve.

Numerical Comparative Statics Analysis of Changing Price

STEP Open the Excel workbook *DemandCurves.xls* and read the *Intro* sheet,then go to the *OptimalChoice* sheet.

The problem is set up, but the consumer is not optimizing because the MRS does not equal the price ratio and the consumer can move to higher indifference curves by traveling up the constraint.

STEP Run Solver to find the initial solution: $x_1{}^* = 25$ and $x_2{}^* = 16\frac{2}{3}$.

Next, we explore how this initial optimal solution changes as the price of good 1 changes, ceteris paribus. This comparative statics analysis will produce a demand curve.

Before we actually do it, can you anticipate what will happen when we increase the price of good 1? Believe it or not, if you try to figure it out first—before actually seeing it—you will learn more. Take a moment to think: what will happen to the graph on your screen when we increase the price of x_1?

Let's see how you did.

STEP Shock: Change cell B16 to 3.

Figure 4.9 shows how your screen should look. With a higher p_1,the budget constraint rotates in, pivoting on the x_2 intercept. The consumer now has fewer consumption possibilities and needs to re-optimize to find the new optimal solution.

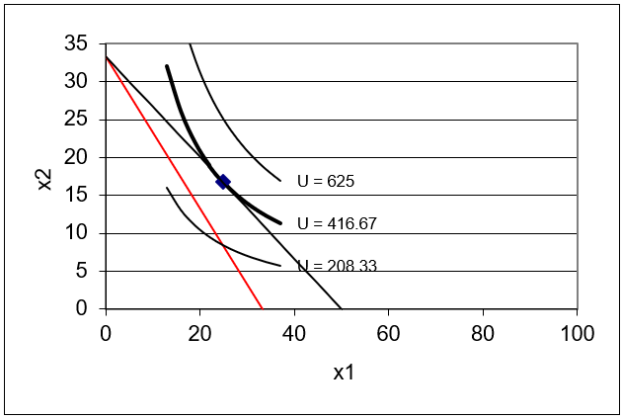


Figure 4.9: New budget line when p_1 rises.

STEP New: Run Solver to find the new optimal solution.

We have completed initial, shock, and new—the last step is to compare. Figure 4.10 shows a table that displays the comparative statics results.

p_1	x_1^*	x_2^*	$\Delta x_1^*/\Delta p_1$	$\%\Delta x_1^*/\%\Delta p_1$	$\Delta x_2^*/\Delta p_1$	$\%\Delta x_2^*/\%\Delta p_1$
2	25	$16\frac{2}{3}$				
3	$16\frac{2}{3}$	$16\frac{2}{3}$	$-8\frac{1}{3}$	−0.67	0	0

Figure 4.10: Comparative statics results of an increase in p_1.

In qualitative terms, we can see that x_1* falls as p_1 rises, but x_2* remains unchanged.

Quantitatively, we can compute the own units response in good 1 as new minus initial x_1*, which is $16\frac{2}{3} - 25 = -8\frac{1}{3}$ divided by 1 (from $3-2$). This is the value displayed in the table. The own units response in x_2 is zero since it did not change.

Responsiveness in percentage terms is the price elasticity of demand. We need to compute the percentage change in $x_1{}^*$ divided by the percentage change in p_1. The numerator is -33% because $\frac{16\frac{2}{3}-25}{25} = -\frac{1}{3}$. The denominator is $\frac{3-2}{2} = 0.5$ or 50%. So, a 50% increase in price, from $p_1 = 2$ to 3, caused a 33% decrease in quantity demanded. Thus the price elasticity of demand is $\frac{-0.33}{0.5} = -\frac{2}{3}$ or roughly -0.67. This number is displayed in the table in Figure 4.10.

The same calculation can be performed on x_2. Since we are considering the effect on *good 2* from a shock to the price of *good 1*, we call this a *cross price* analysis. The term *cross* is used in economics when we examine the effect of i on j; an *own* effect, for example, would be p_1 on x_1.

We quickly realize that the cross price elasticity, the p_1 elasticity of x_2, is zero because the numerator is zero. This is perfectly inelastic or completely unresponsive.

Comparative statics via numerical methods is easier with the Comparative Statics Wizard add-in. If it is not installed, return to the beginning of this chapter to load the CSWiz add-in.

STEP Analyze the effect of a change in p_1 by running the CSWiz add-in and changing the price of good 1 by \$1 increments (for five shocks).

You can see a slightly different comparative statics analysis in the *CS1* sheet. Instead of changing price by one dollar increments, the CS1 sheet was performed with a shock size of 0.1.

STEP Use your comparative statics results to make a *demand curve*, a graph of $x_1{}^* = f(p_1)$. To do this, select the p_1 data in column A, then hold down the *ctrl* key (and keep holding it), while selecting the x_1 data in column C. With cells in columns A and C selected, select the Scatter chart type. Title the graph and label the axes.

Another way to display the comparative statics results is via the *price consumption (or offer) curve*, as shown in Panel A of Figure 4.11 for a utility function that is not Cobb-Douglas and not meant to display the increasing price analysis that you just completed. Instead, a price decrease is shown.

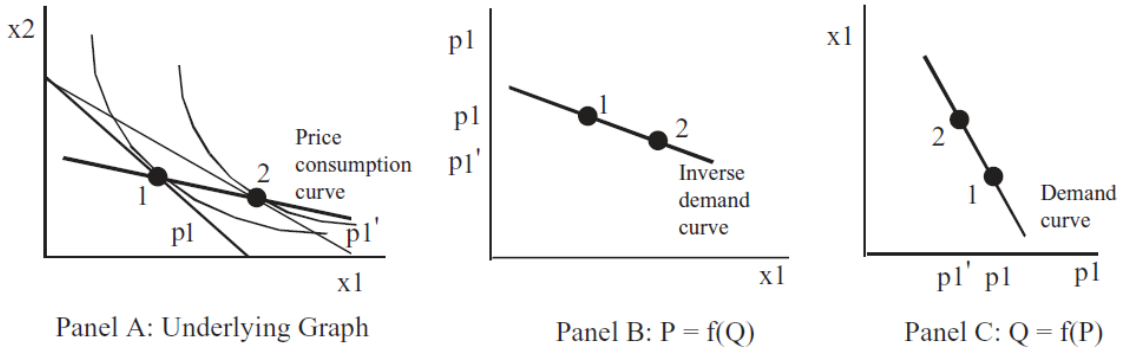


Figure 4.11: Three ways to show effects of p_1 shock.

There is a lot going on in Figure 4.11. The graph on the left (Panel A) shows a price decrease swinging the budget constraint out. It uses numbers to indicate the initial and new optimal solutions.

Panels B and C show demand, but look closely, the axes have been flipped. Instead of graphing x_1 as a function of p_1, the exogenous variable (p_1) is on the y axis in Panel B. This is a backwards, but common presentation in economics. The roots of this strange way of presenting the results can be traced back in the history of economics to Alfred Marshall in 1890.

Modern economists call the graph in Panel B of Figure 4.11 an *inverse demand curve* because it is plotted as $P = f(Q)$. The demand curve, the mathematically correct version, is $Q = f(P)$ because we plot $y = f(x)$ with y as the dependent variable that is determined by x.

In introductory economics, the inverse demand curve is used. The professor just draws a downward sloping line or curve and pronounces that it is obvious that as price goes up, quantity demanded falls (we will soon see that this is not guaranteed). As the level of sophistication rises, especially if we are doing econometrics and trying to estimate a demand curve, economists use the mathematically correct demand curve. Economists are used to both ways of presenting demand. It is confusing at first, but you can get the hang of it pretty quickly.

STEP Read the information in the *CS1* sheet. It explains how the ROUND function was used to create the price consumption curve from the comparative statics results.

Notice that the price consumption curve for changes in p_1 in the Excel workbook is horizontal. This is a property of the Cobb-Douglas utility function and is not especially realistic. The indifference map in Figure 4.11 is not based on a Cobb-Douglas utility function because the price consumption curve is not horizontal.

Another useful Excel skill to master that is especially relevant right now involves controlling the x and y axes. Excel's default is that the leftmost column of selected data goes on the x axis. If we want to make a demand curve with the data in the *CS1* sheet, this is convenient. We select the data in column A (p_1), hold down the *ctrl* key and select the data in column C (x_1). When you make a Scatter chart, Excel puts price on the x axis and quantity on the y axis.

But what if we want to make an inverse demand curve, with p_1 on the y axis? One easy way to do it is by directly editing the SERIES formula in the chart.

STEP Visit vimeo.com/econexcel/using-series-formula to watch a quick, 5-minute video of how the SERIES formula works.

After you watch the video, try it on your demand curve chart. Can you flip the axes by directly editing the SERIES formula? Click on your demand curve, then switch columns A and C in the x and y arguments in the SERIES formula. To see an example of this, click on the series in the chart in the *CS1* sheet.

Analytical Comparative Statics Analysis of Changing Price

We take the opportunity here to extend our previous analytical work. We could just leave p_1 as a letter since we want to derive a demand curve, but we will be more aggressive and leave all exogenous variables as letters. This will give us the most general answer we can get.

We rewrite the constraint and form the Lagrangean.

$$\max_{x_1,x_2,\lambda} L = x_1^c x_2^d + \lambda(m - p_1 x_1 - p_2 x_2)$$

Although it seems more formidable than when numbers are used in place of letters, we can apply the usual strategies for taking derivatives and solving the first-order conditions to find the optimal solution.

We take derivatives and set them equal to zero.

$$\begin{aligned}\frac{\partial L}{\partial x_1} &= cx_1^{c-1}x_2^d - p_1\lambda = 0\\ \frac{\partial L}{\partial x_2} &= dx_1^c x_2^{d-1} - p_2\lambda = 0\\ \frac{\partial L}{\partial \lambda} &= m - p_1x_1 - p_2x_2 = 0\end{aligned}$$

To solve for the optimal values of x_1 and x_2, we move the lambda terms to the right-hand side and divide the first equation by the second. This gets rid of lambda and gives the familiar MRS $= \frac{p_1}{p_2}$ condition, which can then be solved for optimal x_2 as a function of optimal x_1.

$$\begin{aligned}\frac{cx_2^*}{dx_1^*} &= \frac{p_1}{p_2}\\ x_2^* &= \frac{d}{c}\frac{p_1}{p_2}x_1^*\end{aligned}$$

We substitute this expression into the third first-order condition (the budget constraint) and solve for optimal x_1.

$$\begin{aligned}m - p_1x_1^* - p_2\left[\frac{d}{c}\frac{p_1}{p_2}x_1^*\right] &= 0\\ \left(1 + \frac{d}{c}\right)p_1x_1^* &= m\\ x_1^* &= \left(\frac{c}{c+d}\right)\frac{m}{p_1}\end{aligned}$$

This expression contains the demand curve for x_1 because it shows the quantity demanded at a given p_1. It also contains an Engel curve because it shows

how x_1 varies with income. It also shows how x_1 moves when c or d, the consumer's tastes and preferences, change—although, such a graph is unnamed.

Furthermore, this expression can be evaluated for any combination of exogenous variable values. For example, suppose $c = d = 1, p_1 = 2$, and $m = 100$. Then it can be seen easily that optimal $x_1 = 25$. In fact, you can readily see that the reduced form expression for optimal x_1 agrees with the numerical approach using the Comparative Statics Wizard to recalculate the optimal solution at given values of p_1.

We can use our reduced form expression to calculate an own units response to a shock in p_1 by taking the derivative with respect to p_1.

$$x_1^* = \left(\frac{c}{c+d}\right)\frac{m}{p_1}$$

$$x_1^* = \left(\frac{c}{c+d}\right) m\,(p_1)^{-1}$$

$$\frac{dx_1^*}{dp_1} = -1\left(\frac{c}{c+d}\right) m\,(p_1)^{-2}$$

This formidable-looking expression is the instantaneous rate of change of the demand curve at a particular point. Because x_1* is a nonlinear function of p_1, its derivative with respect to p_1 contains p_1. The fact that the demand curve is not a line explains why we get different results when we compute responsiveness with Δ versus d.

STEP Read the *CS1* sheet carefully. Your primary goal is to understand the relationship between Δ in cells F14 and G14 versus the derivative in cells I13 and J13.

The key idea is this: as Δ gets smaller, it approaches d. Thus, earlier, we computed the price elasticity of demand from $p_1 = 2$ to 3 and got -0.67. But the *CS1* sheet shows an elasticity of -0.95 (in G14) as we go from $p_1 = 2$ to 2.1 and when we use the derivative formula, which is based on an infinitesimally small change in p_1, we get an elasticity of -1.

Notice that, unlike the demand curve, $x_1^* = f(p_1)$, the Engel curve, $x_1^* = f(m)$ is a line for the Cobb-Douglas utility function. We say, "x one star is nonlinear in p one" and "x one star is linear in m." Because the Engel

curve is a line, Δm and the derivative with respect to m give identical results. The size of the change in m does not matter if the relationship is linear.

The unit price elasticity is a property of a Cobb-Douglas utility function. We can use the reduced form expression for ${x_1}^*$ to show that we always get a -1 price elasticity.

$$\frac{dx_1^*}{dp_1} \cdot \frac{p_1}{x_1^*} = -1 \left(\frac{c}{c+d}\right) m (p_1)^{-2} \frac{p_1}{\left(\frac{c}{c+d}\right)\frac{m}{p_1}}$$

$$\frac{dx_1^*}{dp_1} \cdot \frac{p_1}{x_1^*} = -1$$

So Cobb-Douglas produces three constant elasticities:

1. Unit income elasticity
2. Unit own price elasticity
3. Zero cross price elasticity

None of these are especially realistic. Cobb-Douglas is common because it is easy to work with, not because it produces sensible elasticities.

A Point Off the Demand Curve?

Unlike an introductory economics course where demand curves appear out of the blue as downward sloping lines or curves, understanding where demand curves come from and what they actually represent are major goals for us.

So far, we have a mechanical understanding of the derivation of demand. Yes, it is true that changing p_1, ceteris paribus, and tracking how ${x_1}^*$ changes is how a demand curve is derived. And, yes, it is true that at every price, quantity demanded is the solution to an optimization problem for that price. But let's try a thought experiment not included in introductory economics.

If we consider what it means to be at a point off the demand curve, such as point Z in Figure 4.12, it helps us understand that the demand curve is really like a ridgeline across the top of a mountain range.

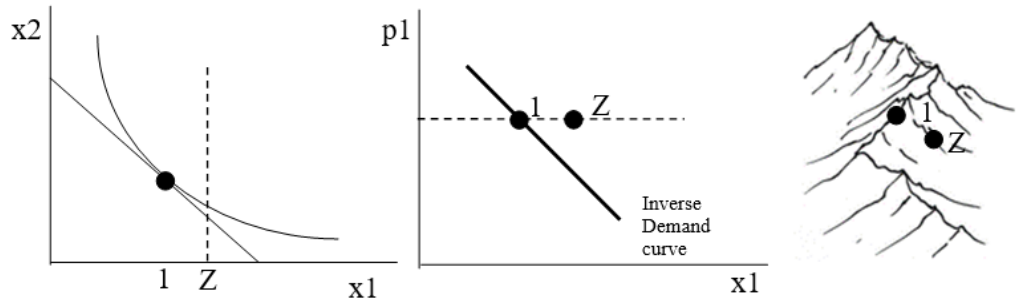


Figure 4.12: Interpreting a point off the (inverse) demand curve..

With a point Z to the right of the inverse demand curve, we know that the consumer is buying too much x_1, as shown by the vertical dashed line in the graph on the left of Figure 4.12. We cannot precisely plot the point Z on the indifference curve graph because we do not know how much good 2 the person is buying at point Z. We do know, however, that she is not optimizing. In other words, at point Z, this consumer is failing to maximize satisfaction and is not on the tangency of the budget line and highest attainable indifference curve.

Considering the meaning of a point off the demand curve reveals that a demand curve is a geometrical object with a special characteristic—every point on the demand curve is a point of maximum utility given prices and income. If we added an axis for utility, the demand curve would show itself as a 3D object that displayed the maximum utility at each given price. In other words, the demand curve is a ridgeline that connects mountain peaks, as shown in the sketch on the right in Figure 4.12.

A Demand Curve Is a Comparative Statics Exercise

Deriving a demand curve is the most important comparative statics exercise in the Theory of Consumer Behavior. Demand and supply (the most important comparative statics exercise in the Theory of the Firm) are at the heart of the market mechanism.

Given a particular functional form for utility, demand curves can be derived via numerical methods, picking off individual points on the demand curve for explicit values of price, ceteris paribus. Slopes and elasticities can be computed.

Demand curves can also be derived via analytical methods by finding the reduced form expression as a function of price (and any other exogenous variables). Slopes and elasticities can be computed by using the derivative.

For Cobb-Douglas utility, we found that ${x_1}^* = (\frac{c}{c+d})\frac{m}{p_1}$. For this reduced form, the numerical and analytical methods yield different values for slopes and elasticities based on changing p_1 because the demand curve is a curve, instead of a line (like the Engel curve). The smaller the discrete change in p_1 used in the numerical method, the closer it gets to the analytical result.

We can also "derive" a demand curve with graphs, as shown in Figure 4.11. We can display the effect of a price change by rotating the budget line and showing the initial and new points of tangency. If we display the p_1 and corresponding optimal amount of x_1 in a separate graph, we have graphically derived a demand curve (or inverse demand curve, if we flip the axes).

Finally, if we work out the implications of a point off the demand curve, we can see the demand curve in a new light—it is actually a 3D object represented in 2D space. All of the points on the demand curve are actually points of maximum utility subject to the budget constraint.

Exercises

1. In the *OptimalChoice* sheet, click the Reset button and reproduce Figure 4.10 with a decrease (instead of an increase) in p_1 from \$2/unit to \$1/unit. Use Word's Table feature to create the table and fill in the cells.

2. Use Word's Drawing Tools to create a graph of the price consumption curve and demand curve for x_1 (as in Figure 4.11) that accurately reflects the shock and results from question 1.

3. What is the difference between a demand curve and an inverse demand curve?

References

The epigraph is from page 103 of George J. Stigler, "The Early History of Empirical Studies of Consumer Behavior," *The Journal of Political Economy*, Vol. 62, No. 2 (April, 1954), pp. 95–113 (www.jstor.org/stable/1825569)

Most economists do not care who first came up with the concept of a demand schedule. Most of those who do care believe that it was Gregory King, a century after Charles Davenant. Stigler was a winner of the Nobel Prize in Economics and a professor at the University of Chicago. He had a lifelong passion for the intellectual history of economics. In this article, he showed that Davenant actually preceded King.

It took a long time to translate demand (and supply) schedules as tables (with columns for price and quantity) into graphs. Fleeming (pronounced flem-ming) Jenkin in 1870 is often given credit for drawing the first demand curve, but there were precursors. Alfred Marshall's *Principles of Economics* (1890) popularized supply and demand graphs. His graphs appeared, however, only in footnotes.

Marshall's *Principles* was the most popular economics book of its era. It is freely available online at www.econlib.org/library/Marshall/marP.html.

Modern economists sometimes mock Marshall for switching the axes, claiming he made a mistake, but this assertion is incorrect. Marshall put price on the vertical axis because he wanted to show market demand and supply curves on a graph as the horizontal sum of individual demand and supply curves, as in footnote 70 from Book III, Chapter IV. Future generations of introductory economics students became locked in to the Marshallian inverse demand and supply curves.

Although you may conclude that Marshall's violation of accepted mathematical convention (i.e., independent variables belong on the x axis) is confusing, the decision was not due to a lack of math knowledge. In fact, Marshall was a brilliant mathematician, earning Second Wrangler (to the future Lord Rayleigh) as an undergraduate at Cambridge in the Tripos competition.

To understand how the role of mathematics has changed in economics, consider the recipe Marshall gave a friend for using math in economics: "1) Use mathematics as a shorthand language, rather than as an engine of inquiry. 2) Keep to them till you have done. 3) Translate into English. 4) Then illustrate by examples that are important in real life. 5) Burn the mathematics. 6) If you can't succeed in 4 burn 3. This last I did often." (A. C. Pigou, Memorials of Alfred Marshall, 1925, p. 427.)

Quasilinear utility functions are not particularly realistic, but they are very easy to work with.

Hal Varian

4.4 More Practice with Deriving Demand

This section derives the demand curve from two different utility functions, quasilinear preferences and perfect complements, to provide practice deriving demand curves. Nothing new here, just practice applying the tools, techniques, and concepts of the economic way of thinking.

Quasilinear Preferences

We begin with the analytical approach. Rewrite the constraint and form the Lagrangean, leaving p_1 as a letter so we can derive a demand curve.

$$\max_{x_1,x_2,\lambda} L = x_1^{1/2} + x_2 + \lambda(140 - p_1x_1 - 10x_2)$$

STEP Follow the usual Lagrangean procedure to solve this problem. For help, refer back to section 4.2 where we solved this same problem except with m instead of p_1.

You should find reduced form expressions like this:

$$x_1^* = \frac{25}{p_1^2}$$

$$x_2^* = 14 - \frac{2.5}{p_1}$$

The first expression, ${x_1}^* = \frac{25}{p_1^2}$, is a demand curve for ${x_1}^*$ because it gives the quantity demanded of x_1 as a function of p_1. If we rewrite the equation in terms of p_1 like this, $p_1^2 = \frac{25}{{x_1}^*} \rightarrow p_1 = \frac{5}{\sqrt{{x_1}^*}}$ then we have an inverse demand curve, with price on the y axis as a function of quantity on the x axis.

The derivative of $x_1{}^*$ with respect to p_1 tells us the slope of the demand curve at any given price.

$$x_1^* = 25p_1^{-2}$$
$$\frac{dx_1^*}{dp_1} = -2 \cdot 25p_1^{-3} = -\frac{50}{p_1^3}$$

The own price elasticity of demand is:

$$\frac{dx_1^*}{dp_1} \cdot \frac{p_1}{x_1^*} = -\frac{50}{p_1^3}\frac{p_1}{\frac{25}{p_1^2}} = -2$$

The constant elasticity of demand for good 1 is a property of the quasilinear utility function. Notice that 2 is the reciprocal of the exponent on x_1 in the utility function. In fact, with $U = x_1^c + x_2$, the price elasticity of demand for x_1 is $-\frac{1}{1-c}$ for values of c that yield interior solutions.

The expression for optimal x_2 is a cross price relationship. It tells us how the quantity demanded for good 2 varies as the price of good 1 changes. The equation can be used to compute a cross price elasticity, like this:

$$\frac{dx_2^*}{dp_1} \cdot \frac{p_1}{x_2^*} = \frac{2.5}{p_1^2}\frac{p_1}{14 - \frac{2.5}{p_1}} = \frac{2.5}{p_1\left(14 - \frac{2.5}{p_1}\right)} = \frac{2.5}{p_1\left(\frac{14p_1 - 2.5}{p_1}\right)} = \frac{2.5}{14p_1 - 2.5}$$

Unlike the own price elasticity, the cross price elasticity is not constant—it depends on the value of p_1. It is also positive (whereas the own price elasticity was negative). When p_1 rises, optimal x_2 also rises. This means that goods 1 and 2 are *substitutes*.

Complements, on the other hand, are goods whose cross price elasticity is negative. This means that an increase in the price of good 1 leads to a decrease in consumption of good 2.

Demand can also be derived via numerical methods.

STEP Open the Excel workbook *DemandCurvesPractice.xls*, read the *Intro* sheet, then go to the *QuasilinearChoice* sheet.

The consumer is maximizing satisfaction at the initial parameter values because the marginal condition, MRS $= \frac{p_1}{p_2}$, is met at the point 6.25,12.75 (ignoring Solver's false precision) and income is exhausted.

We can explore how this initial optimal solution varies as the price of good 1 changes via numerical methods. We simply change p_1 repeatedly, running Solver at each price, while keeping track of the optimal solution at each price. The Comparative Statics Wizard add-in handles the tedious, cumbersome calculations and outputs the results in a new sheet for us.

STEP Run the Comparative Statics Wizard on the *QuasilinearChoice* sheet. Increase the price of good 1 by 0.1 (10 cent) increments.

You can check your comparative statics analysis by comparing your results to the *CS1* sheet, which is based on 1 (instead of 0.1) dollar size shocks. Of course, the numbers will not be exactly the same since the Δp_1 shock size is different.

The columns of price and optimal x_1 are points on the demand schedule. The numerical approach via the CSWiz essentially picks individual points on the demand curve for the given prices. If you plot these points, you have a graph of the demand curve.

The analytical approach, on the other hand, gives the demand function as an equation. You can evaluate the expression at particular prices and generate a plot of the demand curve.

The two approaches, if done correctly, will always yield the same graphical depiction of the demand curve. They may not, however, yield the same slopes or elasticities.

STEP Using your results, create demand and price consumption curves. Compute the own unit changes and elasticities for x_1* and x_2*.

The *CS1* sheet shows how to do this if you get stuck. You can click on cells to see their formulas. Think about how the formulas work and how they compute the answer.

It is critical that you notice that your own unit changes and elasticities are closer to the instantaneous rates of change in columns I and J of the *CS1*

sheet because you have smaller changes in p_1 and, for this utility function, ${x_1}^*$ is nonlinear in p_1.

Take a moment to reflect on what is going in the calculations presented in the *CS1* sheet. The color-shaded cells invite you to compare those cells.

Now, let's walk through this slowly.

STEP Click on cell F13 to see its formula.

It is computed as the change in optimal x_1 for a \$1 increase in p_1. There is a decrease of about 3.47 units when price increases by 1 unit.

STEP Click on cell I12 to see its formula.

It is computed by substituting the initial price, \$2/unit, into the expression for the derivative (displayed as an equation above the cell). The result of the formula, -6.25, is the instantaneous rate of change. In other words, there will be a 6.25-fold decrease in optimal x_1 given an infinitesimally small increase in p_1.

STEP Go to your CSWiz results and, if you have not done so already, compute the change in optimal x_1 for a \$0.1 increase in p_1.

You should find that your slope is about -5.8. The change in optimal x_1 is about 0.58, but you have to divide by the change in price, 0.1, to get the slope. Notice that your answer is much closer to the derivative-based rate of change (-6.25). This is because you took a much smaller change in price, 0.1, than the one dollar change in price in the *CS1* sheet and you are working with a curve.

STEP Return to the *CS1* sheet and compare cells G13 and J12.

The same principle is at work here. Because the demand curve is nonlinear, the two cells do not agree. Cell G13 is computing the elasticity from one point to another, whereas cell J12 is using the instantaneous rate of change (slope of the tangent line) at a point.

If you compute the price elasticity from 2 to 2.1 (using your CS results), you will find that it is much closer to -2.

Finally, you might notice that unlike the Cobb-Douglas utility function, which produced a horizontal price consumption curve (PCC), the quasilinear utility function in this case is generating a downward sloping price consumption curve. In fact, the slope of the price consumption curve tells you the price elasticity of demand: Upward sloping PCC means that demand is inelastic, horizontal PCC yields a unit elastic demand (as in the Cobb-Douglas case), and downward sloping PCC gives elastic demand (as in this case).

Perfect Complements

We begin with the analytical approach.

$$U(x_1, x_2) = min\{ax_1, bx_2\}$$

For $a = b = 1$, we know that we can find the intersection of the optimal choice and budget lines to get the reduced form expressions for the endogenous variables, ${x_1}^* = \frac{m}{p_1+p_2}$ (which is the same for ${x_2}^*$ since ${x_1}^* = {x_2}^*$).

This solution says that when a and b are the same in a perfect complements utility function, the optimal amounts of each good are equal and found by simply dividing income by the sum of the prices.

The reduced form expression contains Engel and demand curves. Holding prices constant, we can see how m affects consumption. Likewise, holding m and p_2 constant, we can explore how optimal x_1 varies as p_1 changes. This, of course, is a demand curve for x_1.

As usual, we find the instantaneous rate of change by taking the derivative with respect to p_1. The p_1 elasticity of x_1 is the derivative multiplied by $\frac{p_1}{{x_1}^*}$.

$$\frac{d{x_1}^*}{dp_1} = -\frac{m}{(p_1+p_2)^2}$$

$$\frac{d{x_1}^*}{dp_1} \cdot \frac{p_1}{{x_1}^*} = -\frac{m}{(p_1+p_2)^2} \frac{p_1}{\frac{m}{p_1+p_2}} = -\frac{p_1}{p_1+p_2}$$

We can also derive demand for a perfect complements utility function via numerical methods.

STEP Proceed to the *PerfCompChoice* sheet and run the Comparative Statics Wizard with an increase in the price of good 1 of 0.1 (10 cents).

Can you guess what we will do next? The procedure is the same every time: we solve the model then explore how the optimal solution responds to shocks.

STEP Create demand and price consumption curves based on your comparative statics results. Compute the own units changes and elasticities for x_1* and x_2*. The *CS2* sheet shows how to do this if you get stuck.

As before, you will want to concentrate on how your own units changes and elasticities are closer to the instantaneous rates of change than the Δp_1 in columns F and G of the *CS2* sheet because you have smaller changes in p_1 and we are dealing with a nonlinear relationship.

The lesson is clear: whenever the demand curve is not a line, that is, x_1* is nonlinear in p_1, then Δp_1 will not exactly equal dp_1. As the size of the discrete change in price gets smaller, the numerical method result will approach the result based on the derivative.

Although the two methods might not exactly agree, they are usually pretty close. How close depends on the curvature of the relationship and the size of the discrete shock. This means you can always check your analytical work by doing a manual Δ shock and computing the change from one point to another.

Notice also that the price consumption curve is upward sloping and the price elasticity is less than one (in absolute value).

Deriving Demand from the Consumer's Utility Maximization Problem

The primary purpose of this section was to provide additional practice in deriving demand with different utility functions. Clearly, the demand curve is strongly influenced by the utility function that is being maximized given a budget constraint.

Two examples were used to demonstrate how the analytical and numerical methods are related. Calculus is based on the idea of infinitesimally small changes. You can see calculus in action by using the CSWiz to take

smaller changes in price—which drives the numerical method ever closer to the derivative-based result.

Exercises

1. Return to the *QuasilinearChoice* sheet and click the Reset button. Now change the exponent on good 1 from 0.5 to 0.75. Use the Comparative Statics Wizard to derive a demand curve for this utility function.

2. Working with the same utility function as in the first question, derive the demand for x_1* via analytical methods. Use Word's Equation Editor as needed. Show your work.

3. Using your results from questions 1 and 2, compute the own price elasticity via numerical and analytical methods. Do they agree? Why or why not? Show your work and take screen shots as needed.

References

The epigraph is from page 63 of Hal Varian's best-selling, undergraduate textbook, *Intermediate Microeconomics* (7th edition, 2006). In the preface, Varian tackles head on the issue of calculus. "Many undergraduate majors in economics are students who should know calculus, but don't—at least not very well. For this reason, I have kept calculus out of the main body of the text."

The book you are reading at this moment takes a different approach. Calculus is used extensively, but it is made accessible by consistent repetition along with the substantial support of numerical methods. If you are a student who struggles with analytical methods, you will never have a better opportunity to master calculus and algebra. Do the practice problems with care and match the analytical and numerical approaches in each application.

To my knowledge, no one has described heroin as a Giffen good. But the description may be appropriate for those users who are addicted.

Neal Kumar Katyal

4.5 Giffen Goods

Demand curves are derived by doing comparative statics on the consumer's optimization problem: Change price, ceteris paribus, and track optimal consumption of a good.

In introductory economics courses around the world, demand is always drawn downward sloping so that as price rises, ceteris paribus, quantity demanded falls. Economists have long been intrigued, however, by a perplexing possibility: quantity demanded rising as price rises. An upward sloping demand curve! Can this happen? Yes, but it is quite rare and it took decades to figure it out.

We begin with a definition: *Giffen goods* are goods that have upward sloping demand curves. Giffen's connection to this counter intuitive demand relationship—price rises and you want to buy more?—is controversial.

Giffen and the Irish Potato Famine

The Great Irish Famine took place during 1845-1848.

> To put the disaster in proper perspective, the famine killed at least 12 percent of the population over a three-year period. Another 6-8 percent migrated to other countries. In terms of the percentage of population affected, the 1845-48 famine is one of the largest ever recorded. Other famines have killed more people in total because the affected populations were larger, not the percentage of exposure. For instance, the 30 million or more people who perished in the Chinese famine of 1958-62 were 5 percent or 6 percent of the population. (Rosen, 1999, p. S303)

Why did so many people die? This is a difficult question to answer comprehensively. The economics of famine are complicated. The proximate answer is that the Irish ate a *lot* of potatoes and a potato blight destroyed the food source. Rosen (1999, p. S303) says this:

> As difficult as it is to imagine today, on the eve of the famine, per capita consumption of potatoes is reliably estimated to have averaged 9 pounds (40-50 potatoes) per person per day (Bourke 1993). Diets were astonishingly concentrated on potatoes, especially in rural areas. Grain was grown in rural Ireland but was either sent to towns or exported abroad.

When blight wiped out the potato crop, why didn't the Irish eat something else or just import food? This is hard to understand. Books have been written on the subject. The *Biblio* sheet in *GiffenGoods.xls* has references. In fact, Amartya Sen won a Nobel Prize in Economics for his work on famine. It turns out that it is not simply a matter of too little food—amazingly, food can be just a few miles away and yet many people can be starving!

But our focus is on Giffen goods and the story picks up decades after the famine. Although there is no evidence that he ever said anything close to "price increase led to higher quantity demanded," Sir Robert Giffen (1837–1910) is credited with using the behavior of potato prices and quantities to state the claim that quantity demanded rose as prices rose.

Figure 4.13 shows Irish potato prices before, during and after the famine. Although consumption fell when price spiked in 1847 to more than double the 1846 price, somehow the legend grew that quantity demanded increased as prices rose in this time period. Thus, the Irish potato became the canonical example of a Giffen good—even though there is no evidence that price and quantity moved in the same direction.

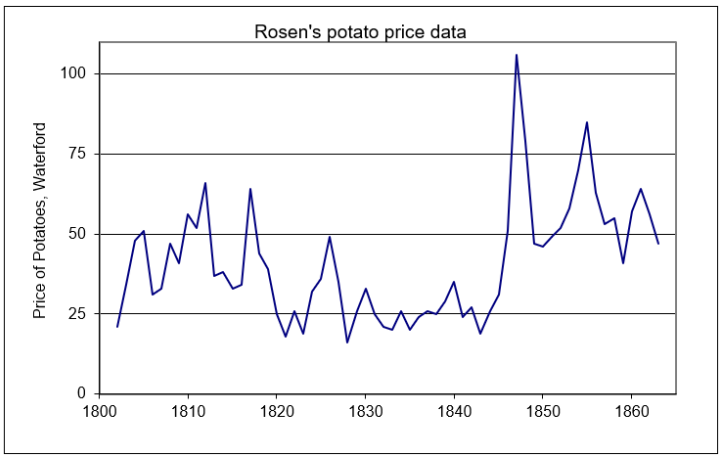


Figure 4.13: Potato price in Waterford, Ireland.
Source: OptimalChoice.xls!OptimalChoice

Economists began arguing over whether or not quantity demanded rose as the price spiked and, even if it did not, whether it was theoretically possible. It would take decades of contentious debate before the matter was settled.

Two Common Mistakes in the Giffen Debate

Before explaining how we could, in theory, get a Giffen good, we need to clear up two mistakes in thinking about Giffen goods. Both mistakes involve violating the strict ceteris paribus requirement that underlies a demand curve. The first mistake has a long history in econometrics and the second is easily corrected once we remember that we must hold everything else constant.

Estimating demand from observed prices and quantities is quite difficult. It turns out that plotting price and quantity data over time and fitting a line is no way to estimate a demand curve.

Suppose that the observed quantity of potatoes sold and consumed really had increased as the price spiked in 1847. Would that have been a good way to support the Giffen good claim? Absolutely not.

The problem is that the price and quantity data in different time periods do not fulfill the ceteris paribus requirement. It is true that price and quantity changed over time, but presumably so did other factors that affect demand and supply.

STEP Open the Excel workbook *GiffenGoods.xls* read the *Intro* sheet, then go to the *ID* sheet and read it carefully. Make sure to click the buttons and think about the charts that are displayed.

This sheet walks you through a simple example and shows why fitting a line to observed market price and quantity data is a really bad move. The heart of the confusion lies in the inability to extract the individual supply and demand curves that produce the observed data. This is called the *identification problem*.

So, even if it is true that we see prices and quantities moving together, that is not a demonstration of Giffen behavior.

The second mistake is less easy to forgive. No complicated issues of estimation are involved. We simply forget that demand requires that the ceteris

paribus condition hold. Suppose you notice that a particular brand of jeans has become increasingly popular and suddenly more people want it as its price rises. Have we discovered a Giffen good?

Absolutely not. We are violating the crucial ceteris paribus part of the definition of a demand curve by failing to hold constant everything except a change in price. In this case, the increased popularity of a particular brand is a shock to the demand curve, shifting it right. This is not a Giffen good because we are not working with a single, fixed demand curve. Instead, as in the second chart in the *ID* sheet, changes in demand are driving new equilibrium price–quantity combinations.

Having seen two common mistakes in trying to understand and show Giffen behavior, both involving violation of the strict ceteris paribus condition, the natural question then is: Can true Giffen goods, ones that meet the specific requirements of a demand function, exist? The answer is yes.

Giffen Goods in Theory

The left graph in Figure 4.14 shows the canonical graph of the Theory of Consumer Behavior displaying a Giffen good, while the right shows its associated upward sloping demand curve. Notice that the indifference curves require a little tweaking and somewhat odd placement to make x_1 be a Giffen good. Remember that indifference curves cannot cross, but they do not have to be similarly shaped and equally separated. For x_1 to be Giffen, point 2 in Figure 4.14 has to lie to the left of point 1 so that the decrease in p_1 leads to a decrease in optimal x_1.

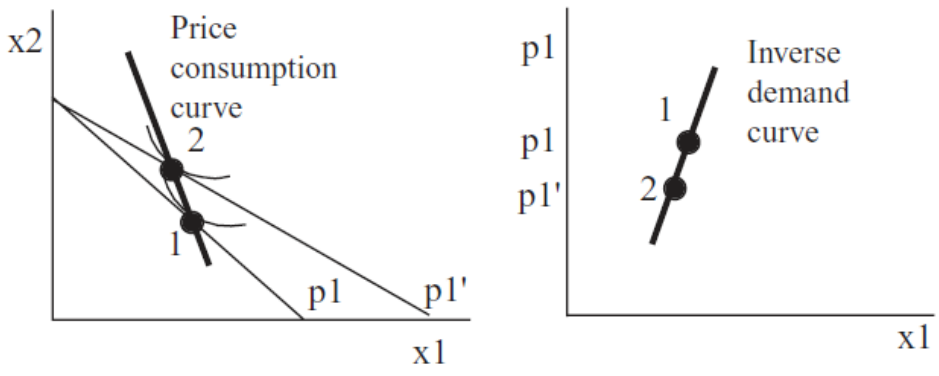


Figure 4.14: A Giffen good.

Do not be confused by the decrease in x_1. Quantity demanded fell, but so did price. Thus, we have a *positive relationship* between price and quantity demanded (they are moving together) and an upward sloping demand curve. This is a Giffen good.

To be crystal clear, it is not the fact that optimal x_1 decreased that tells us we have a Giffen good, but that it decreased as price fell. If we started at point 2 and raised the price, the budget constraint would swing in, and we would move to point 1, with an increase in optimal x_1. We would have *Giffenness* because x_1 rose as p_1 increased, We would be traveling up the upward sloping demand curve.

A version of Figure 4.14 is depicted in every microeconomics book that discusses Giffen goods and, make no mistake, this is a canonical graph in micro theory. But dead graphs on a printed page (or computer screen) force the reader to reconstruct individual elements and can be difficult to disentangle. With Excel at our disposal, we can walk through a numerical example to gain complete mastery of the concept of Giffenness.

STEP Proceed to the *Optimal1* sheet and look at the utility function.

The sheet models a Giffen good. The utility function is admittedly quite complicated, but a simple functional form like Cobb-Douglas or quasilinear is never going to produce Giffenness.

$$u(x_1, x_2) = \begin{cases} ax_1 - \frac{b}{2}x_1^2 + cx_2 + \frac{d}{2}x_2^2 & \text{for } 0 \le x_1 \le a/b \\ \frac{a^2}{2b} + cx_2 + \frac{d}{2}x_2^2 & \text{for } x_1 > a/b \end{cases}$$

The *U1* sheet shows that this functional form meets the requirements of well-behaved preferences. The coefficients have been set to values that do not violate the axioms of revealed preference in the range we are working in. The indifference curves, for example, will never intersect.

Another example of a utility function that exhibits Giffen behavior is $U = ax_1 + \ln x_1 + \frac{x_2^2}{2}$. This is implemented in the *Optimal2* sheet. We will use the *Optimal1* sheet here and save the *Optimal2* sheet for Q&A work. These are just two of the many functional forms that meet the requirements of well-behaved utility that could exhibit Giffen behavior.

The *Optimal1* sheet opens with $x_1 = 44$ and $x_2 = 11$. A single indifference curve is displayed and it does not have the curvature we have been used to seeing. Recall that perfect substitutes are straight lines, so we can infer that this utility function is expressing preferences with a high degree of substitutability between the two goods.

Without running Solver, we know this is the optimal solution because the MRS equals the price ratio.

STEP It is hard to see that the budget line is just touching the indifference curve, but if you click the Zoom In button, you will see that the tangency condition is clearly met.

Since we are working on Giffen behavior, we want to explore the effects of a change in price on the quantity demanded. We will increase the price of x_1 and see how the consumer responds. Before we do, think through what will happen. How will the constraint change and where must the new tangency point lie if x_1 is a Giffen good?

STEP Change p_1 to 1.1. What happens?

The budget line pivots around the y intercept. It may look like a parallel shift, but it really is not.

STEP Click the Zoom Out button to see that the price increase has, as expected, rotated the budget line in.

The 44,11 initial optimal bundle is no longer affordable. The consumer must re-optimize.

STEP Run Solver. What happens?

Figure 4.15 shows the result. Optimal consumption of good 2 has collapsed from 11 to around 1.5 and the consumer now wants to buy 48.6 units of good 1, which is more than the initial amount of 44.

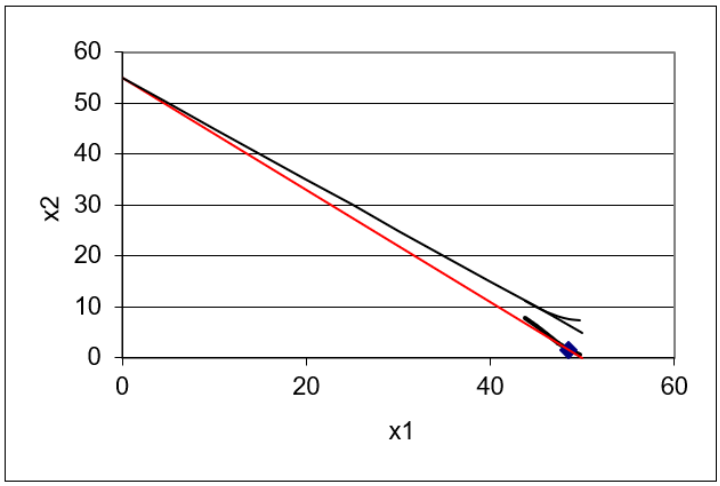


Figure 4.15: A numerical example of Giffen behavior.
Source: GiffenGoods.xls!Optimal1

This is amazing! The price of good 1 went *up* by 10 cents (from 1 to 1.1) and the optimal amount of good 1 *increased* by 4.6 units (from 44 to 48.6). Price rose, ceteris paribus, and so did quantity demanded!

This is a concrete, numerical example of a Giffen good. We can use the Comparative Statics Wizard to explore more carefully the demand curve resulting from this bizarre utility function.

STEP Use the Comparative Statics Wizard to trace the demand curve from 0.1 to 3. Set cell B16 to 0.1, then apply 300 (yes, 300) shocks by increments of 0.01 with the CSWiz add-in. Finally, create a graph of the inverse demand curve, p_1 as a function of x_1*.

Your results should look like Figure 4.16, which is also in the *CS1* sheet. That is certainly a strange looking demand curve. It is Giffen in a range. In other words, a Giffen good is not intrinsically and everywhere a Giffen good. Giffenness is a local phenomenon. The demand curve pictured in Figure 4.16 has three different behaviors. As price rises from zero, quantity demanded falls. This continues until a price of about 70 cents. From there, penny increases lead to increased consumption of good 1. In this range, x_1 is a Giffen good. There is a third region, at prices such as $2 and $3, where the good is not Giffen.

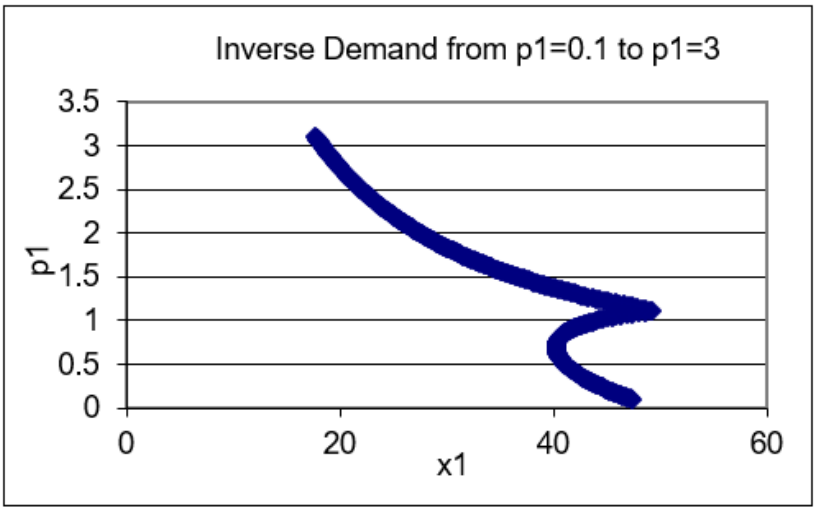


Figure 4.16: The inverse demand curve for x_1.
Source: GiffenGoods.xls!CS1

So, this example has shown that Giffen goods are not only possible, they can be modeled by the Theory of Consumer Behavior. We now know that there are utility functions that reflect well-behaved preferences that generate Giffen behavior.

Giffen Goods in Theory and Practice

A Giffen good is a strange creature in economics. The phenomenon of quantity demanded rising as price increases was first purportedly sighted during the Irish potato famine and named after Sir Robert Giffen, even though there is no evidence that Giffen actually claimed seeing quantity demanded rise as prices rose, ceteris paribus.

Certainly there are utility functions that give rise to Giffen goods. Certainly individual consumers may have well-behaved preferences that yield Giffen behavior. But has a Giffen good ever been spotted? Do Giffen goods exist in the real world in the sense that a market demand curve is upward sloping? This is the subject of much debate. Ceteris paribus is a difficult requirement to meet.

The actual sighting of a Giffen good in the real world remains contentious. We know for sure that the original example, potatoes during the Great Irish

Famine, was flawed and there is little evidence that it was a Giffen good. The *Biblio* sheet has a few references that can start you learning more about the history of Giffen goods in economics.

The next section gives an even deeper explanation for Giffen goods. It establishes the specific conditions needed for Giffenness to occur.

Exercises

1. Use the results in the *CS1* sheet to find the price range for which we see Giffen behavior. Report your answer and describe your procedure.

2. Use the *Optimal1* sheet utility function and parameter values to find the optimal solution via analytical methods. Show your work. Note that $x_1 < \frac{a}{b}$,so the utility function is

$$U = ax_1 - \tfrac{b}{2}x_1^2 + cx_2 + \tfrac{d}{2}x_2^2$$

3. Use Word's Drawing Tools to reproduce Figure 4.14, depicting x_1 as a Giffen good, but use a p_1 increase (instead of a decrease).

References

The epigraph comes from page 2436 of Neal Kumar Katyal, "Deterrence's Difficulty," *Michigan Law Review*, Vol. 95, No. 8. (August, 1997), pp. 2385–2476, repository.law.umich.edu/mlr/vol95/iss8/3/.

The *Biblio* sheet in *GiffenGoods.xls* has a list of references on Giffen goods. Scroll down to see suggested readings on the Irish potato famine, the history of Giffen goods in economics, and modern-day efforts at finding Giffen goods. Click on a link if anything catches your eye and seems worth exploring.

Eugene (or Eugen or Yevgeni) Slutsky [1880 – 1948] intended to become a mathematician, but he was expelled from the University of Kiev for participating in student revolts.

Gonçalo L. Fonseca

4.6 Income and Substitution Effects

Without a doubt, the demand curve is the most important idea in the Theory of Consumer Behavior. We have derived the demand curve analytically and numerically. The demand curve tells us the optimal amount to buy at a given price. It also tells us how quantity demanded will change as price changes, ceteris paribus.

This section remains focused on the demand curve, extending the analysis of the consumer's optimal response to a change in price. The core concept is that the total effect on quantity demanded (given by the demand curve) for a given change in price can be broken down into two separate effects, called *income and substitution effects*.

Our attention is still on the change in quantity demanded as price changes, ceteris paribus, but by breaking apart the observed response when price changes, we get a deeper explanation of demand. We also explain how we might get a Giffen good.

Intuition

Before diving into complicated graphs and math, let's review the story behind income and substitution effects. Seeing the big picture improves your chances of really understanding what income and substitution effects are all about.

Suppose that, ceteris paribus, price rises. We know the consumer has to re-optimize. We know the consumer will choose a new optimal combination of goods. We can see the consumer buy a different amount after the price changes. If we simply compute the change in the amount purchased of x_1 before and after the price change, we are comparing two points on the demand curve. This is called the *total effect* of a price change.

The breakthrough idea is that the increase in price has two channels by which it affects the consumer. One channel focuses on the fact that a price increase is like a decrease in purchasing power. After all, given an income level, if prices double, then I can buy half of what I bought before. My income has not changed, but my purchasing power has fallen just the same as if my income had been cut in half. The *income effect* reflects the fact that price changes affect optimal quantity demanded by altering purchasing power.

The other channel is called the *substitution effect*. The idea is that a price change in one good alters the relative prices faced by the consumer and induces substitution of the relatively cheaper good for the relatively more expensive one. When p_1 rises, x_1 is relatively more expensive than x_2 and so I am naturally going to avoid x_1 and be attracted to x_2.

Figure 4.17 shows the two channels below the total effect—they are submerged and not directly observed. Added together, they make up the total effect.

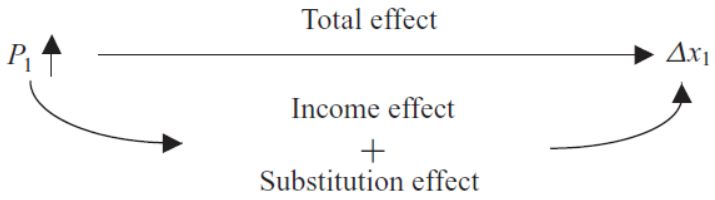


Figure 4.17: The basic idea behind income and substitution effects.

We will see that the income effect can be either positive or negative, but the substitution effect is *always negative* (assuming well-behaved preferences). When price goes up, the substitution effect says "buy less." Of course, if price falls, the reverse occurs and, according to the substitution effect alone, consumption increases.

The reason the income effect is ambiguous in sign is the fact that there are normal and inferior goods. If the good is normal, then optimal x_1 rises as income increases, but if the good is inferior, then consumption and income are inversely related.

Finally, it helps to know the underlying motivation behind the discovery of income and substitution effects. Economists were arguing about the existence of Giffen goods. The Law of Demand said price and quantity were inversely

related. Income and substitution effects explained under which conditions Giffen behavior (an upward sloping demand curve) is possible. We will see that if the income and substitution effects work together, then the demand curve is guaranteed to be downward sloping. Understanding income and substitution effects will allow us to give a more refined, precise definition of the Law of Demand.

Numerical Example of Income and Substitution Effects

STEP Open the Excel workbook *IncSubEffects.xls*, read the *Intro* sheet, and proceed to the *OptimalChoice* sheet.

We have the usual Cobb-Douglas utility function with a conventional budget line. We have done this problem before and the initial optimal solution is $25, 16\frac{2}{3}$.

STEP Decrease p_1 by 1 to \$1/unit (in cell B17).

Figure 4.18 displays what is on your screen. The red line is the familiar new budget line (after the price decrease). There is, however, a dashed line that has not been used before. This dashed line represents the outcome of a thought experiment.

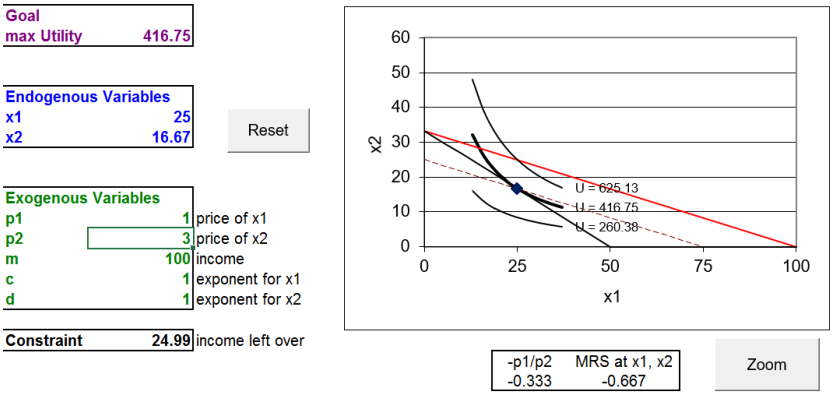


Figure 4.18: Decreasing p_1.
Source: IncSubEffects.xls!OptimalChoice

STEP Click the Zoom button to see a second graph of the situation. It has the axes scale adjusted so you can see better what is going on.

The dashed line is critical to understanding the splitting of the total effect into income and substitution effects. It has the same slope as the new budget line, yet it goes through the initial optimal solution. What we have done is pretend to take away enough income from the consumer to enable him to buy the initial bundle with the new, lower p_1.

We took away income (shifting down the budget constraint relative to the new budget line) because the fall in price implies an increase in purchasing power. Had there been a price rise, we would have had to increase income to compensate for the price increase.

We will find a tangency solution on the dashed line and this will allow us to split the total effect into the income and substitution effects.

Of course, nothing like this actually happens in the real world. When the price falls, the consumer re-optimizes, buying a new optimal bundle, and that is the end of the story. But for the purposes of understanding the demand curve, we figure out what the consumer would buy at the imaginary dashed line and we use that to split the total effect into the substitution and income effects.

But this is all way too abstract. Let's actually do it so you can see how it works. To figure out how much income to take away to cancel out the changed purchasing power from the price change, we use the *Income Adjuster Equation*.

$$\Delta m = x_1{}^* \Delta p_1$$

Applied to this problem, we know that $x_1{}^*$ is 25 (from the initial optimal solution) and the change in p_1 is -1 (because the price fell from 2 to 1, so $new - initial$ is $1 - 2$); thus, we have:

$$\Delta m = x_1{}^* \Delta p_1$$

$$\Delta m = [25][-1] = -25$$

The minus tells us that we have to take away income. The dashed line is based on an income of \$75, $p_1 = 1$, and $p_2 = 3$.

In summary, we have three budget lines when we work with income and substitution effects: (1) the usual initial line, (2) the usual new line from the change in price, and (3) the imaginary (dashed) line that has been adjusted to pass through the initial optimal solution.

We find the usual new optimal solution so we can compute the total effect first, then we use the dashed line to find the income and substitution effects.

STEP With $p_1 = 1$, run Solver.

Figure 4.19 shows that the consumer chooses the $50{,}16\frac{2}{3}$ combination. Thus, we have two points to consider so far:

- Point A: Initial: At $m = 100, p_1 = 2, {x_1}^* = 25, {x_1}^* = 16\frac{2}{3}$.
- Point C: New: At $m = 100, p_1 = 1, {x_1}^* = 50, {x_1}^* = 16\frac{2}{3}$.

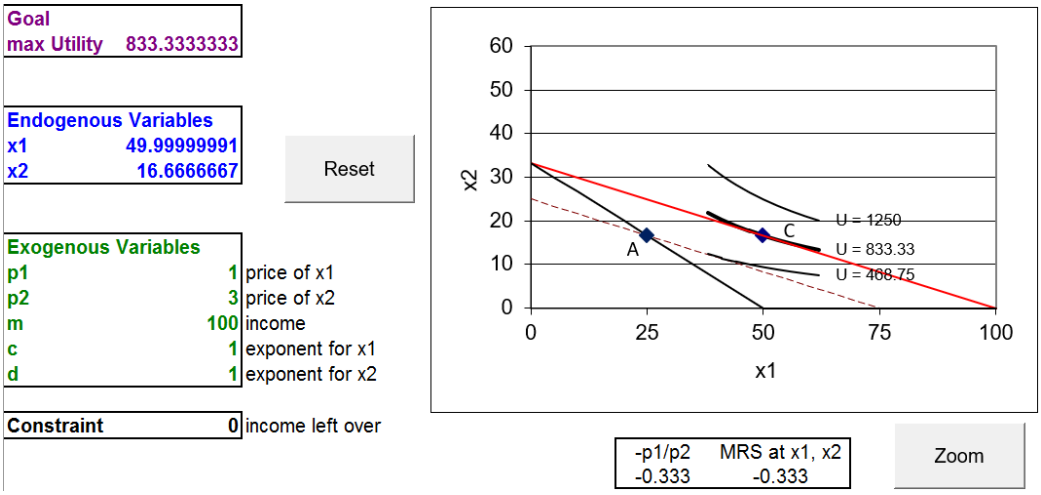


Figure 4.19: New optimal solution at $p_1 = 1$.
Source: IncSubEffects.xls!OptimalChoice

Notice that Excel displays three difference curves around the current optimal solution, but there are actually an infinite number of curves going through every point in the quadrant. With $c = d = 1$ being held constant, the indifference map is not changing in any way. We are simply displaying different indifference curves whenever x_1 and x_2 in cells B12 and B13 change.

Points A and C are two points on the price consumption curve and two points on the demand curve. The total effect of a \$1/unit decrease in the price of good 1 can be found by measuring the movement from A to C: for x_1, the total effect is +25 units and for x_2, the total effect is zero (${x_2}^* = 16\frac{2}{3}$ before and after the price shock).

The total effect can be directly observed. With the initial price, we can see the consumer purchase 25 units of good 1 and $16\frac{2}{3}$ of good 2. We see the price of good 1 fall by \$1/unit and watch the consumer respond by buying 25 units more of x_1 and leaving the amount of x_2 unchanged.

We are now ready for the key move. We will hypothetically take away exactly \$25 of income so we can find the optimal solution on the imaginary, dashed line. The consumer does not actually have income taken away. It is a thought experiment. Working out what the consumer would do in this hypothetical situation allows us to split the total effect into its constituent parts.

STEP Change income to \$75 (notice that the budget line now lies on top of the dashed budget line) and run Solver.

You can safely ignore the steeper line in the chart—all we want is point B, the optimal solution with the dashed budget line. Solver tells us that point B is 37.5,12.5. This gives us three points to consider:

- Point A: Initial: At $m = 100, p_1 = 2, {x_1}^* = 25, {x_2}^* = 16\frac{2}{3}$.
- Point B: Unobserved: At $m = 75, p_1 = 1, {x_1}^* = 37\frac{1}{2}, {x_2}^* = 12\frac{1}{2}$.
- Point C: New: At $m = 100, p_1 = 1, {x_1}^* = 50, {x_2}^* = 16\frac{2}{3}$.

Look carefully at the three points and concentrate on how points B and C differ: C uses new p_1 with original m, while B is based on new p_1 with adjusted m (adjusted in a special way so that the dashed line goes through point A).

With these three points, we can compute total, income, and substitution effects for x_1 and x_2. The three effects are shown by arrows on the axes of Figure 4.20. This is a complicated graph. Take your time and read it with care. Try to separate the different elements and lines to different parts of the problem: initial (A), new (C), and intermediate positions (B).

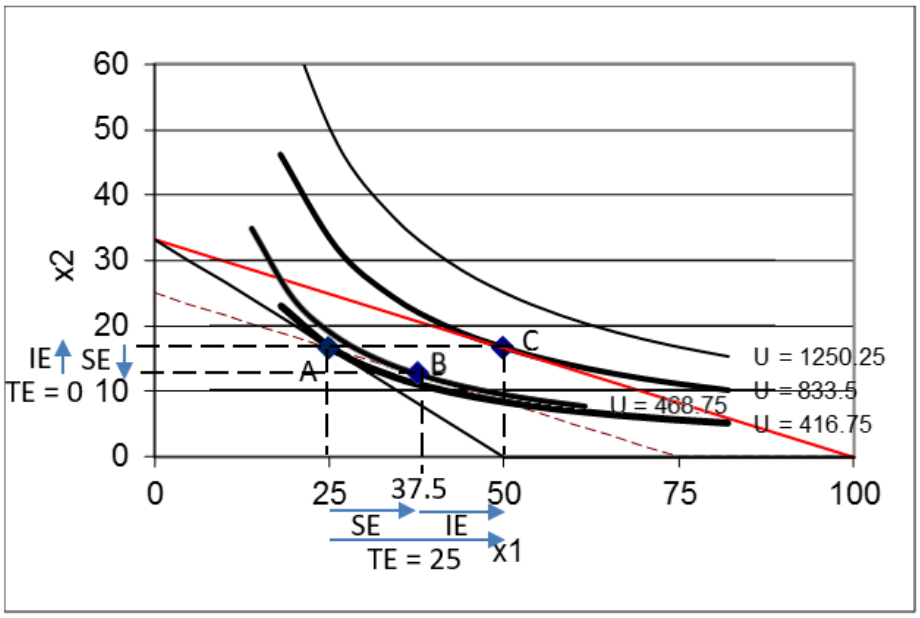


Figure 4.20: Total (TE), income (IE), and substitution (SE) effects.

There are effects measured from one point to another for both x_1 and x_2. These Δs are calculated the usual way as $new - initial$. For x_1, we find:

- SE: A to B: $37\frac{1}{2} - 25 = 12\frac{1}{2}$
- IE: B to C: $50 - 37\frac{1}{2} = 12\frac{1}{2}$
- TE: A to C: $50 - 25 = 25$

Notice that the total effect (TE) can be found by computing the difference from A to C ($50 - 25 = 25$) or taking advantage of the fact that SE + IE = TE, so $12.5 + 12.5 = 25$. The effects for x_1 are all computed along the x axis in terms of units of x_1.

Analyzing the effect on x_2 of a change in p_1 gives us cross income and substitution effects for x_2, which are shown by arrows on the y axis, in Figure 4.20.

- SE: A to B: $12\frac{1}{2} - 16\frac{2}{3} = -4\frac{1}{6}$
- IE: B to C: $16\frac{2}{3} - 12\frac{1}{2} = 4\frac{1}{6}$
- TE: A to C: $16\frac{2}{3} - 16\frac{2}{3} = 0$

On x_2, the income and substitution effects work against each other. The substitution effect, from A to B, lowers the amount of x_2 since p_1 fell, making x_2 more expensive relative to x_1. But when we move from B to C, the income effect exactly cancels out the SE. The fall in p_1 has increased our purchasing power and, since x_2 is a normal good, we want to buy more of it.

It is a property of the Cobb-Douglas utility function that the cross IE and SE effects cancel each other out, leaving a zero total effect. This is not a usual or common result and it demonstrates how the functional form imposes structure on the demand curve.

Let's return now to x_1 and focus on its substitution effect, which we know is always negative. This leads immediately to a question: If the SE is always negative, then why is it +12.5 in Figure 4.20?

The answer to this apparent contradiction is that the negative refers to the relationship, not the actual value of the SE. Given that price fell, an increase in quantity purchased is consistent with a negative effect because it is the relationship between the two variables that is being described as negative.

Likewise, the sign of the income effect can be tricky. The key is to pay attention to which shock variable is being considered. The income effect measured as the response to a change in income is positive, in this case, because as I move from B to C, my income is increased and I respond by increasing my optimal consumption of good 1.

Now you might ask, "If the two effects work together, then how is the substitution effect negative and the income effect positive?" This is because we defined the income effect as the response to a change in income, like the movement from point B to C in Figure 4.20. But, if you remember, this example began with a decrease in the price of good 1. The decrease in the price of good 1 can be interpreted as an increase in income, in the sense of greater purchasing power. If we tie the 12.5 increase in good 1 from the income effect to the *decrease* in price of good 1, we see that this negative relationship reinforces the negative substitution effect and gives a negative total effect.

Now that we know how the income and substitution effects combine to form the total effect of a price change, we can show how easy it is to compute them from a reduced form solution.

We first have to solve the model analytically and get a reduced form expression as a function of m and p_1. We have done this before for a Cobb-Douglas utility function and found

$$x_1{}^* = (\frac{c}{c+d})\frac{m}{p_1}$$

If we substitute in $c = d = 1$, we have

$$x_1{}^* = \frac{m}{2p_1}$$

At $m = 100$ and $p_1 = 2$, $x_1{}^* = 25$. This is the initial solution (point A).

If p_1 falls to \$1/unit, then we plug in $m = 100$ and $p_1 = 1$, which gives the new solution (point C), $x_1{}^* = 50$. The total effect is $50 - 25 = 25$.

To find the SE, we need point B. We use the reduced form expression to compute quantity demanded with adjusted m (\$75) and new p_1 (\$1/unit).

$$x_1{}^* = \frac{m}{2p_1} = \frac{[75]}{2[1]} = 37.5$$

Once we have point B, we have split the total effect from A to C and we can compute the SE and IE by going from A to B and B to C, respectively. The SE is $37.5 - 25 = 12.5$ and the IE is $50 - 37.5 = 12.5$. These results agree with our earlier work.

Income and Substitution Effects via Graphs

Income and substitution effects are complicated. Figure 4.20 is not easy to understand. There are three budget lines and a lot going on. So what is so important about income and substitution effects that makes it worthwhile to master them?

Income and substitution effects hold the key to explaining how we can get a Giffen good. They mark real progress in economics, settling a long debate about whether or not upward sloping demand curves are possible. We will deconstruct the income and substitution effect graph (Figure 4.20), examining each layer one at a time, to show the source of Giffen behavior.

We begin with Figure 4.21. On the left we have the initial optimal solution and the right displays a single point on the demand curve (not shown).

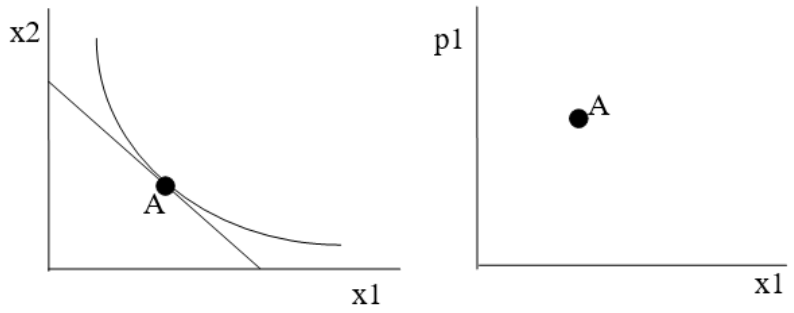


Figure 4.21: The initial solution.

Next, we decrease the price of good 1, as shown in Figure 4.22, which creates a new budget line. We know the consumer will re-optimize and choose a new optimal solution along the new, flatter line, but Figure 4.22 does not show this new solution quite yet. Instead, it shows the point B solution on a dashed line with the income that would have to be taken away to cancel out the increased purchasing power from the price decrease.

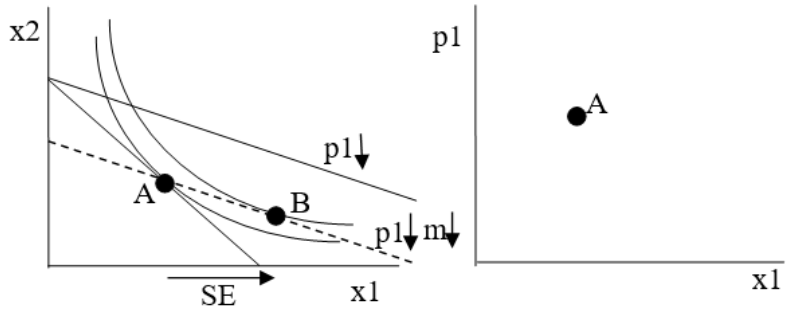


Figure 4.22: A p_1 decrease and imaginary budget constraint.

Figure 4.22 shows the optimal solution, point B, for the hypothetical situation with lower p_1 and adjusted m. The rightward pointing arrow is the SE for x_1 is the substitution effect, from point A to B on the x axis. The dashed line has a flatter slope (new p_1 is less than initial p_1) through point A. This guarantees that B is to the *right* of A. This is why the SE is always negative.

It is impossible to draw a point B to the left of A without making the indifference curves cross. With MRS $= \frac{p_1}{p_2}$ at A, lowering p_1 and adjusting m so dashed line goes through A, means the consumer must move southeast to find the highest indifference curve tangent to the dashed line.

Now, we are ready to show point C. We have a known negative substitution effect and all that remains to be done is to find the indifference curve tangent to the new budget line (with lower p_1). The key insight is that there are several possible positions for point C. Figure 4.23 shows three possibilities.

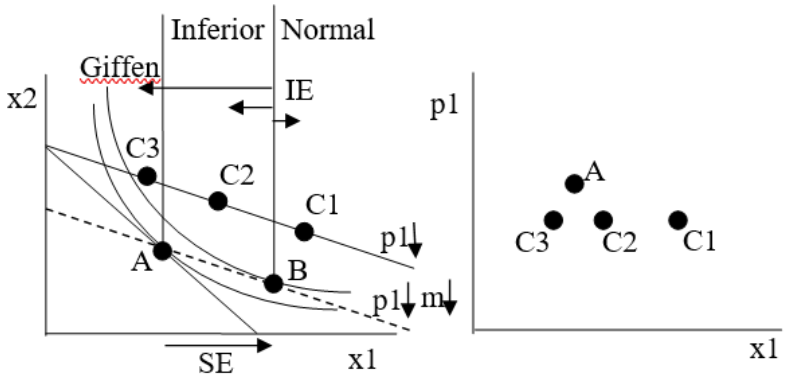


Figure 4.23: Understanding Giffen behavior.

Figure 4.23 shows that the final position of point C depends on whether the good is normal or inferior, with a subcategory of inferior goods that are Giffen.

- C1: Good 1 is a normal good so the income effect from B to C works together with the movement from A to B and we end up at point C1. In this case, and for any point C to the right of B, we get a downward sloping demand curve.

- Good 1 is an inferior good so the income and substitution effects work against each other. The movement from B to C will be to the left and leave us with a point C to the left of B. There are two possibilities:

    1. C2: The income effect pushes the consumer to buy less x_1, but it is less than the substitution effect (which leads to buying more x_1 as p_1 falls). We end up at point C2 between A and B and the demand curve is still downward sloping.
    2. C3: The income effect not only works against the substitution effect, it is stronger, swamping it. Point B to C moves in the opposite direction than A to B and and is bigger than A to B. This leaves the consumer to the left of B at point C3. The demand curve is upward sloping. This is a Giffen good.

It can be difficult to draw a Giffen good correctly because the indifference curves cannot cross. So, in Figure 4.23, the space available for point C3 is tight—C3 can only fit to the left of A and to the right of the indifference curve that is shown tangent to B.

Figure 4.23 also makes clear that it is the indifference curves, which come from the utility function, that determine how quantity demanded responds to a change in price. How a good generates utility (i.e., whether utility is Cobb-Douglas, quasilinear, perfect complements, or another functional form) determines whether it is normal, inferior, or Giffen.

The decomposition of the total effect into income and substitution effects provides the condition which must hold for Giffen behavior: the income effect must work against the substitution effect and be bigger. We can reinforce this key insight with a mathematical expression that gives more detail on exactly how we get Giffenness.

The Slutsky Equation

In 1915, decades after the supposed spotting of a Giffen good during the Irish potato famine, Eugen Slutsky published a paper in an Italian journal that showed how to decompose the total effect of a price change into income and substitution effects. He had a mathematical expression that showed how it was possible to get an upward sloping demand curve!

Unfortunately, his work went unnoticed. Twenty years later, John R. Hicks (a Nobel laureate in 1972) and R. G. D. Allen rediscovered the ideas in Slutsky's paper. Sometimes, the idea of income and substitution effects are referred to as Slutsky-Hicks or Slutsky-Hicks-Allen. We will keep it simple and call it the Slutsky Equation.

The Slutsky Equation, which we will not derive, says in mathematical terms something that we already know: The total effect of a price change can be expressed as the sum of a substitution and an income effect. It turns out that there are several ways to express the decomposition with a Slutsky Equation. Here are two versions:

$$\frac{\Delta x_1}{\Delta p_1} = \frac{\Delta x_1^{SE}}{\Delta p_1} + \frac{\Delta x_1^{IE}}{\Delta p_1}$$

$$\frac{\Delta x_1}{\Delta p_1} = \frac{\Delta x_1^{SE}}{\Delta p_1} - x_1 * \frac{\Delta x_1}{\Delta m}$$

Both equations say the same thing: the total effect, $\frac{\Delta x_1}{\Delta p_1}$, is equal to the substitution effect, $\frac{\Delta x_1^{SE}}{\Delta p_1}$, plus the income effect. Where they differ is how they express the income effect.

Look carefully at the denominators. The income effect in the first equation has a Δp_1 denominator, like the other two terms. What Slutsky figured out was that the income effect of price change, $\frac{\Delta x_1^{IE}}{\Delta p_1}$, could be written as $-x_1 * \frac{\Delta x_1}{\Delta m}$. In other words, the income effect channel of the price change can be expressed as the amount of good 1 initially purchased times the change in x_1 as income changes (the slope of the Engel curve). Notice the minus sign, which picks up the fact that when price falls, that is like an increase in income.

Now we can really see how to get a Giffen good, which has an upward sloping demand curve so $\frac{\Delta x_1}{\Delta p_1} > 0$. Since the first term, the substitution effect is always negative, we definitely need an inferior good so that $\frac{\Delta x_1}{\Delta m} < 0$ so that the second term is positive. Obviously, if the good is extremely inferior, so that $\frac{\Delta x_1}{\Delta m}$ is much less than zero, we might get a Giffen good.

But the Slutsky Equation reveals another way to get Giffen behavior. A large opposing income effect can be obtained by the good being inferior and the consumer buying a lot of it so that $-x_1 * \frac{\Delta x_1}{\Delta m}$ is a big positive number to outweigh the negative substitution effect. If the good is merely inferior, but the consumer buys little of it, then it less likely to be Giffen.

This is why we look for Giffen behavior in *staples*, basic commodities that comprise a large share of the budget. Potatoes for the Irish, rice for Asians, and tortillas for Mexicans are three examples that economists have examined for Giffen behavior. For a poor person, these items could be consumed in large quantities, yet, as income rises, quantity demanded falls so they are inferior goods. The combination of a large $x_1{}^*$ and $\frac{\Delta x_1}{\Delta m} < 0$ could produce a large, positive $-x_1 * \frac{\Delta x_1}{\Delta m}$ term that is bigger than the negative substitution effect.

Remember how we generated Giffen behavior with *GiffenGoods.xls* in the previous section? We increased the price from \$1/unit to \$1.1/unit and optimal x_1 rose from 44 to 48.6, while optimal x_2 fell dramatically from 11 to around 1.5. Notice how x_1 is a staple, dominating the amounts purchased of the two goods.

We know its Giffen, but is x_1 also inferior? Let's find out.

STEP Open *GiffenGoods.xls* and proceed to the *Optimal1* sheet. Click the Reset button and run Solver to make sure you are at the optimal initial solution of 44,11. Increase m to 60 and run Solver. What happens?

Yes, as we know must be true (since we know x_1 is a Giffen good), x_1 is an inferior good: optimal x_1 fell (to 39) as income increased to \$60. Giffenness requires that x_1 be inferior and this example also reflects the fact that concentration of the consumer's budget on an inferior good contributes to the production of a Giffen response.

The *Biblio* sheet in *GiffenGoods.xls*, from the previous section, had several references to papers trying to find Giffen goods, yet the jury is still out. What is unquestioned, however, is the theoretical requirement: it must be an inferior good so that the IE is in the opposite direction and larger than the SE.

The Slutsky Equation also enables us to fine tune a statement that is, strictly speaking, false. Introductory economics students around the world learn the *Law of Demand*: when price increases, ceteris paribus, quantity demanded must fall. In other words, holding everything else constant, quantity demanded and price are inversely related and demand is always downward sloping.

This is fine, at the introductory level, where we do not want to confuse beginning students, but we know that an upward sloping demand curve is possible—it is called a Giffen good. They are a violation of the "Law" of Demand and we know they could exist. When their price rises, so does quantity demanded.

Can we rehabilitate the Law of Demand so there is no exception? Yes, we can. Our knowledge of income and substitution effects points the way. We can more precisely define the Law of Demand. By inserting a qualifying clause, we can get the Law of Demand to be exactly right: *If the good is normal, then quantity demanded falls as price rises, ceteris paribus.* That is guaranteed to be true because a normal good has an income effect that works together with the substitution effect. Thus, there is no way to get Giffenness.

The Cobb-Douglas utility function cannot give Giffen behavior. The reduced form solution, ${x_1}^* = (\frac{c}{c+d})\frac{m}{p_1}$, means that $\frac{d{x_1}^*}{dm} = (\frac{c}{c+d})\frac{1}{p_1} > 0$ so the income

effect, $-x_1{}^* \frac{dx_1{}^*}{dm}$, is negative. This means the IE and SE are both negative and work together so there is no way the Cobb-Douglas utility function can generate Giffenness.

TE = SE + IE

Income and substitution effects are used by economists to better understand the demand curve and to explain Giffen behavior. By disassembling the total effect of a price change, the Slutsky Equation shows how a Giffen good can arise if the income effect opposes and swamps the substitution effect (which generates an upward sloping relationship between price and quantity demanded).

Given a utility function and budget constraint, we find the initial optimal solution (point A). A price change will lead to a new optimal solution (point C) which we can use to compute the total effect. We can then use the Income Adjuster Equation to find a hypothetical point B that splits the total effect into substitution and income effects.

Given a reduced form expression of $x^* = f(p, m)$, we can find points A, B, and C by evaluating the expression at the appropriate p and m values to compute points A, B, and C.

The Slutsky Equation is a mathematical presentation of income and substitution effects. The math gives us the insight that the income effect, $-x_1{}^* \frac{\Delta x_1}{\Delta m}$, is composed of initial optimal x_1 times the response of x_1 to an income change. This reveals that Giffenness is more likely to be found in inferior goods that also attract a high concentration of the consumer's budget.

There are even more ways to express the Slutsky Equation than the two used in this section. Instead of altering income to allow the consumer to buy the initial bundle of goods, you can change income to allow the consumer to be on the initial indifference curve. This is sometimes referred to as the Hicks substitution effect.

Exercises

1. Reproduce, using Word's Drawing Tools, Figures 4.21, 4.22, and 4.23, explaining each graph in your own words.

2. Repeat question 1, with one key change: apply a price *increase* in good 1 (instead of a price decrease).

3. In stating the Law of Demand, some economists choose to include a condition that the good is normal, like this: If the good is a normal good, then price and quantity demanded are inversely related, ceteris paribus. Why is the normal good clause needed?

4. Given the demand function, $x_1{}^* = 20 + \frac{m}{20p_1}$, compute the total, income, and substitution effects when price falls from \$5 to \$4/unit, with income of \$1000. Show your work.

5. Use the *Optimal1* sheet in *GiffenGoods.xls* to find points A, B, and C for a shock in p_1 from \$1 to \$1.1/unit. Compute the TE, SE, and IE for $x1$. Show your work and explain what you did.

References

The epigraph is from the biography of Slutsky available at the New School's History of Economic Thought website, www.hetwebsite.net/het/. The site was created and is maintained by Gonçalo L. Fonseca. There are sketches of hundreds of economists, links to other resources, and descriptions of various schools of thought in economics. The intellectual history of economics is fascinating and this website is a wonderful place to browse.

I never saw Slutsky's work until my own was very far advanced . . . Slutsky's work is highly mathematical, and he does not give much discussion about the significance of his theory.

J. R. Hicks

4.7 More Practice with IE and SE

This chapter uses a quasilinear utility function to provide practice working with income and substitution effects. There is a surprising twist when using the quasilinear functional form. See how fast you can figure it out.

STEP Open the Excel workbook *IncSubEffectsPractice.xls*, read the *Intro* sheet, then go to the *OptimalChoice* sheet.

Notice that the absolute value of the MRS is less than the price ratio. Because the slope of indifference curve at 16.25,10.75 is less than the slope of the budget constraint, we know the consumer should travel northwest along the budget constraint, buying more x_2 and less x_1, until the MRS $= \frac{p_1}{p_2}$.

STEP Run Solver to find the initial optimal solution. Figure 4.24 shows this result.

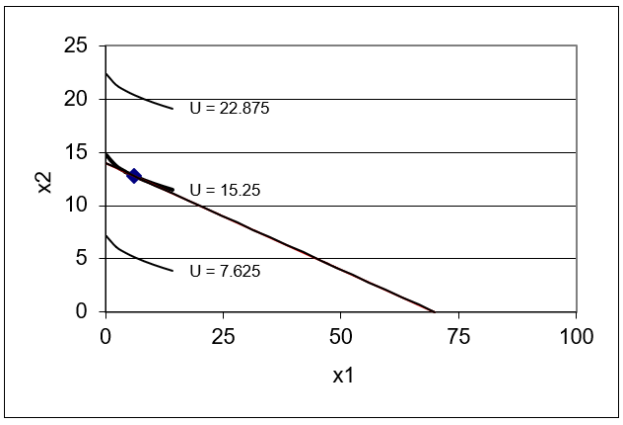


Figure 4.24: Initial optimal solution.
Source: IncSubEffectsPractice.xls!OptimalChoice

STEP Proceed to the *CS1* sheet. It shows a comparative statics analysis of an increase in the price of good 1 from \$2/unit to \$7/unit in \$1 increments. It also charts the results as an inverse demand curve for x_1.

The demand curve tracks the total effect of a price change. When the price of good 1 rises from \$2 to \$3, the quantity demanded falls from $6\frac{1}{4}$ to $2\frac{7}{9}$. By subtracting the new from the initial value, we see that the total effect is a decrease of $3\frac{17}{36}$ units of x_1, displayed in cell F13 as -3.47222.

Income and substitution effects explain how this total effect came to be by dismantling the total effect into two parts that add up to the total.

The substitution effect tells us how much less the consumer would have purchased when price rises strictly from the fact that the relative prices of the two goods have changed. We compute how much income we have to give the consumer to cancel out the reduced purchasing power caused by the price increase to focus exclusively on the relative price change. The substitution effect is always negative.

Figure 4.25 shows a typical decomposition of the total effect (TE) into the substitution effect (SE) and income effect (IE) with indifference curves suppressed to highlight the budget lines under consideration.

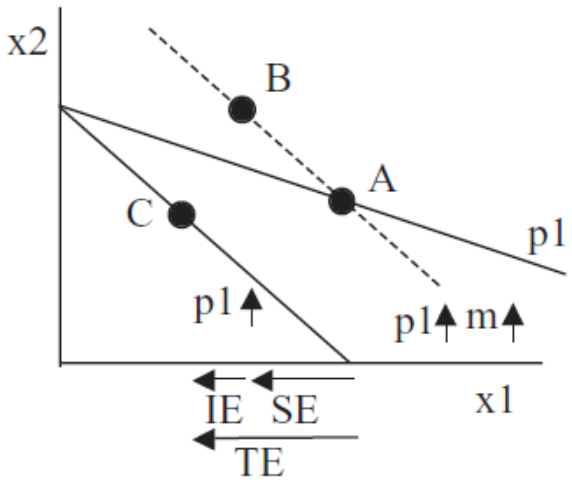


Figure 4.25: Typical TE, SE, and IE with p_1 increase.

From point A, price rose and the consumer will now be at point C on the new budget line (labeled $p_1 \uparrow$). The dashed line is the result of a hypothetical

scenario in which the consumer has been given enough income to purchase the initial bundle A. Notice how the original budget line and the dashed line go through point A. The dashed line has a higher price, but also a higher income. Thus, the movement from point A to point B reflects solely the different relative prices in the goods, without any change in purchasing power. This is the substitution effect.

While the substitution effect is focused on relative prices, the income effect is that part of the response in quantity demanded when price changes that is due to changed purchasing power. From point B, a decrease in income from the dashed to the new budget line leads to a decrease in x_1 (at point C). Thus, x_1 is a normal good from point B to C in Figure 4.25 and the two effects are working in tandem. The demand curve is guaranteed to be downward sloping for this price change.

In the *CS1* sheet, we have seen that the demand curve is downward sloping because quantity demanded falls when price rises. But an open question still remains: Do the income and substitution effects work as in Figure 4.25?

We know point A, the initial optimal solution, is $x_1{}^* = 6.25$ when $p_1 =$ \$2/unit and point C is about 2.78 units of x_1 when price rises to \$3/unit. We need point B to do the income and substitution effects analysis.

The first step in finding point B is to use the Income Adjuster Equation to compute how much income to give the consumer in order to cancel out the effect of the reduced purchasing power.

$$\Delta m = x_1{}^* \Delta p_1$$

$$\Delta m = [6.25][+1]$$

STEP On the *OptimalChoice* sheet, set cell B16 to 3.

The chart updates, showing the new budget constraint in red (swinging in since price rose) and the dashed line. To find point B, we need the optimal solution for the dashed line constraint so we need to change in income on the sheet.

STEP Set cell B18 to 146.25. This applies the dashed line budget constraint to this problem. Run Solver to find point B.

Your result might surprise you. Solver says the optimal solution is about 2.78 for x_1, but that is the same answer we had for point C. What is going on here?

We turn to analytical work to shed light on this mysterious result. Following the procedure in section 3.2, we found this reduced form solution for the quasilinear utility function, $U = x_1^c + x_2$:

$$x_1{}^* = (\frac{p_1}{cp_2})^{\frac{1}{c-1}}$$

We use the initial values of c and p_2 in the *OptimalChoice* sheet to simplify things a bit:

$$x_1{}^* = (\frac{p_1}{[0.5][10]})^{\frac{1}{[0.5]-1}} = (\frac{p_1}{5})^{\frac{1}{-0.5}} = (\frac{p_1}{5})^{-2} = (\frac{5}{p_1})^2 = \frac{25}{p_1^2}$$

This is the same kind of expression, $x_1{}^* = f(p_1, m)$, that we used in the previous section for a Cobb-Douglas utility function, $x_1{}^* = \frac{m}{2p_1}$, to find points A, B, and C.

You might be puzzled. Exactly where is m for the quasilinear reduced form expression for x_1? It is not there, although a mathematician might say that we could easily include it by writing the reduced form expression like this:

$$x_1{}^* = \frac{25}{p_1^2} + 0m$$

The fact that m does not affect optimal x_1 for a quasilinear utility function is the source of the surprising result for point B. We can apply the usual procedure for finding points A, B, and C with a reduced form expression to show this.

Point A is the initial optimal x_1 solution so we plug in $p_1 = 2$ and find $x_1{}^* = \frac{25}{2^2} = 6.25$.

Point C is the new optimal x_1 solution so we plug in $p_1 = 3$ and find $x_1{}^* = \frac{25}{3^2} = \frac{25}{9} = 2\frac{7}{9}$.

Point B is found using new p_1 and adjusted m, \$146.25. But notice that adjusted m is irrelevant because it does not affect x_1. Point B is $x_1{}^* = 2\frac{7}{9}$, the same as point C.

Figure 4.26 shows what is going on here. Unlike the typical case, there is no income effect at all with quasilinear utility, so TE = SE. As usual, the

substitution effect is the move from point A to B and the income effect is the movement from B to C. The IE is zero because C is directly below B. The total effect is A to C.

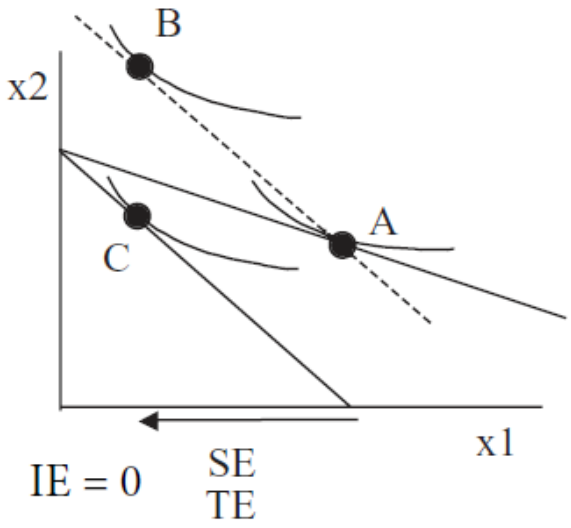


Figure 4.26: TE, SE, and IE with quasilinear utility.

It is the utility function that is driving this result. A utility function with the functional form $U = f(x_1) + x2$ has no income effect because the indifference curves are vertically parallel. If you shift the budget line via an income shock, the new tangency point will be directly above or below the initial point. In other words, the income consumption curve is vertical. Thus, the total effect is composed entirely of the substitution effect. This is the curious twist produced by the quasilinear functional form.

We saw that the income consumption curve is vertical and Engel curve is horizontal in section 4.2 (see Figure 4.7). Economics is certainly cumulative and ideas learned are often worth remembering because they tend to show up again.

Finally, notice that we now know that quasilinear preferences cannot yield Giffen behavior. After all, if the substitution effect is always negative and the income effect is zero, there is no way for the total effect to ever be positive.

Quasilinear Preferences Yield Zero Income Effects

Splitting a total effect into income and substitution effects works for any utility function. After finding the total effect, the Income Adjuster Equation can be used to determine the income needed to cancel out the change in purchasing power from the price change (i.e., setting the imaginary, dashed budget line). Finding the optimal solution with the new price and adjusted income budget constraint determines point B and allows us to split the total effect in two parts.

Of course, the component parts, SE and IE, need not be equal nor share the same sign. We know that Giffen goods arise when the income effect opposes and swamps the always negative substitution effect.

In the case of quasilinear preferences, we have a situation where there is no income effect. The Slutsky decomposition still applies, however, with the total effect being entirely composed of the substitution effect.

Exercises

1. Click the Reset button on the *OptimalChoice* sheet and apply a price decrease for good 1 from \$2/unit to \$1.90/unit. Compute the total, substitution, and income effects. Show your work.

2. Use Word's Drawing Tools to draw a graph similar to Figure 4.26 that shows the total, substitution, and income effects from the 10 cent decrease in price from question 1.

Questions 3 and 4 are difficult. Revisit questions 2 and 3 in *EngelCurvesPracticeA.doc* (in the *Answers* folder in the *MicroExcel* archive) for more detail on the corner solution for this utility function at low levels of income.

3. With quasilinear utility, the income consumption curve is vertical and the Engel curve horizontal only above a threshold income level. At very low levels of income, we get a corner solution. Click the Reset button on the *OptimalChoice* sheet and set income to 10. This will generate a corner solution. Compute the total, substitution and income effects from a 10 cent price increase in good 1 (from 2 to 2.1). Show your work.

4. Use Word's Drawing Tools to draw a graph depicting your results for question 3.

References

The epigraph comes from page 19 of the second edition of *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory* by John R. Hicks. This remarkable book was explicitly cited in the press release announcing that Hicks had won the Nobel Prize in Economic Science in 1972 (with Kenneth Arrow). "In his most well-known work, the monograph, *Value and Capital*, published in 1939, Hicks abandoned this [formal] tradition and gave the [general equilibrium] theory an increased economic relevance." See www.nobelprize.org/prizes/economic-sciences/1972/press-release/.

As mentioned in the previous section, the history of income and substitution effects is complicated. Hicks (and Allen) figured out that the total effect could be decomposed into income and substitution effects in the 1930s, two decades after Slutsky's work. Once Slutsky was rediscovered, Hicks and Allen gave him credit and made the economics profession aware of his contribution. Hicks wrote in *Value and Capital* that "The present volume is the first systematic exploration of the territory which Slutsky opened up" (p. 19).

Usually the first question anyone asks about a proposed new tax is "Who pays?" and about a tax cut is "Who benefits?"

Joel Slemrod and Jon Bakija

4.8 A Tax-Rebate Proposal

This section examines a tax-rebate plan that provides further practice with the logic of income and substitution effects. This application shows that they are more than an intellectual curiosity.

The heart of the idea is for the government to reduce consumption of a particular good, for example, gasoline, without hurting the consumer.

The idea is to tax a good and then turn around and rebate (give back) all of the tax revenue to the consumer. Can we alter the consumer's choices without lowering satisfaction? We keep things simple by ignoring administrative costs of collecting the tax and rebating it so the tax and rebate leaves the consumer's income unchanged. Proponents point out that the government is not making any money (all of the tax revenue raised is refunded back) so the consumer is not going to be hurt.

Opponents contend that this scheme will have no effect because the rebated tax will immediately be spent on the taxed good and we will end up right where we started.

Who is right? We use the Theory of Consumer Behavior to find out. Along the way, income and substitution effects will come into play.

A Concrete Example

STEP Open the Excel workbook *TaxRebate.xls* and read the *Intro* sheet, then go to the *QuantityTax* sheet.

We have a Cobb-Douglas utility function with an option to apply a per unit (quantity) tax on good 1. The workbook opens with no tax and the consumer maximizing satisfaction by buying the bundle 25,50, yielding $U^* = 1250$.

We begin by applying a quantity tax.

STEP Change cell B21 to 1. Notice that a new budget line appears. The consumer cannot afford the original bundle and must re-optimize. Run Solver to find the new optimal solution.

You should find that the consumer will now buy the bundle $16\frac{2}{3}$,50 and maximum utility falls to 833.33. Cell B22 shows that the government collects \$16.67 (\$1/unit tax on the 16.67 units purchased).

The idea behind the tax-rebate proposal called for rebating the tax revenue so that the consumer would not be hurt by the tax. We need to implement the rebate part of the proposal.

STEP Change cell B18 to 116.67. This shifts the budget constraint out. Run Solver to find the optimal solution.

You should find that the consumer optimizes by purchasing 19.445 units of x_1 and 58.335 units of x_2.

This result presents us with a problem. This is not the tax-rebate scheme the government envisioned. After all, the government is collecting more tax revenue (\$19.445) than the consumer is getting as a rebate (\$16.67).

Instead of giving the consumer \$16.67, let's give her \$19.445. What does the consumer do in this case?

STEP Change cell B18 to 119.445. This shifts the budget constraint out a little bit more. Run Solver to find the optimal solution.

Now the consumer buys a little more x_1, just over 19.9 units. But we still do not have a revenue neutral policy. We need to increase m again. This process of repeatedly doing the same thing is called *iteration*.

STEP Set the cell B18 value to \$100 (initial m) plus the amount of tax revenue in cell B22. Run Solver.

You can see that we are converging because the increases to income keep getting smaller and smaller. There is a tax rebate that yields an optimal x_1 that generates a tax revenue that exactly equals the tax rebate. The value of this tax rebate is \$20.

STEP Set cell B18 to \$120. Run Solver.

You should see that the optimal solution is 20,60 and maximum utility is 1200. If Solver is off by a little bit (this is false precision), you can enter 20 and 60 in cells B11 and B12. Since they buy 20 units of x_1, the consumer is paying \$20 in tax. Since they are getting a tax rebate of \$20 (m is set is 120), the tax they pay is exactly canceled out. We are ready to evaluate this program.

Who's Right?

Proponents argued that by taxing the good and then turning around and rebating (giving back) the tax revenues to the consumer, we can alter the consumer's choices without lowering satisfaction. Since the government is not making any money (all of the tax revenue raised is refunded back), the consumer is not going to be hurt.

Clearly the supporters of the tax-rebate proposal are wrong. The consumer had an initial $U^* = 1250$ and now has a new $U^* = 1200$. While we cannot meaningfully say that utility has fallen by 50 (because utility is measured on an ordinal, not cardinal scale), we can say that utility has fallen. Thus, in fact, the consumer is hurt by the tax-rebate proposal.

Critics, on the other hand, believed that this scheme will have no effect since the rebated tax will immediately be spent on the taxed good and we will end up right where we started.

Because the consumer went from an initial bundle of 25,50 to 20,60 after the \$20 tax-rebate, it is obvious that the critics are wrong also. This consumer has altered purchasing plans and is, in fact, buying less x_1.

So, wait, who's right—the critics or the supporters of the scheme? Neither. They are both wrong. Income and substitution effects will help us explain why.

We return to the original problem without a tax or rebate and the initial solution of 25,50. The \$1/unit tax is just like a price increase. We can find point B and compute the substitution and income effects from such a price change.

We first use the Income Adjuster Equation.

$$\Delta m = x_1{}^*\Delta p_1$$

$$\Delta m = [25][+1]$$

This result says that a \$25 increase in income to \$125 will allow us to buy the initial bundle.

STEP Set income in cell B18 to 125 (and confirm that there is a \$1/unit tax in cell B21) and run Solver.

The optimal solution is $20\frac{5}{6}, 62\frac{1}{2}$. We have points A, B, and C so we can compute total, substitution, and income effects of the \$1/unit price increase due to the tax without any rebate.

- SE (A to B): $20\frac{5}{6} - 25 = -4\frac{1}{6}$
- IE (B to C): $16\frac{2}{3} - 20\frac{5}{6} = -4\frac{1}{6}$
- TE (A to C): $16\frac{2}{3} - 25 = -8\frac{1}{3}$

Figure 4.27 displays these results with each point signifying a tangency between the budget line and an indifference curve (not drawn in to make it easier to read the graph).

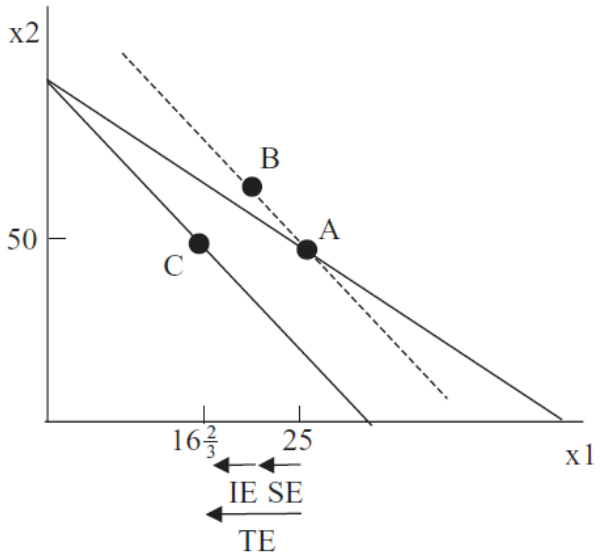


Figure 4.27: TE, SE, and IE for tax without rebate.

The tax-rebate proposal is closely related to Figure 4.27. The tax is like a price increase that moves the consumer from A to C and the rebate is like an income effect that moves the consumer from C to B.

However, if you look carefully, the changes in income are not the same. In the tax-rebate proposal, the revenue-neutral rebate is \$20, whereas in our income and substitution effect work we gave the consumer \$25 to be able to purchase the original bundle. A \$25 rebate is not revenue neutral because the consumer buys only $20\frac{5}{6}$ units of x_1 so the government ends up losing revenue. The rebate has to be \$20 to be consistent with the break-even logic of the proposal.

In addition to the income and substitution effects, Figure 4.28 adds point D, which shows the optimal solution given the tax-rebate proposal. Point D (at coordinate 20,60) has utility of 1200, which is, of course, lower than point B (the combination $20\frac{5}{6}, 62\frac{1}{2}$ yields just over 1300 units of utility). More importantly for the purposes of evaluating the proposal, utility at point D is less than utility at point A (where 25,50 generates $U^* = 1250$).

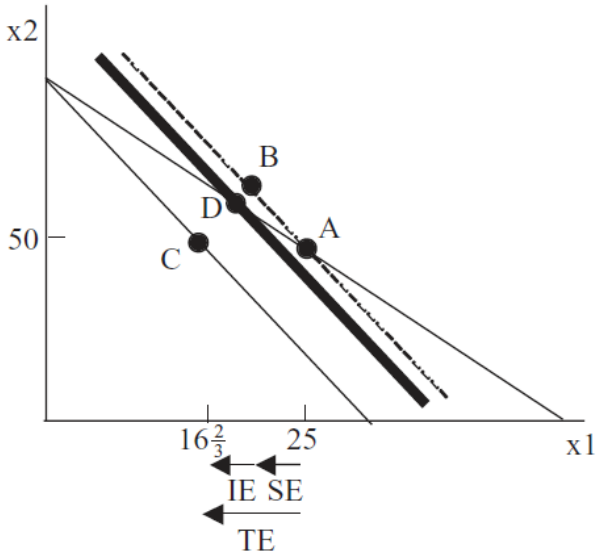


Figure 4.28: Understanding the tax-rebate proposal.

The key to the analysis lies with point D in Figure 4.28. It has to be on the initial budget line to fulfill the revenue-neutral condition of the proposal. But we know point A was the initial optimal solution on that budget line, so we can deduce that the consumer prefers point A to point D (and any other

point on the initial budget line) and will suffer a decrease in satisfaction if the tax-rebate proposal is implemented.

Tax-rebate Schemes

Taxes are often used to pay for government services and fund programs deemed worthy by society, but they can also be corrective. Taxes on specific products can discourage particular activities (think cigarettes and smoking).

Simultaneously taxing a good and rebating the tax revenue periodically appears as a policy proposal (often with regard to gasoline). Proponents claim the rebate cancels out the price increase from the tax. The scheme is related to income and substitution effects. The tax is like a price increase and the rebate is like an income effect.

Although similar to income and substitution effects, there is one important difference in tax-rebate proposals: a revenue-neutral rebate does not return enough income to allow the consumer to buy the pre-tax bundle or to reach the pre-tax level of satisfaction. Thus, the consumer cannot reach the initial level of satisfaction.

It is true, however, that a tax-rebate policy will alter consumption patterns. Whether the loss in utility is compensated by the changed consumption pattern is a different question.

Exercises

1. Analytically, we can show that the demand curves for goods 1 and 2 with a Cobb-Douglas utility function (where $c = d$) are ${x_1}^* = \frac{m}{2(p_1+Q_Tax)}$ and ${x_2}^* = \frac{m}{2p_2}$. Use these demand functions to compute the income, substitution, and total effects for x_1 for a \$1/unit tax. Show your work.

2. We know that the tax-rebate scheme gives back too little income to return the consumer to the initial level of utility (1250 units). With a \$1/unit tax, find that level of rebate where the consumer is made whole in the sense that $U^* = 1250$. Describe your procedure in answering this question.

3. At point D in Figure 4.28, is the MRS greater or smaller in absolute value than the price ratio before the tax-rebate scheme is implemented? How do you know this?

References

The epigraph is from page 87 of the fifth edition of *Taxing Ourselves: A Citizen's Guide to the Debate over Taxes* published in 2017 by Joel Slemrod and Jon Bakija. The book does not discuss the tax-rebate proposal covered in this chapter, but it is an excellent, user-friendly guide to the ever-present debate over taxes.

Government spending, taxing, and budgeting is part of the subdiscipline of economics called Public Finance. If you are interested in government's role in the economy or tax reform (including flat or consumption tax proposals), the history of the income tax in the United States, or how economists evaluate and judge taxes, this book is a great place to start.

Chapter 5

Endowment Models

Introduction to the Endowment Model

Intertemporal Consumer Choice

An Economic Analysis of Charity

An Economic Analysis of Insurance

> Our consumers could simply sit down and consume their endowments. But one consumer might, for example, be endowed with a lot of some good that she is not particularly fond of. She may wish to exchange some of that good for something she likes more.
>
> David M. Kreps

5.1 Introduction to the Endowment Model

This chapter introduces a wrinkle to the standard consumer theory model that greatly enhances its applicability. Instead of treating income as a given cash amount, we model the consumer as having a given initial endowment of goods that can be traded for other goods. This transforms the consumer into a combined consumer and seller.

Although the power of this approach may not be immediately obvious, we will see that a wide variety of examples such as saving/borrowing, charitable giving, and much more can be handled with this modification.

The Budget Constraint in an Endowment Model

Instead of the usual income (m) variable, an Endowment Model is characterized by a budget constraint that equates expenditures and revenues from sales out of the initial endowment.

$$p_1 x_1 + p_2 x_2 = p_1 \omega_1 + p_2 \omega_2$$

The term on the right-hand side says that the consumer has a given amount of each good, ω_1 and ω_2 (this is Greek letter *omega* so we have omega-one and omega-two). Because the initial amounts of each good are given, ω_1 and ω_2 are exogenous variables.

The starting amount of each good, the coordinate pair ω_1, ω_2, is called the *initial endowment*. If we multiply the initial amount of each good by the price of that good, as done in the right-hand side of the budget constraint equation, we get a dollar-valued amount that represents the total income that can be raised by selling the entire endowment.

Thus, the budget constraint says that spending (on the left-hand side) must equal the value of the consumer's assets (on the right-hand side).

The classic example to illustrate someone operating with an endowment model constraint is a farmer who goes to market with his crop. He sells his produce and, with the revenue obtained by selling, buys other goods. The core idea is that the farmer is a buyer *and* a seller.

Perhaps a more modern example is eBay. People sell all kinds of products and turn around and buy different products. It is a massive online garage-sale community. Once again, the core idea is that eBayers sell *and* buy.

In an Endowment Model, what the agent can buy depends on how much revenue is generated by sales. High prices for goods to be sold are a good thing from the agent's point of view because they generate a lot of revenue with which to buy other goods.

Because Endowment Models transform the consumer into a combined buying-selling agent, we can get different results than we saw in the Standard Model. One critical difference is that price increases lead to decreases in quantity demanded (assuming the good is normal), as usual, but as price keeps rising, we can cross the zero barrier and get *negative* quantity demanded! We will see that the agent switches from being a buyer to being a seller. This is a key idea.

Let's put these abstract ideas into concrete examples so we can understand what is going on with the Endowment Model.

STEP Open the Excel workbook *EndowmentIntro.xls*, read the *Intro* sheet, then go to the *MovingAround* sheet. Follow the instructions on the sheet to learn how we can create a budget line from a single point.

Just like the Standard Model, the agent faces a consumption possibilities frontier, also known as the budget line, that shows the feasible combinations. Bundles beyond the line are unattainable.

STEP Proceed to the *Properties* sheet.

Notice how we can use the value of the endowment to measure the agent's "income." Starting with 35,10 and $p_1 = 2, p_2 = 3$, the value of the endowment is \$100 (\$70 from x_1 and \$30 from x_2). The most x_2 the agent can have is $33\frac{1}{3}$, the y intercept and the maximum x_1, the x intercept, is 50.

The highlighted circle in the graph (reproduced as Figure 5.1) represents the initial endowment. From the initial allocation of 35,10, the agent can move northwest, selling x_1 and buying x_2. Or, the agent can decide to acquire even more x_1 by selling x_2 and buying x_1, which means traveling in a southeasterly direction. The slope of the constraint is the usual price ratio.

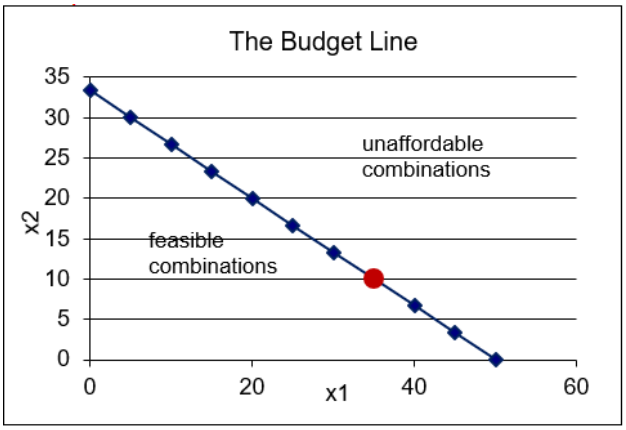


Figure 5.1: Endowment Model budget constraint.
Source: EndowmentIntro.xls!Properties

What will the consumer do in terms of buying and selling? In other words, where will the agent end up on the budget line? We do not know because we do not have any information on this agent's preferences. Before we tackle that problem, however, we need to see how the budget constraint changes when an exogenous variable is shocked.

STEP Proceed to the *Changes* sheet. Change p_1 (in K9) from 2 to 5.

This is different than before. Instead of the budget constraint pivoting about the y intercept (as in the standard, cash-income model), your screen should look like Figure 5.2. The budget constraint has pivoted or rotated as it did before, but the rotation is around the initial endowment. This is a critical difference.

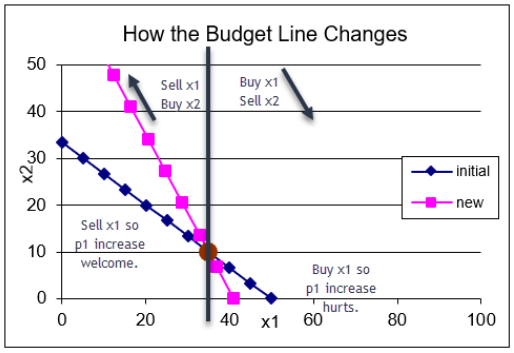


Figure 5.2: Endowment Model p_1 increase.
Source: EndowmentIntro.xls!Changes

The way the budget constraint has changed reveals important information. The price increase has improved the agent's consumption possibilities if she is planning on traveling northwest on the constraint. This makes sense because she would be a seller of good 1 and, with the higher price, she would have more money with which to buy good 2.

On the other hand, if she is a buyer, then we get the usual result that the budget line has rotated in and reduced the consumption possibilities.

STEP Click the Reset button and change p_1 (in K9) from 2 to 1.

Notice how the budget line has swiveled around the endowment again, but this time the agent is worse off if she is a seller and better off if she is a buyer.

STEP Click the Reset button and change p_2 (in K10) from 3 to 6. The result is exactly the same as when you changed p_1 (in K9) from 2 to 1.

This reveals a lesson: All that matters in the Endowment Model are relative prices, $\frac{p_1}{p_2}$. So $p_1 = 1, p_2 = 3$ is the same as $p_1 = 2, p_2 = 6$ and $p_1 = 10, p_2 = 30$ and any p_1 and p_2 whose p_1/p_2 ratio is $\frac{1}{3}$.

Finally, we consider shifts in the budget constraint. We cannot shift m (cash income) like we did in the Standard Model, but we can shock the initial endowment quantities of goods and this acts like a shift in income.

STEP Click the Reset button and change ω_1 (in K13) from 35 to 50.

The chart now looks like the usual increase in income in the Standard Model.

STEP Click the Reset button and change ω_2 (in K14) from 10 to 2.

This generates a downward shift in the budget constraint. So price changes cause rotations (or pivots or swivels) and endowment shocks produce shifts.

The budget constraint in an Endowment Model plays the same role as the budget constraint in the Standard Model. It describes the agent's consumption possibilities. Unlike the Standard Model, however, where price changes caused rotation around the x or y intercept, price shocks in the Endowment Model lead to swiveling around the initial endowment. It makes sense that the initial endowment is going to remain the same as prices change because the agent is neither buying nor selling at the initial endowment so the price does not matter at that point.

To get shifts in the budget constraint, we will have to change either ω_1 or ω_2. This changes the initial endowment point and allows the agent to buy and sell from the new endowment point, creating a new budget line.

Now that you understand the budget constraint, we are ready to solve the agent's constrained utility maximization problem with the Endowment Model.

The Initial Solution

The utility side of the Endowment Model is the same as the Standard Model. The agent's preferences are shown by indifference curves that are represented mathematically by a utility function.

The agent seeks to maximize utility given the budget constraint. As usual, we can solve this problem numerically and analytically.

STEP Proceed to the *OptimalChoice* sheet. Figure 5.3 shows what this sheet looks like when you first open it.

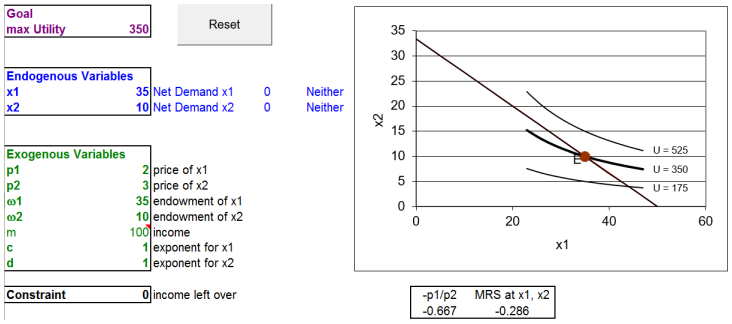


Figure 5.3: The initial view of the optimization problem.
Source: EndowmentIntro.xls!OptimalChoice

Notice how the organization is the same as the Standard Model. There is a goal, endogenous variables (in blue) and exogenous variables (in green). The agent seeks to maximize utility, represented by a Cobb-Douglas functional form, by choosing the amounts of x_1 and x_2 to consume, subject to the budget constraint.

The graph is also similar, with the addition of point E, representing the initial endowment. There are three representative indifference curves (there are an infinity of such curves, one through every point in the quadrant).

Although much is familiar, Figure 5.3 and your computer screen do have some notable innovations. Cells B18 and B19 have been added to the list of exogenous variables. They represent the given initial endowment. Cell B20 has a formula that computes m, which is not bolded to indicate that it is derived from other exogenous variables.

In addition, cells C11:E12 are new. Let's find out what they tell us.

STEP Click on D11 to see its formula, = x1_ - w1_.

The underscore (_) is used to distinguish the names, x1 and w1, from the cell addresses, X1 and W1. Lowercase w is the closest English character to ω.

More substantively, the formula computes *net demand*, how much the consumer wants to buy or sell. It takes *gross demand*, the optimal amount of the good the agent wishes to end up with, that is, the values of x_1 and x_2 and subtracts the initial endowment amounts. There is a gross and net demand for each good.

On opening, the net demand for x_1 is zero because B11 is set at 35, which is equal to the agent's initial endowment of good 1. Suppose the agent decided to buy three units of good 1.

STEP Change B11 to 38.

Net demand for good 1 is now plus three. That makes sense because the consumer started with 35 units of good 1, but wants to have 38, so three more must be purchased.

Of course, the combination 38,10 is unattainable. The consumer must sell some x_2 in order to be able to buy three units of x_1. How much needs to be sold? Two units.

STEP Change B12 to 8.

The agent is back on the budget line and net demand for good 2 is negative. Cell E12 reports that the agent is a seller of good 2. Clicking on cell E12 reveals an IF formula that displays Buyer or Seller depending on whether net demand is positive or negative.

Compare the MRS on your screen to the MRS at the initial position from Figure 5.3. Was buying three units of good 1 with the proceeds from the sale of two units of good 2 a smart move?

No. The MRS at 38,8 is farther away from the price ratio than the MRS at 35,10. The graph reveals that we moved to a lower indifference curve when we moved to 38,8.

We need to head the other way. The agent needs to travel up the budget line, to the northwest, selling good 1 and buying good 2. How much should be sold and bought?

STEP Run Solver to find the initial solution.

Utility is maximized when gross demands are 25 and $16\frac{2}{3}$ of goods 1 and 2, respectively. Net demands are -10 and $6\frac{2}{3}$. This means the agent sells 10 units of good 1 and uses the \$20 in revenue to buy $6\frac{2}{3}$ units of good 2.

This is the same solution as in the Standard Model with $m = \$100$. That makes sense, since the value of the initial endowment is \$100.

We can confirm this result with analytical methods. We follow the recipe for the Lagrangean method of solving constrained optimization problems.

We will work on a general form of this problem, leaving all exogenous variables as letters to get a reduced form expression that we can evaluate for any combination of exogenous values. We rewrite the constraint to that it is equal to zero and form the Lagrangean.

$$\max_{x_1,x_2,\lambda} L = x_1^c x_2^d + \lambda(p_1\omega_1 + p_2\omega_2 - p_1x_1 - p_2x_2)$$

The third step is to take derivatives with respect to each choice variable and in the final step we set the three derivatives equal to zero to get the first-order conditions, which we need to solve for $x_1{}^*, x_2{}^*$, and λ^*.

$$\begin{aligned}
\frac{\partial L}{\partial x_1} &= cx_1^{c-1}x_2^d - p_1\lambda = 0 \\
\frac{\partial L}{\partial x_2} &= dx_1^c x_2^{d-1} - p_2\lambda = 0 \\
\frac{\partial L}{\partial \lambda} &= p_1\omega_1 + p_2\omega_2 - p_1x_1 - p_2x_2 = 0
\end{aligned}$$

Our solution strategy involves moving the lambda terms to the right-hand side and dividing the first equation by the second to cancel lambda (and giving the familiar MRS $= \frac{p_1}{p_2}$ condition). This equation can then be solved for optimal x_2 as a function of optimal x_1.

$$\begin{aligned}
\frac{cx_2^*}{dx_1^*} &= \frac{p_1}{p_2} \\
x_2^* &= \frac{d}{c}\frac{p_1}{p_2}x_1^*
\end{aligned}$$

Although it looks like it, this is not the answer for x_2 because it has x_1 in it. The reduced form solution must be a function of exogenous variables alone. Substitute this expression for x_2 into the third first-order condition (the budget constraint) and solve for optimal x_1.

$$p_1\omega_1 + p_2\omega_2 - p_1x_1^* - p_2\left[\frac{d}{c}\frac{p_1}{p_2}x_1^*\right] = 0$$
$$\left(1+\frac{d}{c}\right)p_1x_1^* = p_1\omega_1 + p_2\omega_2$$
$$x_1^* = \left(\frac{c}{c+d}\right)\frac{p_1\omega_1 + p_2\omega_2}{p_1}$$

This expression can be evaluated for any combination of exogenous variable values. For example, if we use the parameter values in the *OptimalChoice* sheet, we can compute that optimal $x_1 = 25$. This agrees perfectly with the numerical approach.

Furthermore, this expression shows the quantity demanded at a given p_1, ceteris paribus, so it can be used to display a demand curve for x_1. There is, of course, a similar expression for good 2.

In the Standard Model, the reduced form solution was ${x_1}^* = (\frac{c}{c+d})\frac{m}{p_1}$. The Endowment Model's solution is the same, except instead of m in the numerator, we have $p_1\omega_1 + p_2\omega_2$. This makes sense since the value of the initial endowment is $p_1\omega_1 + p_2\omega_2$.

With an Endowment Model, we can subtract the initial amount of good 1 to obtain a net demand curve.

$$nd_1 = {x_1}^* - \omega_1 = \left(\frac{c}{c+d}\right)\frac{p_1\omega_1 + p_2\omega_2}{p_1} - \omega_1$$

Comparative Statics with the Endowment Model

We can do comparative statics analyses analytically or numerically. The reduced form expression can be used to explore the rate of change of optimal x_1 with respect to any exogenous variable. For example, we can take the derivative with respect to p_1.

This is more complicated than usual because p_1 appears in two places. We could use the Product Rule, but it is easier to do some reorganizing and simplify things before we take the derivative.

First, we move p_1 from the denominator. This will enable us to use our usual derivative rule.

$$x_1{}^* = (\frac{c}{c+d})\frac{p_1\omega_1 + p_2\omega_2}{p_1} = (\frac{c}{c+d})(p_1\omega_1 + p_2\omega_2)p_1^{-1}$$

But we can also multiply p_1 through to cancel the p_1 in the $p_1\omega_1$ term.

$$x_1{}^* = (\frac{c}{c+d})(p_1\omega_1 + p_2\omega_2)p_1^{-1} = (\frac{c}{c+d})(\omega_1 + p_1^{-1}p_2\omega_2)$$

Then we can expand to leave p_1 isolated in a single term so that the derivative with respect to p_1 is straightforward.

$$x_1{}^* = (\frac{c}{c+d})(\omega_1 + p_1^{-1}p_2\omega_2) = (\frac{c}{c+d})\omega_1 + (\frac{c}{c+d})p_1^{-1}p_2\omega_2$$

Now, when we take the derivative with respect to p_1, we apply our usual derivative rule and bring the exponent down and subtract one from the second term. The first term has a derivative with respect to p_1 of zero since it does not contain p_1.

$$\frac{dx_1{}^*}{dp_1} = (-1)(\frac{c}{c+d})p_1^{-2}p_2\omega_2$$

We can evaluate this expression at the initial values of the exogenous variables to get an instantaneous rate of change in optimal x_1 as p_1 changes. Plugging in $c = d = 1, p_1 = 2, p_2 = 3$, and $\omega_2 = 10$ gives -3.75. This means that an infinitesimally small increase in p_1 would decrease x_1 by 3.75-fold.

But what does that number tell us? Is it a lot in the sense of a big response to a price shock? The slope provides no answer to this question. We need percentage changes—elasticity—to answer this question.

We can multiply the slope by the initial ratio of $\frac{p_1}{x_1{}^*}$ to compute the p_1 elasticity of $x_1{}^*$.

$$\frac{dx_1{}^*}{dp_1}\frac{p_1}{x_1{}^*} = ((-1)(\frac{c}{c+d})p_1^{-2}p_2\omega_2)(\frac{p_1}{x_1{}^*})$$

We evaluate this expression at $p_1 = 2$ (and the initial values of the other exogenous variables).

$$\frac{dx_1{}^*}{dp_1}\frac{p_1}{x_1{}^*} = ((-1)(\frac{c}{c+d})p_1^{-2}p_2\omega_2)(\frac{p_1}{x_1{}^*}) = -3.75(\frac{2}{25}) \approx -0.3$$

The elasticity does tell us that the quantity demanded of x_1 is quite price insensitive at the initial solution. An elasticity less than one (in absolute

value) is said to be inelastic and the closer to zero, the lower the responsiveness.

Unlike the Standard Model, where a Cobb-Douglas utility function gives a unit price elasticity, we get a non-unitary elasticity here because a change in p_1 appears in the denominator and numerator in the reduced form. In the numerator, the change in price is affecting the value of the agent's endowment whereas in the Standard Model, income is fixed.

We can also use numerical methods to explore the comparative statics properties of an own price change.

STEP Use the Comparative Statics Wizard to *decrease* p_1 by 0.1 (10 cents) for 15 shocks (from 2 to 0.5). Be sure to keep track of net demands and the buyer/seller position in the endogenous variables by using the *ctrl* key to select non-contiguous cells, as depicted in Figure 5.4. You want to track cells B11:B12 and D11:E12.

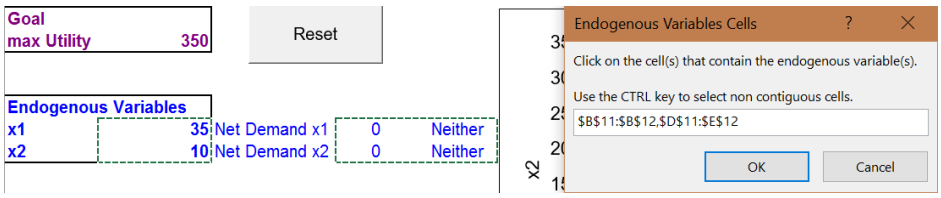


Figure 5.4: Selecting endogenous variables with CSWiz.

The *CSP1* sheet shows what your results should look like. There are several notable outcomes.

When the price fell from 90 cents to 80 cents, the agent switched from selling x_1 and buying x_2 to buying x_1 and selling x_2. The price of x_1 got so low that even though the agent starts with a lot of x_1 (compared to x_2), it is better to buy more x_1. The budget line gets flatter as p_1 falls, making buying x_1 a better choice than selling it.

Notice the behavior of maximum utility (column B) as price falls. The agent was a seller at first so falling prices hurt. Below 90 cents, however, the agent is a buyer of x_1 and falling p_1 increases utility.

The *CSP1* sheet also shows slope and elasticity computations. From p_1 \$2/unit to \$1.90, the slope (yellow background) and elasticity (orange back-

ground) measures are close, but different than at $p_1 = 2$ (using the derivative). This is due to the fact that optimal x_1 is non-linear in p_1. In other words, $x_1{}^* = f(p_1)$ is not a line, but a curve (as clearly shown in the chart below the data).

The Endowment Model Extends the Standard Model

The Endowment Model is the Standard Model of the Theory of Consumer Behavior with an initial endowment of goods instead of cash income. This transforms the consumer into the dual-role of seller and buyer of goods. The driving force in the agent's decision making remains utility maximization. Many of the ideas behind the Standard Model (such as equating the MRS and price ratio) carry over to the Endowment Model. Of course, the framework for presenting and understanding the model, comparative statics analysis, remains the same.

It may seem that replacing income with an initial endowment is a minor twist, but we will see that the Endowment Model enables analysis of a wide range of choice problems.

Exercises

1. Perform a comparative statics analysis of c, the exponent on x_1, using the Comparative Statics Wizard. Use increments in c of 0.1. State the effect of changing c on $x_1{}^*$. Describe your procedure and take screen shots of your results as needed.

2. Use your comparative statics results to find the c elasticity of $x_1{}^*$ from 1 to 1.1. Show your work.

3. Use the reduced form expression in this chapter to find the c elasticity of $x_1{}^*$. Show your work.

4. Compare your answers from questions 2 and 3. Explain why they are the same or differ.

References

The epigraph is from page 188 of David M. Kreps *A Course in Microeconomic Theory* (1990). If you are interested in graduate study in economics,

this book is worth browsing. In the preface, Kreps says (p. xv), "The primary target for this book is a first-year graduate student who is looking for an introduction to microeconomic theory that goes beyond the traditional models of the consumer, the firm, and the market." Kreps allows that it could be used for undergraduate majors taking an "advanced theory" course or "mathematically sophisticated students," but he warns that, "The book presumes, however, that the reader has survived the standard intermediate microeconomics course."

The Endowment Model is taking us close to the next level of microeconomic theory. Google "graduate micro theory" for more advanced micro books.

To learn more about Masters and PhD programs in economics, search for "graduate economics rankings" and be sure to visit the American Economics Association's website at www.aeaweb.org/resources/students/grad-prep.

The term impatience carries with it the presumption that present goods are preferred. But I shall treat the two terms (impatience and time preference) as synonymous.

Irving Fisher

5.2 Intertemporal Consumer Choice

Suppose the government wants to stimulate saving by workers so they won't be poor when they retire. Individual Retirement Accounts (IRAs) and 401(k) (their section in the tax code) plans enable savings to grow tax free, so the interest rate earned is higher than if returns were taxed. A higher interest rate should stimulate more saving. But how much more?

Typically, estimates of the interest rate elasticity of savings are positive, but quite small, say 0.15. If someone had this elasticity, would attempts to stimulate saving by increasing the interest rate be effective?

No, because the low interest rate elasticity of savings means that saving is not responsive to changes in the interest rate. Suppose the interest rate doubles so we have a huge 100% change. Because the elasticity is 0.15, that means we will see only a 15% increase in savings. A more realistic 10% increase in the interest rate would generate a small 1.5% increase in savings. The small elasticity tells us that shocks to the interest rate are not going to move the amount saved by very much.

This is an example of interpreting an elasticity. Computing an elasticity is important (and you will continue to see examples of how to do it), but understanding what an elasticity is telling us is even more critical.

Now that we know the elasticity is low and what that means, this leads to a second question: What would make the interest rate elasticity of savings be so small? The rest of this chapter offers an application of the Endowment Model to answer this question. In addition, income and substitution effects play a major role in the explanation. There is no doubt about it, learning economics is a cumulative undertaking—the same ideas keep popping up again and again.

The Intertemporal Choice Model

Intertemporal choice means the agent faces a decision that spans across time periods. Saving over the years working means less consumption, but that allows for more consumption when retired. We model the agent as deciding what to consume every year over their lifespan.

Just as when we modeled the consumer buying just x_1 and x_2 instead of many goods and services, we make a simplifying assumption that collapses many time periods into two: present and future. In the present, right now, the agent works and in the future, one year later, she does not (she retires).

In addition, there is another implied simplifying assumption: the agent knows with certainty how long she will live. She is born and works as one-year old, is retired as a two-year old and dies on the last day of her second year. She decides, as soon as she is born, how much she will consume in year 1 (the present) and year 2 (the future).

Instead of having two goods x_1 and x_2, we have consumption of a single good in the present, c_1, and the future, c_2. The price of the single good is \$1/unit so if you have, say, \$40, you can buy 40 units. There is no inflation so the price is the same in both time periods.

Notice the usual modeling technique at work here—realistic details are simply assumed away. Most people's lives unfold as follows: Childhood becomes teen-aged years, and then a long period of working adult life eventually turns to retirement years and death. The Intertemporal Choice Model collapses all of that into two time periods. It also assumes away complications from not knowing exactly when we die.

Faced with criticisms about the unrealistic nature of the model, economists respond by saying that we are not interested in realism. We reduce the complex real world to a model that can be analyzed with comparative statics to produce testable predictions. For economists, the goal is not to describe reality, but to predict via comparative statics. We strip away all complications to create an unreal, incredibly simple model that contains the kernel of the problem so we can work out how the agent responds to shocks.

Modeling is not easy. There is science (and math) and art involved. Users and consumers of these models need sharp critical thinking skills—sometimes important elements are assumed away.

We continue building the model by defining the initial endowment as the amount of present and future income you start with. The initial endowment in the first year is m_1 and in the second year m_2. The first year's initial endowment is income from working and the second year's initial endowment is income from sources like Social Security. Thus, it makes sense that $m_1 > m_2$, which says that income is higher during the working than the retired year. Since the price is \$1/unit, the initial endowment incomes are also initial endowment consumption in the two periods.

We are ready to work on the optimization problem itself. We follow the usual approach, modeling the budget constraint, then satisfaction, then putting the two together to find the initial solution. Of course, after finding the initial optimum we will do comparative statics analysis, where we will answer the question: What causes the interest rate elasticity of savings to be so small?

The Budget Constraint

STEP Open the Excel workbook *IntertemporalChoice.xls* and read the *Intro* sheet, then go to the *MovingAround* sheet.

The consumer begins at the initial endowment point, 80,20, where 80 represents her income and consumption in time period 1 (remember that the price of the good is \$1/unit). Income and consumption of 20 in time period 2 is lower (given that she is not working). These numbers are arbitrary and do not have any special meaning.

A critical concept for the Endowment Model is that the agent does not have to stay at the initial position. In this application, she can move by saving or borrowing. Saving means you consume less in the present and carry over the unconsumed portion into the future. Saving is like selling present consumption and buying future consumption.

Suppose she saves 30 units of consumption in year 1 by saving \$30. What would be her position in the second year?

STEP Change cell B19 to 50. This implements the plan to increase future consumption, but look at cells B21 and B22. Instead of simply reallocating from 80,20 to 50,50, by saving 30 units, she got an extra 6 units in interest on her savings.

If you save \$30 for one year at 20%, you end up with \$56. The \$30 you saved (called the *principal*) and *interest* earned of \$30 x 20% = \$6 makes your savings worth \$36 in the future and we add this to the \$20 of initial future income to get the grand total of \$56.

There is an equation that gives us the value of c_2 for any chosen value of c_1.

$$c_2 = m_2 + (m_1 - c_1) + r(m_1 - c_1)$$

The equation says that the amount of consumption in time period 2 equals the initial endowment amount in time period 2, m_2, plus the principal saved, $m_1 - c_1$, plus the interest earned on the amount saved, $r(m_1 - c_1)$. We can rewrite this in a simpler form by collecting the savings term.

$$c_2 = m_2 + (1 + r)(m_1 - c_1)$$

This is the equation of the budget constraint in this model. It shows that the intercept is $m_2 + (1 + r)m_1$ and the slope is $-(1 + r)$ (just multiply through by $(1 + r)$). The slope tells us that saving \$1 will yield $1 + r$ dollars in time period 2.

What would be the maximum consumption possible in time period 2? We have two ways to answer this question.

STEP Change cell B19 to 0. She consumes nothing now and ends up with 116 units in the future.

"But she will starve if she consumes nothing in period 1." That would be another constraint that is not being modeled. We are not saying she will consume nothing in the present time period, we are merely exploring the consumption possibilities.

Saving everything (the same as consuming nothing in the present) can also be found by computing the value of the y intercept. We can evaluate $m_2+(1+r)m_1$ at $m_1 = 80, m_2 = 20$, and $r = 20\%$, yielding $20+(1+0.2)80 = 116$. This is the same answer that we got with Excel.

The y intercept tells us the *future value* of the agent's initial endowment, measuring income in both periods in terms of time period 2.

Instead of saving, the agent can borrow. Suppose the agent decided to consume more than 80 units in time period 1. How could she do this? Easy: use her time period 2 income to borrow from it. As before, however, we have to be careful. The interest rate plays a role.

STEP Change cell B19 to 90. She borrows \$10 from her future income.

Does she end up with 90,10—subtracting 10 from c_2 and adding it to c_1? No way. As Excel shows, she has to pay interest on the borrowed funds. If she borrows \$10, she ends up with only \$8 in the future because she has to pay back the principal (\$10) and the interest (\$2).

What is the most she could consume in time period 1?

STEP Change cell B19 to 100. What happens?

She cannot do this. She cannot choose negative x_2. She does not have enough future income to enable 100 units of time period 1 consumption.

STEP Continue entering numbers in cell B19 until you drive c_2 (in cells B23 and B24) to zero.

The x intercept is $96\frac{2}{3}$. It is the *present value* of her endowment, measuring income in both periods from the standpoint of time period 1.

STEP Proceed to the *Properties* sheet.

Our work in the *MovingAround* sheet makes it easy to understand the budget line displayed in the *Properties* sheet. Clearly, given an initial endowment, movement up the budget line is saving and down is borrowing.

These are just consumption possibilities. We do not know what this person will do until we incorporate her preferences. We do know she can be anywhere on the constraint (including the initial endowment point). It all depends on her indifference map and where the highest attainable indifference curves lie.

STEP Proceed to the *Changes* sheet. Change the interest rate, cell L8, to 50%. Your screen will look like Figure 5.5.

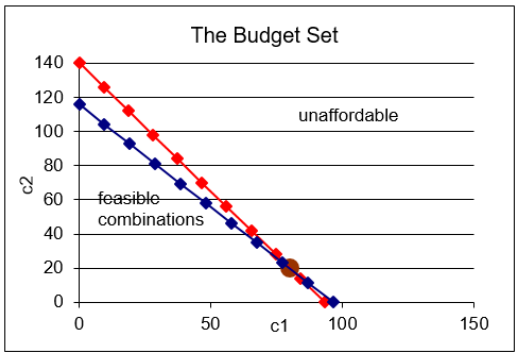


Figure 5.5: Increasing r.
Source: IntertemporalChoice.xls!Changes

Our work with the Endowment Model in the previous section enables us to easily interpret the result. As before, the budget constraint swivels around the initial endowment point.

Above the initial endowment point, the increase in r is a good thing, increasing consumption possibilities. If the agent is a saver, the shock is welcome.

Borrowers, however, would not be happy with an increase in r. This is a price increase to present consumption and reduces consumption possibilities for borrowers.

STEP Click the Reset button. Change m_1 and m_2 to see how these shocks are like an income shock. It maintains the slope, but shifts the budget constraint.

Now that we understand how the budget constraint works, we are ready to turn to the agent's goal, maximizing utility.

Preferences

The agent has preferences over present and future consumption that can be captured by the indifference map.

We use the usual Cobb-Douglas function form to express preferences as a utility function.

STEP Proceed to the *Preferences* sheet. Compare the utility functions with $d = 0.5$ and $d = 0.1$. The utility function allows us to model different preferences.

Figure 5.6 shows two different agents with different rates of time preference for future consumption. The person on the right exhibits a strong preference for present consumption, while the person on the left is more willing to wait.

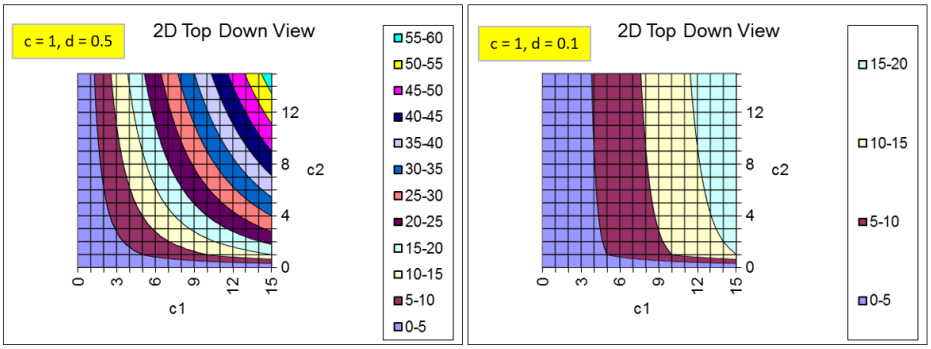


Figure 5.6: Modeling rates of time preference.
Source: IntertemporalChoice.xls!Preferences

A more immediate gratification personality is represented on the right side of Figure 5.6. We would say this person is more impatient—he likes present much more than future consumption. The exponent d is much smaller than c, which means inputs into the utility function through c_2 provide much less utility than via c_1.

The steep indifference curves reveal that he is willing to trade a great deal of future consumption for a just a little more present consumption. His MRS at a given point (for example, 6,6) is higher (in absolute value) than the MRS of the person on the left.

We do not say the person on the right has "bad preferences" (although the language used in this example, such as *impatience* does seem to connote disapproval). Economists take preferences as given. We are not supposed to judge them as right or wrong. A person with preferences that substantially ignore the future is treated the same as someone who does not like broccoli or likes the color blue.

There is a complication here, however, in that a person's rate of time preference almost certainly changes over time. A young person may not save much

because she does not value the future, but she may regret her decision when she gets older. Deciding whose preferences should rule, young or old you, is a difficult philosophical problem.

With the budget line and preferences, we can now solve the constrained utility maximization problem.

Finding the Initial Solution

STEP Proceed to the *OptimalChoice* sheet. Figure 5.7 shows the initial display. The current bundle is 80,20—the initial endowment point. The agent is not maximizing satisfaction subject to the budget constraint. The indifference curve is clearly cutting the budget line and, therefore, the agent should move northwest up the budget line to maximize utility.

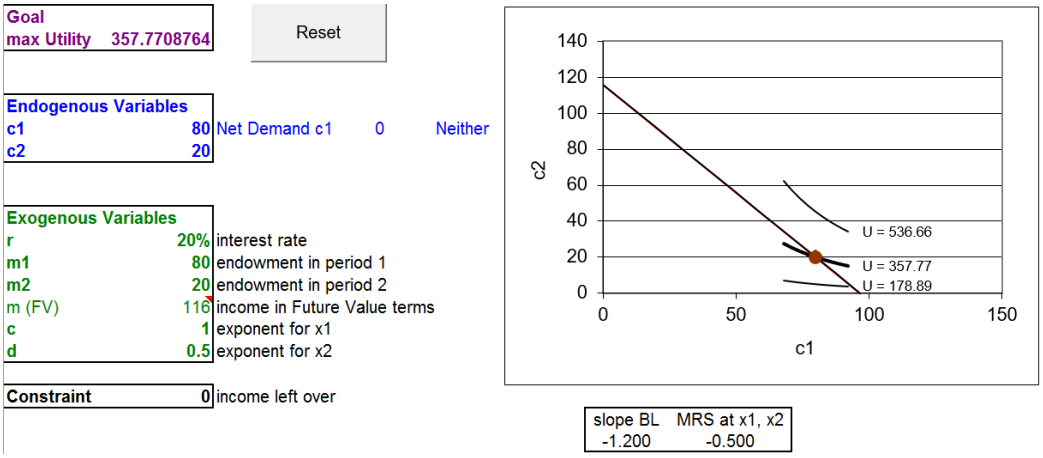


Figure 5.7: An inefficient position.
Source: IntertemporalChoice.xls!OptimalChoice

STEP Run Solver to find the initial solution.

The agent opts for the point $64\frac{4}{9}, 38\frac{2}{3}$. This means she has decided to save $15\frac{5}{9}$ of her present consumption. She chooses this present and future combination, implying this level of saving, because this maximizes utility subject to the budget constraint.

Notice that the negative net demand is interpreted as saving. It is computed as optimal c_1 minus the initial endowment of present consumption. As mentioned earlier, saving is like selling present consumption to buy greater future consumption. We often drop the minus sign so we do not get confused by increases and decreases in saving.

Comparative Statics

We focus on r. We want to know how savings will respond when r changes. Remember our question: Why is the interest rate elasticity of savings so low?

Before we begin our comparative statics analysis, we need to be clear about the language used. Since the shock variable, r, is measured as a percent, things can get confusing once we start working on responses and elasticities. We need to keep clear the difference between a *percentage point change* and *percent change*. They sound the same, but the former is a difference (Δ), $new - inital$, and the latter is a percent computation, $\frac{new-initial}{initial}$.

So, if r increases from 20% to 30%, that is a 10 percentage point change since we compute 30 - 20, but a 50 percent change: $\frac{30-20}{20}$. The same language would be used if we were working with unemployment rates. An increase from 5% to 6% is a one percentage point increase and a 20% increase.

The finance literature uses basis points for differences in variables measured in percents. There are 100 basis points in one percentage point. If a bond yield rises from 3.25% to 3.35%, that is an increase of 10 basis points.

STEP Run the Comparative Statics Wizard, changing the interest rate by 10 percentage points (0.1) increments. Keep track of c_1, c_2, net demand, and whether the person is a saver or borrower (cells D11 and E11).

Your results should be similar to those in the *CSr* sheet.

STEP Use your CSWiz results to compute the interest rate elasticity of savings from r = 20% to 30%.

We find that the interest rate elasticity of savings from r = 20% to 30% is about 0.11. (Check the formula in cell I15 in the *CSr* sheet if needed.) That is quite low. A 50 percent increase in r only increased savings by a little over 5 percent.

This elasticity is similar to the 0.15 elasticity at the beginning of this chapter. Why is this happening? Why is saving so unresponsive to changes in the interest rate?

The answer lies in the income and substitution effects. For savings, the income and substitution effects from a change in r work in opposite directions (when c_1 is a normal good). Thus, they tend to cancel each other out and the total effect ends up being small.

To head off serious misunderstanding, you need to know right now that this does not mean that we are dealing with a Giffen good. We will see that we are dealing with cross effects when r rises for a saver and Giffen goods are defined in terms of own effects. Also, c_1 and c_2 are both normal goods in a Cobb-Douglas utility function so we know we can't get Giffenness.

STEP To see how the income and substitution effects apply to this problem, return to the *OptimalChoice* sheet. Suppose r increases to 300%. Change B16 to this absurdly high interest rate.

This huge change enables us to see clearly what is happening on the graph. The budget line swivels in a clockwise direction, getting much steeper. Remember that the slope is $-(1+r)$ so an increase in r makes the line steeper. This is good for savers and bad for borrowers.

STEP After changing cell B16 to 300%, run Solver to find the new initial solution.

Solver gives the new optimal solution, $c_1{}^* = 56\frac{2}{3}$ and $c_2{}^* = 113\frac{1}{3}$, when $r = 300\%$. Optimal savings has increased from \$15.56 to \$23.33, so that is good news, but this is a pretty weak response to the massive increase in the interest rate from 20% to 300%.

Figure 5.8 shows the initial solution (point A) and the new optimal solution (point C). It also includes a dashed line that is parallel to point C's budget line, but goes through point A. This, of course, is the line that is used to separate the total effect into income and substitution effects using point B.

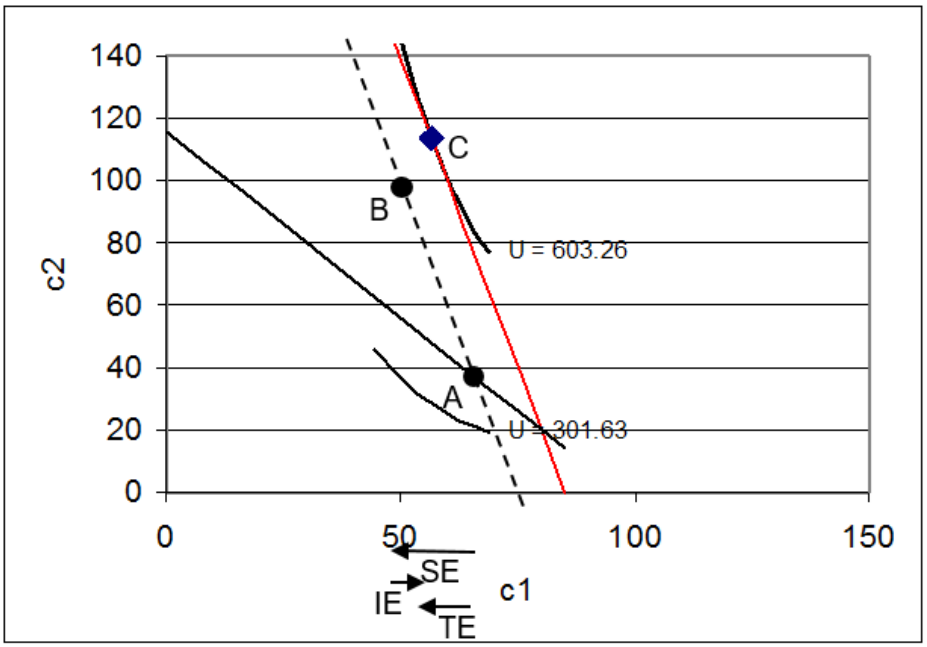


Figure 5.8: Income and substitution effects.
Source: IntertemporalChoice.xls!OptimalChoice: cell F52

How much income (m_1) did we have to take away (hypothetically, of course) to cancel out the income effect of the higher interest rate? We can use Excel to answer this question.

STEP With $r = 300\%$, enter the initial solution (point A). To minimize rounding error, use a formula with fractions. So, enter "= 64 + 4/9" in B11 and "= 38 + 2/3" in B12. Now, start decreasing m_1 (in cell B17). Your goal is to find that value of m_1 so that the initial solution is on the budget line—i.e., the constraint cell is zero.

A little experimentation should convince you that $m_1 = 69\frac{1}{9}$ is the value that puts the dashed budget line through the initial solution.

If you want to be daring, you could use Solver. Call Solver, then click the Reset All button. The objective is the constraint cell (B23) and you want to make the value of it zero by changing m_1 (B17). Solver gives the same answer as above.

Or, you could use the budget constraint to find the m_1 needed to buy the original optimal bundle with $r = 300\%$. Simply plug in the initial optimal

solution along with the new value of r (and initial m_2) and solve for m_1. You are finding the value of m_1 that would enable you to buy the initial optimal combination with the higher interest rate. The analytical answer agrees with the numerical approach.

STEP Now, with r = 300% and $m_1 = 69\frac{1}{9}$, run Solver to find point B.

Be careful with the interpretation of savings for point B. Remember that income is not really $m_1 = 69\frac{1}{9}$, but 80. This means that at point B, the agent would save \$30.59, not \$19.07 as displayed in cell D11.

Figure 5.9 shows the results in a table. You can see Figures 5.8 and 5.9 side by side by scrolling down to row 50 or so in the *OptimalChoice* sheet. Look at how the substitution effect leads to a large increase in savings, but the income effect cancels out part of this increase.

Point	Description	c1*	c2*	Savings*
A	Initial solution	64.44	38.67	15.56
B	Imaginary	49.41	98.81	30.59
C	New solution	56.67	113.33	23.33

Effect	Movement	Amount
SE	A to B	15.03
IE	B to C	-7.26
TE	A to C	7.77

Figure 5.9: Total, income, and substitution effects.
Source: IntertemporalChoice.xls!OptimalChoice: cell M51

The income and substitution effects provide an explanation for the low interest rate elasticity of savings. What is happening is that the two effects are working against each other when r rises and the agent is a saver.

Does this mean c_1 is an inferior good? No. The reason why the effects are opposing each other is because, for savers, an increase in the interest rate is like a decrease in the price of future consumption so the effects on c_1 and savings are actually cross effects. Look carefully at Figure 5.8. In the region of the graph with points A, B, and C, it is as if we decreased p_2,and rotated the budget line up clockwise (with a steeper slope).

Saving and Borrowing Explained

The Intertemporal Choice Model is an application of the Endowment Model in the Theory of Consumer Behavior. The model says that the agent chooses the amount to consume in time periods 1 and 2 in order to maximize satisfaction given a budget constraint.

The model explains saving (or borrowing) as an optimizing move on the part of an agent who is trading off present and future consumption.

The model can also explain why the interest rate elasticity of savings is often estimated as a positive, but small number, which means that saving is quite unresponsive to the interest rate. The explanation rests on the fact that the income effect opposes the substitution effect for c_1 and savings (for those with negative net demand for c_1).

Exercises

1. Solve the problem in the *OptimalChoice* sheet using analytical methods. In other words, find the reduced form expressions for optimal c_1, c_2, and saving from

$$\max_{c_1,c_2} u(c_1, c_2) = c_1^c c_2^d$$
$$\text{s.t. } c_2 = m_2 + (1 + \text{r})(m_1 - c_1)$$

   Show your work.

2. Use the parameter values in the *OptimalChoice* sheet (with $r = 20\%$) to evaluate your answers for question 1. Provide numerical answers for the optimal combination of consumption in time periods 1 and 2 and for optimal saving.

3. Do your answers from question 2 agree with Excel's Solver results? Is this surprising? Explain.

4. Use your reduced form solution from question 1 to compute the interest rate elasticity of savings at r = 20%.

5. In working through this chapter, you found the interest rate elasticity of savings from r = 20% to 30%. Why is the elasticity computed at a point (in question 4 above) different from this elasticity?

References

The epigraph is on page 66 of Irving Fisher, *The Theory of Interest: As Determined by Impatience to Spend Income and Opportunity to Invest It* (first edition, 1930; reprinted 1977 by Porcupine Press).

Joseph Schumpeter had high praise for Fisher: "[S]ome future historian may well consider Fisher as the greatest of America's scientific economists up to our own day" (*History of Economic Analysis*, 1954, p. 872). Schumpeter chose to ignore Fisher's "propagandist activities (temperance, eugenics, hygiene, and others)," but he did point out that Fisher's reputation as an economist was negatively affected: "Fisher, a reformer of the highest and purest type, never counted costs—even those most intensive pain costs which consist in being looked upon as something of a crank—and his fame as a scientist suffered correspondingly" (*History of Economic Analysis*, 1954, p. 873).

For a recent biography of Fisher, who seems to be enjoying a rehabilitation of sorts, see Robert W. Dimand (2019), *Irving Fisher*.

The empirical evidence on the interest rate elasticity of savings is mixed (which is actually evidence that it is not large). For a dated, but perhaps comprehensible example, see Irwin Friend and Joel Hasbrouck, "Saving and After-Tax Rates of Return," *The Review of Economics and Statistics*, Vol. 65, No. 4. (November, 1983), pp. 537–543, www.jstor.org/stable/1935921.

The literature on the effect of Individual Retirement Accounts and other plans (such as 401(k)) on saving is truly vast. A Google Scholar search on "individual retirement accounts saving" produces hundreds of thousands of hits. This topic would make an excellent paper or undergraduate senior thesis.

> The Prophet said: "Charity is a necessity for every Muslim." He was asked: "What if a person has nothing?" The Prophet replied: "He should work with his own hands for his benefit and then give something out of such earnings in charity."
>
> Prophet Muhammed

5.3 An Economic Analysis of Charity

The phrase "an economic analysis of" is code for "using the framework of optimization and comparative statics to study observed behavior." In this case, we use the Endowment Model from the Theory of Consumer Behavior to study charitable giving.

How can economics have anything to say about giving away money? Isn't charity something really nice people do, not the selfish, rational maximizers that inhabit economics? Doesn't this mean that thinking like an economist is useless for studying charity?

These questions are based on a common misunderstanding that economics applies only to a subset of the world. So, the mistaken thinking goes, you can use economics to study certain things like banking or unemployment, but not war or marriage. This is wrong because modern economics is not defined by content, but by method. Anything involving choice, like going to war or getting married or brushing your teeth or joining a church can be analyzed with the tools of economics.

We will see that the economic approach offers a different view of charitable giving. By casting the problem as a choice—how much to give is the key endogenous variable—we can apply the optimizing and comparative statics framework of economics. We do not claim this is the only or even the best perspective, but it does provide another way to understand charity.

Basic Facts about Giving

Each year, people all around the world give away a lot of money, goods, and time (as volunteers). Humans are sympathetic when people close to them are in distress. All religions encourage charity and caring for people less fortunate.

Giving USA provides data on philanthropy in the United States. Figure 5.10, from the *2018 Annual Report*, shows the breakdown of the $410 billion that were contributed to charities in 2017. To help understand what this number means, we can compare total contributions to the size of the economy and we find a giving rate of about 2.1% of GDP.

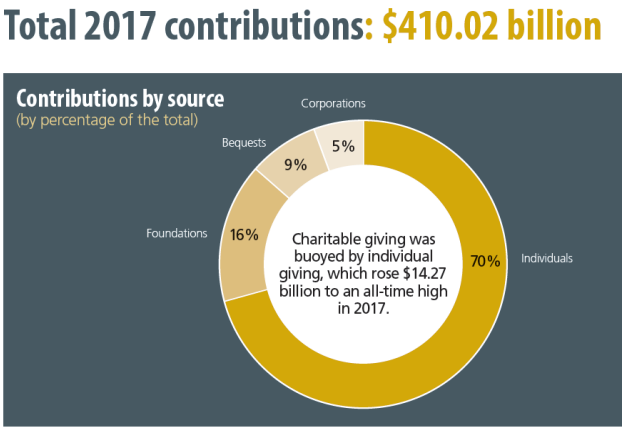


Figure 5.10: Charitable giving by source of contribution.
Source: Giving USA 2018 Annual Report

The *2018 Annual Report* contextualizes total giving by tracking giving over time, shown in Figure 5.11. Total giving jumped in the mid 1990s and reached its highest level in 2017. That is good news.

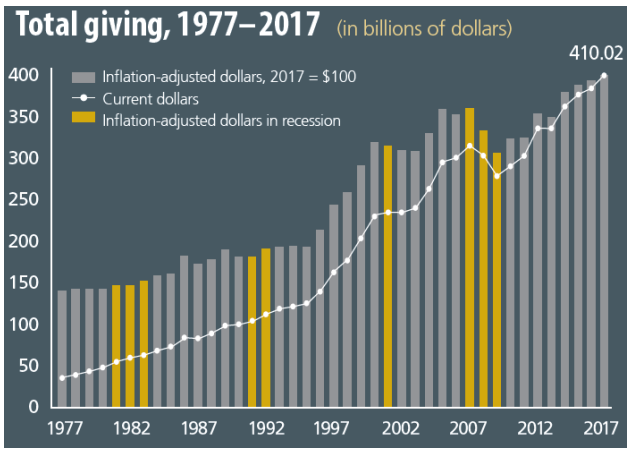


Figure 5.11: Charitable giving over time.
Source: Giving USA 2018 Annual Report

The Internal Revenue Service is another source of data on charitable giving because taxpayers claim deductions when they give to charity to lower the tax owed. The IRS also collects data on non-profit organizations which do not pay tax, but they have to file Form 990. IRS data can be found at www.irs.gov/statistics.

Charitable giving data shows that it not only varies over time, there is also tremendous individual variation. Many people give nothing, others give a little, and a few people donate a lot. Religions encourage members to tithe, giving 10% of their income. Upon death, some people give substantial fractions of their estates to charity, while others hand it all to their heirs.

There are many questions we can ask about charitable giving, but our top three are:

1. Why do people give to charity?
2. What determines how much they give?
3. How can charitable giving be stimulated?

Because this is an economic analysis of charity, we are going to answer these questions by using the method of economics. We will set up and solve an optimization problem. This will provide the economic explanation for why people give and what determines how much they give. We will see that charitable giving can be stimulated by changing exogenous variables, ceteris paribus.

Our model will do the usual stripping away of realistic details, making incredible simplifying assumptions, to enable us to solve the model and play comparative statics games. Keep your eye on the procedure as we set up, solve, and compute our key measure—the tax break elasticity of giving.

An Endowment Model of Giving

As usual, we begin with the budget constraint, then we model preferences, and we use both to find the initial solution to the problem of maximizing satisfaction subject to the budget constraint.

The optimization problem is entirely from the donor's point of view. It is the donor, the giver, who decides how much, if any, to grant to the beneficiary, the recipient.

Figure 5.12 depicts the donor's budget constraint in this application. The initial endowment is the coordinate pair that represents the donor's consumption (on the y axis) and the beneficiary's consumption (on the x axis). There is only one good (which represents consumption of all goods) and its price is \$1/unit. So, if the donor has \$100 and the beneficiary only \$10, we know the initial endowment is at the point 10,100.

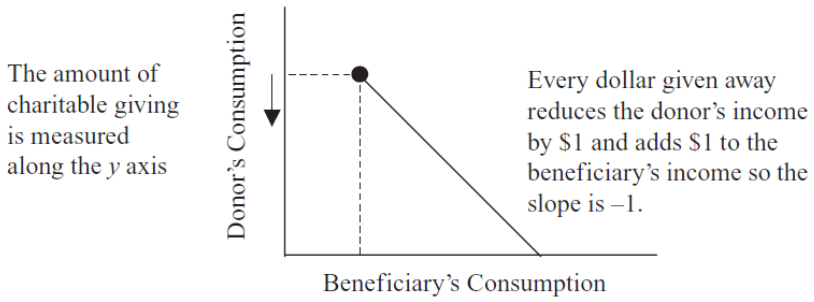


Figure 5.12: The budget constraint.

Giving is modeled as moving down the budget line in Figure 5.12. If the donor gives \$20 away, then she will have \$80 and the beneficiary will have \$30. Of course, the donor could give all of her money away, choosing to be at the x intercept. It is easy to see that the donor decides how much, if any, to give, by choosing a point on the budget line which determines both the donor's own consumption and the beneficiary's consumption.

Thus, at any point on the budget line, we can compute the amount of giving as simply the vertical distance (along the y axis) from the initial endowment to the point on the budget line. If the donor decides to stay on the initial endowment point, then they give nothing to the beneficiary.

The slope of the budget line is -1 because there is a dollar-for-dollar exchange from the donor to the beneficiary.

Notice that this budget line does not extend left or northwest from the initial endowment because that would imply taking money from the beneficiary. The donor cannot do that.

Finally, because we will (of course) be doing comparative statics analysis, we point out that a tax break for those who donate money means that the budget line will have a shallower slope. If the donor gives \$1 and is rewarded,

for example, with a 30¢decrease in taxes, then the recipient gets \$1, but the donor actually gave only 70¢. The slope is not -1, but $-(1 - TaxBreak)$. By adjusting the tax break, we can see how the agent responds.

This is too abstract. It is time to go to Excel to understand how the tax break really works.

STEP Open the Excel workbook *Charity.xls* and read the *Intro* sheet, then go to the *MovingAround* sheet.

All you see is a single point at 20,80—this is the initial endowment. The donor gives nothing and there is no tax break.

STEP Change cell C5, the amount the donor gives, to 20. The beneficiary gets the 20, adding it to his initial 20, and new red dot is at 40,60. The slope of the constraint is -1, displayed in I5.

Without a tax break, every dollar given is subtracted from the donor and added to the beneficiary. But the tax code incentivizes giving by lowering the donor's tax liability.

STEP Change E5, the amount of the tax break, to 40%. The red dot jumped up. Hit *ctrl-z* a few times to move back forth between zero and a 40% tax break.

With or without the tax break, the beneficiary still gets 20, but a tax break on charitable donations affects how much the donor actually gave up. With a 40% tax break, the sheet shows that the donor really gave up only 12 because taxes are lowered by 8 (40% of 20). Thus, the slope of the constraint is -0.6.

Wait, if the donor gives 12 and the recipient gets 20, who makes up the difference? The government. The beneficiary gets the full donation, but the donor pays less tax to the government. Clearly, by manipulating the tax break, the government can make charitable giving less expensive to donors.

So, if the tax break increases, what happens to the budget line? Think it through. You can check yourself when we get to the *OptimalChoice* sheet.

But before we get there, we have to consider the donor's preferences. The constraint is only about possibilities. To know what the donor will do, we need to know the donor's utility function.

The neat trick here is to enable the beneficiary's consumption to affect the donor's satisfaction. The way we model giving is to have the self-interested agent care about others.

The usual Cobb-Douglas functional form will represent the donor's satisfaction derived from her own consumption and the beneficiary's consumption.

$$U = BeneficiaryCon^{c} DonorCon^{d}$$

As usual, the exponents allow us to model different preferences. If c and d are equal, the donor gets as much satisfaction from her own consumption as the beneficiary's consumption. She is a saint. Although possible, this is unlikely. Most people get more satisfaction from their own consumption and, thus, d is greater than c.

We will use the *OptimalChoice* sheet with different exponent values to see the effect on the graph, but it is worth thinking through two scenarios. What would happen to the indifference curves, starting from $c = d$ as we lowered c? What would happen to the indifference curves if c fell all the way to zero? Again, thinking this through and testing yourself is good way to learn—you can check your answer in the *OptimalChoice* sheet.

It is worth remembering that preferences are not right or wrong. We take them as given and we model the agent as maximizing based on given preferences. It can be difficult to do this—we naturally disapprove of someone who doesn't care about others.

Another source of confusion is that preferences can and do change, but that is not to say that they are chosen by the agent. Changes to preferences are like shocks to other exogenous variables—they are imposed by forces outside the agent's control and then the agent re-optimizes in the new environment.

STEP Proceed to the *OptimalChoice* sheet to see how the donor's optimization problem can be implemented in Excel.

The sheet shows a mathematical expression of the constrained utility maximization problem. The constraint is different than usual. If we write the constraint as an equation, we need to compute the y intercept and incorporate the fact that the donor cannot take from the recipient (the empty space in the northwest corner of Figure 5.12).

We cannot use the usual Lagrangean method to deal with this complicated constraint because it only works with equality constraints. There is an analytical method called Kuhn-Tucker that can be used, but it is beyond the scope of this book.

Fortunately, the numerical method is still available. For Excel and Solver, the complicated constraint is easily handled by adding a second constraint (cell B26) and incorporating it as an inequality—this allows the donor to choose m_1 or greater for the beneficiary. The usual budget line constraint is in cell B25. Applying both constraints gives Solver the equivalent of Figure 5.12 and it has no trouble finding the optimal solution.

Figure 5.13 shows the starting position. The endogenous variables are consumption by beneficiary and donor. These are chosen by the donor to maximize utility subject to the budget constraint.

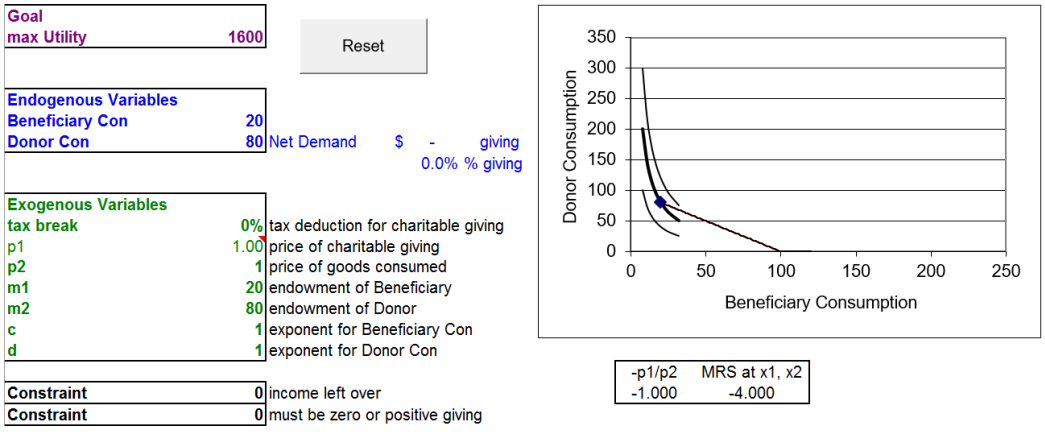


Figure 5.13: Donor with $c = d$ opening position.
Source: Charity.xls!OptimalChoice

The exogenous variables include the amount of the tax break (initially set at zero so the slope of the budget constraint is -1), prices normalized to one, the initial endowment, and the impact of donor and beneficiary consumption on the donor's utility.

With $c = d$, the donor cares as much about the beneficiary as herself and the MRS $> \frac{p_1}{p_2}$ at the initial endowment. We know the donor can increase her satisfaction by traveling down the budget line. For example, suppose the agent decided to donate \$10. How would this affect the chart?

STEP Change cell B11 to 30 and B12 to 70.

The MRS is now closer to the price ratio and utility has risen (from 1600 to 2100). The agent has moved down the budget line and is on a higher indifference curve.

STEP Run Solver to find the initial optimal solution.

The agent chooses the point 50,50 to maximize utility (at 2500), which means she donates \$30 to the beneficiary. The net demand is the amount of giving and we express it as a dollar amount and as a percentage of the donor's income (cell D13).

This is one mighty nice donor. She has an incredibly high giving rate of 37.5%. Because $c = d$, she cares as much about the beneficiary as she does herself. It makes common sense that she picks an equal 50,50 split as her optimal solution.

Comparative Statics

There are several shocks to consider. We start with preferences.

STEP Change the exponent for the beneficiary's consumption to 0.2.

This answers the earlier question about the effect of c on the indifference curves: they become much flatter as c falls, ceteris paribus. With $c = 0.2$, the donor does not care as much about the beneficiary as before.

The shape of the indifference curve is tied to the MRS. With $c = 0.2$, the MRS at 50,50 has fallen to 0.2 (in absolute value). The low MRS and flat indifference curve mean that the donor is willing to trade only a little of her consumption for a lot of additional beneficiary consumption.

The culmination of lowering c is a donor who does not care about the beneficiary at all. With $c = 0$, the indifference curves became horizontal, MRS is zero, and beneficiary consumption is a neutral good.

It is obvious that the donor with $c = 0.2$ is not going to be as generous as before when $c = 1$, but how much will they give?

STEP Run Solver. Figure 5.14 displays the result.

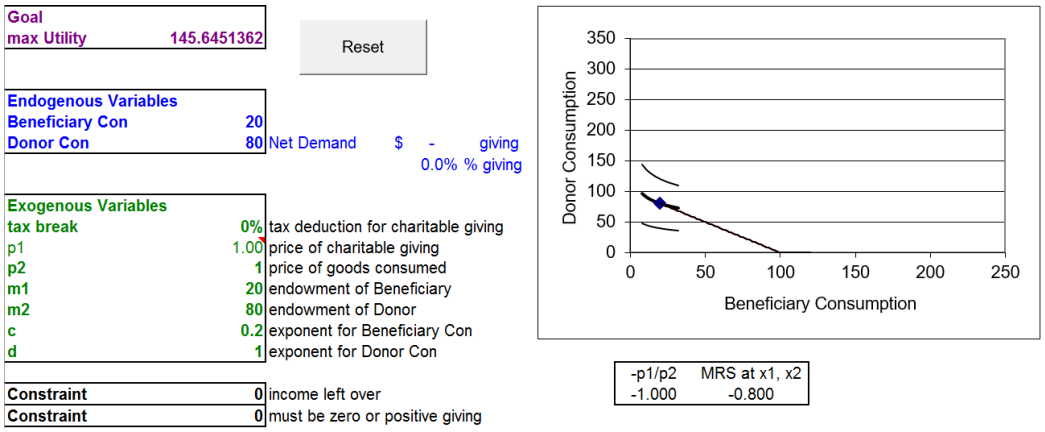


Figure 5.14: Donor with $c = 0.2$ corner solution.
Source: Charity.xls!OptimalChoice

The result is a surprise. The best the agent can do is to donate nothing so that is what she does. Even though the MRS does not equal the price ratio, this donor is optimizing. This is a corner solution.

Our work thus far provides answers to two of the three questions we initially asked.

1. Why do people give to charity? To maximize satisfaction. A donor gives because the consumption of others affects his or her utility. Notice that giving is perfectly compatible with self-interest. The economic model says that the donor feels good when she gives and that is why she gives.

2. What determines how much they give? Clearly preferences matter. How much the donor cares about others (the exponent c in the donor's utility function) plays a major role. Of course, the constraint also matters. Donor's income, beneficiary's income, and the slope of the constraint affect the amount of giving.

3. How can charitable giving be stimulated?

Let's work on the third question. We could try to convince people to care more about others, increasing c (certainly this is a primary goal of religion), but a way to stimulate giving is to lower the price of giving.

As we saw earlier, dollars given to charity reduce the donor's taxable income and reduce tax owed. If the donor is in a 30% tax bracket, every dollar donated to charity saves the donor 30 cents in taxes. Thus, the beneficiary receives the dollar, but the donor is actually paying only 70 cents—with Uncle Sam picking up the remaining 30 cents.

What effect will a 30% tax break have on the budget constraint and charitable giving of a donor with $c = 0.2$? Apply the shock in Excel and find out.

STEP Change the *tax_break* variable (B16) to 30% and note that p_1 becomes 0.70 and the budget line swings out.

The new red budget line is flatter than the original because of the tax break. This answers the earlier question about the effect of a tax break on the budget constraint: the bigger the tax break, the more the line swings and flattens out. This is just like lowering p_1 in the Standard Model.

Notice that the MRS is greater than the slope of the new budget line. This agent can improve her utility by traveling down the constraint. This means she will donate to the beneficiary, as shown in Figure 15.15.

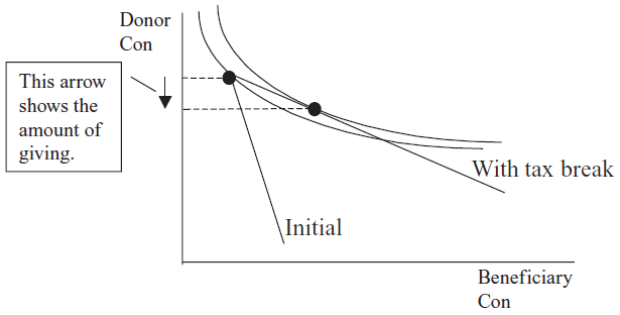


Figure 5.15: The effect of a tax break on giving.

But exactly how much giving does the tax break generate? Let's find out.

STEP With $c = 0.2$ and *tax_break* = 30%, run Solver.

In this case, the tax break has induced charitable giving. It is hard to see on the graph, but the MRS $= \frac{p_1}{p_2}$ condition (under the chart) tells you the indifference curve is now tangent to the budget line. Figure 15.5 shows what happened.

With a tax break of 30%, we get \$1.67 of giving which is 2.1% of the donor's income (the American giving rate in 2018).

We can also explore how responsive our donor would be to further shocks in the *tax_break*. We will compute the tax break elasticity of giving.

STEP Change the *tax_break* cell to 40%.

That's a 10 percentage point change in the tax break and a rather hefty 33% change. The budget line swings out a little bit more, but it is hard to see the change in the chart. We know, however, since MRS does not equal $\frac{p_1}{p_2}$, that we need to re-optimize.

STEP Run Solver.

Charity increased from \$1.67 to \$3.33. That is a big response—a doubling or 100% increase in giving was generated from a 33% increase in the tax break. That is a tax break elasticity of giving of 3.

STEP Proceed to the *CS1* sheet to see a more detailed comparative statics analysis.

Notice that the shock was 1% point, not 10. Notice also that the elasticity from a tax break of 30% to 31% is about 2.87 (H17), not 3. Even though we do not have a reduced form expression, the fact that the measured elasticity depends on the size of the shock tells us that giving is a non-linear function of the tax break.

But regardless of whether it is 3 or 2.87, that high an elasticity is really good news, right? If giving is super-responsive to a tax break, little tweaks in the tax break will generate big increases in giving.

But we need to be careful in how we interpret our result. We do not know whether these preferences and other exogenous variables are representative of many donors. That is an empirical question that requires real-world data. For example, with $c = 0.5$, tax break increases are much less effective in stimulating more giving.

STEP Click the Reset button, change c to 0.5 and the tax break to 30%, and run Solver. Charitable giving is at \$17.33.

This makes sense since giving is much higher than it was when $c = 0.2$ and $tax\_break = 30\%$. But what is the tax break elasticity of giving?

STEP Change the tax break cell to 40% and run Solver. Charitable giving rises to \$18.67.

Ponder the computation for a moment. There are a lot of numbers floating around. How would you compute the tax break elasticity of giving?

It is the percentage change in giving divided by the percentage change in the tax break. The numerator is $\frac{18.67-17.33}{17.33} \approx 7.7\%$. The denominator is 33% ($\frac{0.4-0.3}{0.3}$—notice that it doesn't matter if you use the percents version, $\frac{40\%-30\%}{30\%}$). Thus, the tax break elasticity is $\frac{7.7\%}{33\%} = 0.23$.

This result is much less favorable for a policymaker looking to increase charitable giving by manipulating the tax break. For this donor, giving is insensitive to tax break increases.

The Theory of Consumer Behavior can explain a wide variety of giving outcomes. Unfortunately, theory alone does not tell us about the magnitude of a particular effect in the real world. By changing c, we see that the tax break elasticity of giving is drastically affected, ranging from extremely elastic (3) to quite inelastic (0.23). We must gather data and employ econometric techniques to estimate the responsiveness of giving as the tax break changes in the real world. Theory does, however, give us a framework for analyzing the problem.

The Economic Approach Is Widely Applicable

Charitable giving can be viewed through the lens of an Endowment Model using the Theory of Consumer Behavior. The initial endowment is the consumption of the donor and the beneficiary. The donor can choose to give part, all, or none of her endowment to the beneficiary. The amount she gives is determined by that point that maximizes her satisfaction subject to the budget constraint.

We can stimulate giving by lowering the price of giving. This rotates the budget line and yields a new optimal solution. The amount of the increase in giving is an empirical question that cannot be answered by theory alone.

If we view giving as the solution to an optimization problem, we are doing an economic analysis of giving. "An economic analysis" is a phrase often used to communicate that the behavior under consideration will be cast in the framework of optimization and comparative statics.

Many people think economics is about stocks, business, and money. This content-based definition of economics is too limited. Economics is a method of analysis and it can be applied to such "non-economic" issues as charity and many, many other areas.

Seeing charitable giving through the lens of economics does not mean that this is the only way to study charity. The hope is that it provides insight and furthers understanding of what is surely a multifaceted, complex process.

Exercises

1. The total change in charitable giving can be explained via the income and substitution effects for giving. For $c = 0.5$, compute the income and substitution effects when the tax break changes from 30% to 40%. Describe your procedure.

2. Use Word's Drawing Tools to draw a rough sketch of the income and substitution effects for giving, labeling points A, B, and C and using arrows to show the income, substitution, and total effects. Do not include the indifference curves to reduce clutter.

3. Income and substitution effects were originally used to explain Giffen goods. If the tax break increase leads to a decrease in charitable giving, is this Giffen behavior? Why or why not?

References

The epigraph is a *hadith*, which the website islam.uga.edu/hadith.html explains is "a saying of Muhammad or a report about something he did." It would have been easy to find a quotation on charity from any religion because a primary purpose of religion is to encourage us to treat each other with kindness.

If you are thinking of giving to a charitable organization, you can do some background research at www.guidestar.org/ (free registration required to access basic reports) and www.givewell.org/.

Kiva.org is a microcredit organization that allows you to make loans to low-income entrepreneurs all around the world.

If you liked the food stamps application and understand the concept that cash is as good as or better than in-kind (the Carte Blanche Principle), check out www.givedirectly.org.

During the early 1960s, Kenneth Arrow and Karl Borch published several important articles that can be viewed as the beginning of modern economic analysis of insurance activity.

Georges Dionne and Scott E. Harrington

5.4 An Economic Analysis of Insurance

Why do people buy insurance?

If you are an economist, the answer is easy: because it makes them better off. According to economists, people solve an optimization problem and it turns out that those who buy insurance end up with greater satisfaction, on a higher indifference curve, than if they did not buy insurance.

We will use an Endowment Model to explain how and why insurance is an optimal choice. We will see yet another application of how to solve a constrained utility maximization problem and perform comparative statics analyses.

But the really deep lesson is that the Theory of Consumer Behavior is amazingly flexible and can answer questions from a wide range of problems. In this chapter, we have explored why people save and borrow, give to charity, and, now, buy insurance.

First, we will set up the problem with the usual constraint, indifference curves, and initial optimal solution (with MRS equal to the slope condition). The presence of *risk*, a probability that an event occurs, throws a curveball into the analysis, but we will convert things into our usual framework.

Second, we will do comparative statics. For example, we derive a demand curve for insurance. We can explore the effects of a higher *premium*, the price of insurance, on the quantity of insurance demanded. We are on the lookout for the premium elasticity of insurance.

An Endowment Model of Insurance

There are three parts to every optimization problem. In this case, we have the following:

1. *Goal*: maximize satisfaction (as represented by the utility function).
2. *Endogenous variables*: consumption in two states of nature, good and bad; by choosing the amount of insurance, we control two choice variables at once.
3. *Exogenous variables*: initial assets, potential loss, probability of loss, insurance premium, and preferences over the states of nature.

As usual, we start with the constraint, then turn to preferences, and finally use the constraint and utility function to find the initial solution.

STEP Open the Excel workbook *Insurance.xls* and read the *Intro* sheet, then go to the *Constraint* sheet.

The idea is that you have an asset, say your car or house, which may suffer a given amount of damage from an accident, called the *PotentialLoss*, with a known probability, π (the Greek letter, pi) that the damage occurs. Initially, the *PotentialLoss* is \$10,000, which is only a fraction of the value of the house.

You can buy K dollars of insurance, this is the *InsuredAmount*, by paying a price (called a premium) of γ (the Greek letter, gamma) per \$100 of insurance coverage. On opening, you are not buying any insurance.

If you buy insurance, then if the accident occurs, you get reimbursed for the loss. You can buy insurance in \$100 increments, up to the *PotentialLoss*, in which case you would be fully insured. The trade-off is that you have to pay for insurance up front, before you know if the accident will happen or not.

After you decide how much insurance to buy, there are two possible outcomes, known as *states of nature*: the bad and good outcomes.

STEP Click on cell B18 to see the formula for your consumption in the bad outcome.

The *ConsumptionBad* outcome means the accident actually occurred, leaving your consumption as $InitialAssets - PotentialLoss + K - \gamma K$. You subtract the loss that occurred and the amount you paid for insurance (γK), but you add the amount K that the insurance company pays you because you suffered the accident. You could be fully covered, but you do not have to be. You decide how much insurance to buy.

Your consumption in the good state of nature is simply $InitialAssets - \gamma K$. You do not suffer the accident, but you still have to pay for the insurance.

STEP Click on cell B19 to see the formula for the good outcome.

Cells B23:B25 display in which state of nature you end up. Cell B23 has the formula "=RAND()." This draws a number from a uniform distribution on the interval [0,1].

STEP Hit the F9 key on your keyboard repeatedly to understand Excel's RAND() function works.

Each time you hit the F9 key, Excel draws a random number from 0 to 1 in cell B23. The number drawn is never smaller than zero or bigger than one.

Cell B24 converts the random draw in the cell above it into a zero or a one—zero means the accident did not happen (good outcome) and one means it did (bad outcome). It uses an IF statement to display a "1" (the accident happened) when the random draw is less than 0.01 (the value of π in cell B8).

It is hard to see that anything is really happening in cell B24 because the probability of the accident occurring is so small.

STEP Change π to 50%, then hit F9 a few times. You should be able to see cell B24 flip from 0 to 1 and back again as the random draw is less than 0.5 and greater than 0.5.

Notice that the *FinalAssets* variable, cell B25, depends on whether or not the accident occurred.

Next, let's buy some insurance to see what that does to the spreadsheet.

STEP Click the Reset button and set cell B13 to \$1000. This will cost you \$10.

Notice the values for the good and bad states of nature. You have altered both. If the accident occurs, your consumption is \$25,990, which is \$990 better than the \$25,000 for the bad outcome when you did not buy insurance. Of course, the good outcome is \$10 lower (at \$34,990) in the good outcome because you have to pay for the insurance even when the accident does not occur.

STEP Click the Graph the Constraint button. Click OK to the "4" points default option and read each text box as it appears. At the end, the budget line is displayed (see Figure 5.16).

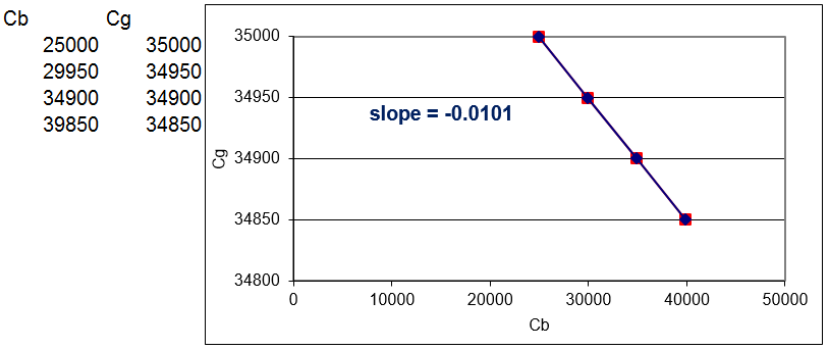


Figure 5.16: The budget line.
Source: Insurance.xls!Constraint

From the initial endowment (C_b, C_g without insurance), you can move down the budget line by buying insurance. You lower your consumption in the good state of nature (C_g is on the y axis), but raise it in the bad state of nature (C_b is on the x axis).

The terms of trade (the slope of the budget line) are determined by gamma (the insurance premium). The slope of the budget line is $-\frac{\gamma}{1-\gamma}$, which with $\gamma = 0.01$ is $\frac{-1}{99} = 0.01$ (the "01" keeps repeating forever). The graph rounds the slope to five decimal places.

Changes in initial assets or potential loss shift the budget constraint. We are interested, however, in deriving a demand curve for insurance so we will shock the insurance premium (the price of insurance). This will pivot or rotate the budget line.

STEP Change the insurance premium to \$1.20 per \$100 of insurance coverage.

You see the familiar swinging in (clockwise rotation) from a p_1 increase. A buyer of insurance would be disappointed in this shock because her consumption possibilities are diminished.

Now that we understand the constraint, we turn to the agent's tastes. We model utility as preferences over the two states of nature. The fact that there is risk involved in which state of nature occurs complicates things.

Instead of having utility simply depend on the amount of consumption in the good and bad outcomes, we include the agent's expectations about the chances of each outcome occurring. Fortunately, our usual Cobb-Douglas functional form can incorporate this new information.

We use the exponents in the Cobb-Douglas functional form to represent the agent's beliefs about the probability of the accident occurring. There are two simplifying assumptions. The first is that the agent accurately gauges the probability of loss, which means we can use π as the exponent in the utility function. The second assumption uses the fact that there are only two mutually exclusive outcomes so the bad outcome occurs with probability π and the good outcome has likelihood $1 - \pi$. The possibility of a partial loss is assumed away.

The utility function is then

$$U = C_b^{\pi} C_g^{1-\pi}$$

The idea behind the utility function is simple: The higher the probability of loss, the more the agent will care about the bad outcome. In terms of the indifference map, the higher π, the steeper the indifference curves. This means the agent cares more about consumption in the bad state of nature as risk rises.

Unlike the Standard Model where the exponents in the Cobb-Douglas utility function can be used to represent changes in preferences, changes in the exponents do not indicate a change in preferences for the utility function with risk. To get a change in preferences, we need an entirely different utility function.

It is beyond the scope of this book, but there is a great deal of research on choosing with random outcomes. The field of behavioral economics was born with the discovery of paradoxes, violations of transitivity and other inconsistencies, when people made choices involving randomness. Our Cobb-Douglas utility function can be written as an expected utility function by simply taking the natural log:

$$lnU = \pi C_b + (1 - \pi) C_g$$

This function reflects risk averse preferences. It is a starting point for modeling attitudes and feelings toward risk and randomness.

STEP Proceed to the *Preferences* sheet to see an implementation of the Cobb-Douglas utility function.

The sheet tries to give a new way of understanding how constrained utility maximization works. It shows consumption in the bad and good states of nature, \$25,000 and \$35,000, respectively, without insurance. This is the initial endowment point.

With $\pi = 1\%$, we can compute the level of utility for the initial endowment combination of consumption in the bad and good states of nature. This is shown in cells D13 and E13. We can also compute the MRS at the initial endowment, displayed in cells G13 and H13.

The *Dead* and *Live* utility and MRS are the same because we are at the initial endowment. The *Dead* cells are numbers. They will not change when we change the cells in column B. The *Live* cells contain formulas. They will update when you change the values of $C_b, C + g$, and π.

STEP Ponder and answer the question in cell A6. Click on the Answer when you are ready. Do the same for B10.

The Live utility and MRS cells change when you change cells B13 and B14. As you moved down from the initial endowment, utility rose and the MRS fell. It got closer to the slope which means we are closer to the optimal solution.

We are ready to find the initial optimal solution.

STEP Proceed to the *OptimalChoice* sheet.

The *OptimalChoice* sheet reproduces the *Constraint* sheet, but it adds the indifference map to the chart and displays the slope of the budget line and the MRS at the bottom of the chart. It also displays the utility in cell B20 from the chosen consumption in the two states of nature.

It is really hard to see what is happening with the indifference curve at the initial endowment and the slope of the budget line.

STEP Zoom in—double-click the y axis and make the minimum bound 34800 and the maximum bound 35200.

You can now see clearly that when MRS > slope of the budget line, the budget line cuts the indifference curve. By moving down the budget line, you can reach higher levels of satisfaction.

STEP Enter 5000 in cell B13 to see where the agent stands when buying $5000 of insurance.

The chart shows movement down the budget line to a higher level of utility. We are closer to the optimal solution, but not there yet because MRS is not equal to the slope of the budget line.

STEP Run Solver to find the optimal solution.

The Solver dialog box is notable for the fact that there are no constraints. The way we implemented the problem in Excel enabled us to directly maximize the utility cell by choosing a single variable, the amount of insurance purchased. We can still use, however, the canonical Theory of Consumer Behavior graph to show the result.

At the optimal solution, the consumer decides to buy $10,000 of insurance. In the bad state, if the accident occurs, the agent is fully covered, so is consumption $35,000? No, because the agent has to pay $100 for the insurance, so consumption would be $34,900 in the bad state.

In the good state, where there is no accident, consumption is also $34,900. This is surprising. Insurance has removed the effect of risk. Consumption is the same in both states. This is an extreme example of diversification.

Diversification is a strategy to lower risk by spreading your wealth over different states of nature. By moving $100 from the good state of nature (buying insurance), the agent has a guaranteed level of utility regardless of whether the accident happens. Without insurance, the expected return is $34,900 since 99% x $35,000 + 1% x $25,000 = $34,900. But the agent has to put up with the risk of every 1 in 100 times getting $25,000. By diversifying, the expected return is the same, $34,900, with absolutely no risk.

Such a perfect result—the complete elimination of risk—relies on the fact

that the two states of nature are perfectly correlated. In the real world, when states of nature are not perfectly correlated (such as the stock market), diversification can lower risk while maintaining the same expected return, but it cannot completely eliminate it.

We know that people buy insurance because it increases satisfaction. This application models choosing the amount of insurance that maximizes utility subject to the budget constraint. Next, we use the model to derive a demand curve for insurance.

Comparative Statics

The procedure is straightforward: we vary the insurance premium (the price of insurance), γ, ceteris paribus, and track the optimal amount of insurance purchased (K) to derive a demand curve for insurance.

We use numerical methods and leave the analytical work for the exercises.

STEP In the OptimalChoice sheet, change γ to \$1.30 per \$100 of insurance. What happens?

The budget line (displayed in red on your screen) gets steeper. The agent needs to re-optimize.

STEP Run Solver to find the new optimal solution.

If you did not zoom in on the y axis as instructed earlier, it is hard to see on the chart, but the cells below the chart confirm that the MRS equals the slope of the budget line when the agent buys \$1847 of insurance.

We can conclude that demand for insurance is downward sloping when the premium rises from \$1.00 to \$1.30 since the amount of insurance purchased fell from \$10,000 to \$1847. That is extremely responsive.

STEP Compute the price elasticity of demand. Proceed to the *CSgamma* sheet to check your answer. Notice that Excel tries to help when you enter the formula by formatting the result as dollars. This is incorrect. Elasticity is unitless.

The *CSgamma* sheet shows that the CSWiz add-in was used to explore the effect of the insurance premium on the amount of insurance purchased. Gamma

was incremented by 0.1 (10 cents) with 10 shocks. Optimal $K, \gamma K, C_b$, and C_g were tracked as γ changed. The sheet includes a chart of $K^* = f(\gamma)$, the demand curve for insurance.

Notice the curious behavior of the model as γ rises: at \$1.40, optimal K becomes negative. This is an Endowment Model. When premium prices get high enough, the agent switches from buying insurance to selling insurance!

If this option is not allowed, you can impose the constraint in Excel that K be greater than or equal to zero. Then, with high premiums, the consumer is at a corner solution and buys no insurance.

Modeling Insurance via the Endowment Model

Insurance is another application of an Endowment Model in the Theory of Consumer Behavior. The usual ideas were applied: the budget constraint, preferences, and MRS equals slope of budget line at the optimal solution. In addition, the usual recipe of the economic approach, finding the initial optimum and then comparative statics, was followed.

But this application does have its own twists and novelties. We used a Cobb-Douglas functional form to model satisfaction where the exponents reflect the probabilities of the states of nature. We also used Excel's Solver without a budget constraint because of the way we implemented the problem in Excel. To be clear, this problem can be solved via the Lagrangean method (see the first exercise question) and we could have implemented a "max U subject to a constraint" model in Excel. We would get, of course, the same answer.

Exercises

1. Use analytical methods to derive a general reduced form solution for K^*. Show your work.

   Although you can use the Lagrangean method, it is easier to maximize the utility directly, substituting in the values for each state of nature.

$$\max_K U = C_b^{\pi} C_g^{1-\pi}$$

The key is that consumption in the good and bad states of nature depends on K:

$$C_b = InitialAssets - PotentialLoss + K - \gamma K$$

$$C_g = InitialAssets - \gamma K$$

We can simply substitute these equations into the utility function and maximize this:

$$\max_K U = [InitialAssets - PotentialLoss + K - \gamma K]^{\pi} [InitialAssets - \gamma K]^{1-\pi}$$

2. Compare the analytical versus numerical approaches by evaluating your answer to question 1 at the initial parameter values in the *Optimal-Choice* sheet. (Click the Reset button if needed.) Do you find that $K^* = \$10,000$?

3. Use your reduced form for K^* to find the probability of loss elasticity of insurance demand at $\pi = 1\%$. Show your work. If you cannot find the reduced form, use

$$K^* = \frac{[\pi - \gamma]\, InitialAssets + [1 - \pi][\gamma] PotentialLoss}{[\gamma][1 - \gamma]}$$

4. Use the Comparative Statics Wizard to find the probability of loss elasticity of insurance demand from $\pi = 1\%$ to 1.1%. Take a picture of your results, including the elasticity calculation.

5. Compare your answers in question 3 and 4. Do these elasticities differ? Why or why not?

References

The epigraph is from the first page of *Foundations of Insurance Economics* by Georges Dionne and Scott E. Harrington, editors, published in 1990. Insurance economics as an organized subfield is quite young, but rapidly growing. It focuses economics, probability, and computer science on applied problems in the world of risk and insurance.

In their wildly popular *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything* (2005), Steven D. Levitt and Stephen J. Dubner include this example from the world of insurance markets:

> In the late 1990s, the price of term life insurance fell dramatically. This posed something of a mystery, for the decline had no obvious cause. Other types of insurance, including health and automobile and homeowners' coverage, were certainly not falling in price. Nor had there been any radical changes among insurance companies, insurance brokers, or the people who buy term life insurance. So what happened? The Internet happened. In the spring of 1996, Quotesmith.com became the first of several websites that enabled a customer to compare, within seconds, the price of term life insurance sold by dozens of different companies. (p. 66)

The freakonomics.com website has podcasts and other resources.

Chapter 6

Bads

Risk Versus Return

Automobile Safety Regulation

Labor Supply

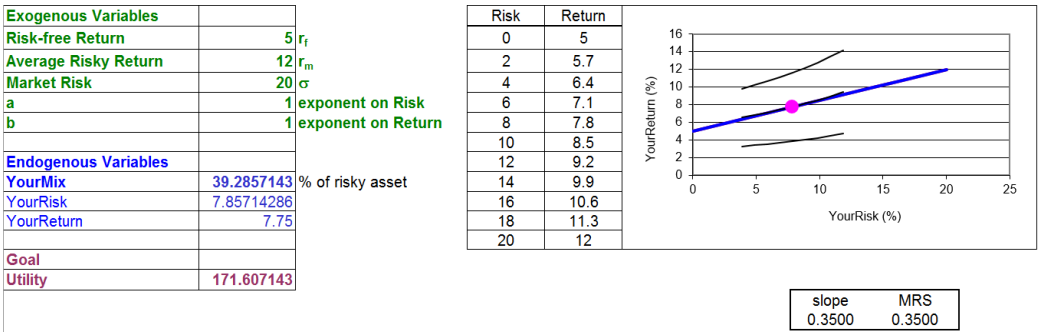
> One of the best-documented propositions in the field of finance is that, on average, investors have received higher rates of return for bearing greater risk.
>
> Burton Malkiel

6.1 Risk Versus Return

In finance, a *portfolio* means the total holdings of stocks, bonds, and other securities of an individual (or other entity, such as a trust or foundation).

Because the investor can decide which securities to include in her portfolio, in other words, because choices are made, we can apply the method of economics. Optimal Portfolio Theory is the name given to the application of the Theory of Consumer Behavior to analyze decisions about which assets to hold.

An important stop on our journey is shown in Figure 6.1, the initial solution to the constrained optimization problem.



Exogenous Variables		
Risk-free Return	5	r_f
Average Risky Return	12	r_m
Market Risk	20	σ
a	1	exponent on Risk
b	1	exponent on Return
Endogenous Variables		
YourMix	39.2857143	% of risky asset
YourRisk	7.85714286	
YourReturn	7.75	
Goal		
Utility	171.607143	

Risk	Return
0	5
2	5.7
4	6.4
6	7.1
8	7.8
10	8.5
12	9.2
14	9.9
16	10.6
18	11.3
20	12

slope	MRS
0.3500	0.3500

Figure 6.1: The initial solution.
Source: RiskReturn.xls!OptimalChoice

There are some strange features in Figure 6.1 and you are not expected to understand it right away. Perhaps the weirdest thing is that the budget constraint and indifference curves are upward sloping. Because risk (on the x axis) is a bad (not a good), the agent substitutes *more* of the bad for *more* of the good (return, on the y axis) on an indifference curve.

There are also, however, elements that are familiar and comfortable in Figure 6.1. There are exogenous (green) and endogenous (blue) variables with a goal. There is a constraint and a few curves with a tangency highlighted that is obviously the optimal solution. And we can see the usual MRS = slope condition below the chart.

Of course, Figure 6.1 is just the initial optimal solution. There is more to do than simply finding the initial solution. That is why Figure 6.1 is an important stop on our journey, but we have more to travel. We want to explore how the optimal solution changes as one of the exogenous variables changes, ceteris paribus. This is called comparative statics analysis.

The procedure that defines the Theory of Consumer Behavior is clear: constraint, preferences, find initial solution, then comparative statics to make statements about how a shock variable affects an optimal choice variable. We will do an elasticity computation and interpretation of the shock. The short way of saying all of this is to just say that we are going to do an economic analysis of portfolio choice.

But since we will be talking about returns from assets, volatility, and the stock market, let's look at some data to make sure we understand some basic facts.

Stock Market Returns

STEP Open the Excel workbook *RiskReturn.xls* and read the *Intro* sheet, then go to the *Data* sheet.

The sheet has returns from the S&P 500 index, a group of 500 large companies, downloaded from www.moneychimp.com/features/market_cagr.htm.

These data are used to show that returns are quite volatile. The sheet also explains the difference between the arithmetic and geometric mean.

STEP Read the explanation in the *Data* sheet, scroll down to see the data (all the way down to 1871), and then click the Next button.

This reveals more material. Keep reading and clicking the Next buttons until you get to the end and then be sure to click the More button.

Of utmost importance is that you understand the volatility in the S&P 500 returns. They swing wildly and unexpectedly, from incredible spurts of 50% to staggering losses of almost negative 50%.

STEP Look in columns A and B of the *Data* sheet at the 1930s, during the Great Depression. Scroll slowly back up, looking at the data.

The volatility in the stock market, measured by the standard deviation, SD, of almost 20%, is unwelcome and unsatisfying. The fear of financial disaster and the risk of losing money lowers utility.

Then why do people put their money in assets like the S&P 500? Because the overall annual return is high—much higher than safer, less volatile assets. For the S&P 500, the overall annual return (as you now know, measured by the geometric mean, GM, or compound annual growth rate, CAGR) is about 9% per year.

The stock market's 9% annual return is much higher than that available from a safe, stable asset that produces consistent annual returns like US Treasury Bills. Cell H10 in the *More* sheet shows that the SD is a mere three percentage points. The variability arises because the yield changes over time, but once you buy a US Treasury note for a particular length of time, you can be quite sure that you will be paid. But right below the SD we see that the overall annual return is one-third of the stock market's return.

The key point is that financial markets offer the investor a menu of options, from low risk, low return to high risk, high return, and the investor chooses. All we need to do is model that choice as an optimization problem.

Optimal Portfolio Theory

The *Compare*, *Mix*, and *Constraint* sheets in *RiskReturn.xls* demonstrate that an investor can mix two assets, a risk-free and a risky asset, to create a portfolio that has a particular combination of risk and return.

The investor is not free to pick any combination of risk and return. They must stay within the constraint imposed by the market. The idea is that you have a fixed amount of money, say $10,000, to allocate across two assets.

The *risk-free asset*, say a US Treasury Bill, has a certain (practically speaking) rate of return, say 5% per year, which is unrealistically high for the current climate. Thus, you are sure to get 5% of \$10,000, or \$500, along with your initial investment of \$10,000 at the end of the year. Each year, a \$10,000 investment is guaranteed to produce \$500 of return.

The *risky asset*, say a mutual fund of stocks, has a greater return, but also volatility in the actual realized return. We will suppose that the actual return will be drawn from a normal distribution centered on 12%, with a spread of 20%. Both of these values are a little higher than the historical experience of the S&P 500 (in the *Data* sheet). Our parameter values mean that the typical realized value in our hypothetical world will be around 12% ± 20% points. It also means you will actually lose money (suffering a negative return) about a quarter of the time.

But this is way too abstract. To understand the meaning of these parameters, let's work on a concrete problem with actual numbers and a clear display of what is going on.

STEP Go to the *Compare* sheet.

The bell-shaped curve is the normal distribution from which each year's return will be drawn. The center and spread are controlled in cells A2 and C2.

The sheet allows you to run the two investments against each other and shows how volatility impacts the annual returns.

STEP Click the Invest One Year button.

For the risk-free asset, cells I3 and L3 show 5% and \$500. In other words, if you place \$10,000 in the risk-free asset, these are the returns on that investment.

The risky asset is different. Cells J4 and M4 show a number that is taken from the normal distribution on the left of your screen, centered on 12 with an SD of 20. Thus, the number in J4 is likely to be around 12, but could easily be in the range −8 to 32 (±1 SD from the average) and roughly 95% of the time will be between −28 and 52 (±2 SDs from 12).

STEP Click the Invest One Year button a few times.

You can clearly see what is happening here. The return from the risk-free asset is always the same, but the risky asset bounces around.

Once you have more than one year of returns, the display shows more information in columns P:S. You can see the arithmetic mean of the returns, SD, the exact geometric mean, and its approximation.

STEP Click the Invest One Year button many times, at least 20.

Notice what is happening to the average of the returns of the risky asset as you keep adding years: The average return is converging to 12% (the average return from the normal distribution in A2). In other words, over the long haul, the risky asset will outperform the risk-free asset. However, in any one year, the risky asset can do pretty badly. Look at your screen to confirm that this is true. You will see some whopper losses (and gains)—just like the real-world S&P 500 data.

STEP Click the Reset button and set the dispersion to 6% (in C2). Repeatedly (many times) click the Invest One Year button.

The SD of the normal distribution controls the variability. The lower SD makes the normal distribution much more spiked. In other words, the draws from the distribution are much more concentrated at the average and it is much less likely that you will see values far from the center of the distribution.

As you get one yearly return after another (keep drawing more returns), it is easy to see that the returns are much closer to 12%. You will rarely lose money with an average of 12% and an SD of 6%.

In finance, risk is denoted by the Greek letter sigma, σ. The SD and σ are the same thing. Both represent risk as volatility and bounce in returns, including the possibility of negative returns. Risk is bad and undesirable. The lower the risk, the better.

What determines the amount of risk in the risky asset? That depends on the asset. We have seen that the S&P 500 has a lot of volatility. From 1871 to 2019, it has experienced an overall annual return of about 9% with an SD of 18%. The *More* sheet showed that other assets have different volatility. So, the investor is given the average and SD parameters of various assets and chooses what to invest in.

Although we ran risk-free and risky assets in the *Compare* sheet, in fact, the choice is not simply between a risk-free and a risky asset. You can combine the two in varying proportions.

For example, you could split your investment and put \$5000 in the risk-free and \$5000 in the risky asset. In this case, your return would be halfway between the risk-free and risky assets:

$$\frac{r_f + r_m}{2} = 8.5\%$$

Although the return is lower than using the risky asset alone, your risk, the variability in returns, would be cut in half also.

STEP Proceed to the *Mix* sheet to see this idea in action.

The *Mix* sheet is the same as the *Compare* sheet, except it has a scroll bar in H1 to control the allocation of your \$10,000 across the two assets.

STEP After you set the scroll bar value (any value will do; pick the one you think makes the most sense for you), click the Invest One Year button many times.

You should be able to see that the average return for your mix (or portfolio) converges on a return that is in between the risk-free and risky assets. In other words, you can choose the return and risk that you get. You must, however, trade them off—more return requires accepting more risk.

STEP Experiment. Use the Reset button to try different mixes and parameter values (yellow-backgrounded cells A2, C2, and F2).

You can copy the *Mix* sheet (right-click the sheet tab, select *Move or Copy*, and check Create a Copy) if you want to compare different scenarios. The more you experiment, the more you learn.

Your work in the *Compare* and *Mix* sheets makes understanding the constraint much easier because you have seen that there are two assets that can be mixed to form a portfolio with a continuous range of risk and return possibilities. This constitutes the constraint for the investor. He or she is free to choose combinations of risk and return, trading higher risk for greater return.

STEP Proceed to the *Constraint* sheet.

There are two endogenous variables, *YourRisk* and *YourReturn*, in cells B14 and B15. These are the risk and return you have chosen, in other words, a single point on the budget line. However, we can create a single variable, *YourMix* (just like in the *Mix* sheet) that controls the proportion of your investment in the two assets and the values of risk and return you select.

Clearly, you can mix the risk-free and risky assets in any combination from 0 to 100%. Zero means you buy just the risk-free asset and 100% means you buy only the stock market.

Do not confuse the exogenous variable *Market Risk* with the endogenous variable *YourRisk*. The riskiness of the risky asset, sigma, is exogenous to the agent. But the agent determines how much risk to take and, therefore, the chosen amount of risk is endogenous.

STEP Change B13 to 20%, 50%, and 90%.

As you change B13, the red dot moves on the constraint. You can put the red dot wherever you like along the line. At 50%, you are setting *YourRisk* to 10% (this is the variability in the 50/50 portfolio) and *YourReturn* to 8.5% (halfway between r_f and r_m).

The equation of the budget line (derived in the *Constraint* sheet) is

$$YourReturn = r_f + \frac{r_m - r_f}{\sigma} YourRisk$$

Clearly, if you choose a risk of zero, then your return is the risk-free return. This is the y intercept. As you accept more risk, your return grows with a slope given by $\frac{r_m - r_f}{\sigma}$

Notice that combinations under the budget constraint are feasible, but will not be selected because more return can always be obtained at the same risk by going straight up. Points to the northwest of the line are more desirable, but are unattainable.

Which mix is the best, the optimal choice? We cannot answer this question with the constraint alone. It tells us only the choices we can make. To answer the question, we need to model preferences.

But before we leave the constraint, let's explore the effect of a change in sigma, Market Risk. This will be our shock variable when we do comparative statics analysis.

Remember when you lowered the SD to 6% and that made the variability in the risky asset go way down? That was a welcome shock. What would happen to the constraint if we applied that shock? Before we do it, ponder the question. Do you have an answer? Let's see how you did.

STEP Change Market Risk, cell B10, to 6.

The budget line rotates up (counterclockwise) around the y intercept. This gives the investor access to higher returns with the same risk or the same return with less risk. Mathematically, it also makes sense since we lowered the denominator in the slope, so the slope term increased, making the line steeper.

STEP Proceed to the *Preferences* sheet to see how we handle risk as a bad.

Our usual Cobb-Douglas functional form can be modified to reflect a bad with a simple tweak:

$$U(YourRisk, YourReturn) = (30 - YourRisk)^a YourReturn^b$$

The clever trick here is subtracting a variable from a constant, which has been chosen to be bigger than the possible values of the variable. By having a constant, 30, which is a bigger number than the relevant range for *Risk* (from zero to 20), as we increase the chosen amount of *YourRisk*, 30–$YourRisk$ falls. This gives us a bad because utility falls as *YourRisk* rises (for $YourRisk < 30$). *YourReturn* is a good—as *YourReturn* rises, so does utility.

The chart shows three representative, upward sloping indifference curves. The investor gets equal satisfaction by the combinations of risk and return on a single indifference curve. If the investor takes on more risk, she must be given more return to compensate.

STEP The agent is free to choose any combination of risk and return that is on the budget line. Change B12 to 50.

Figure 6.2 shows the result. In addition to the three original indifference curves with a black dot, three new curves are displayed along with a red dot. The black dot is the initial 75% mix choice and it produced *Dead Utility* of 153.75 and a *Dead MRS* of about 0.6833.

Dead Utility	Live Utility	Dead MRS	Live MRS
153.75	170	0.683333	0.425

slope
0.3500

Price of risk $=(r_m - r_f)/\sigma$

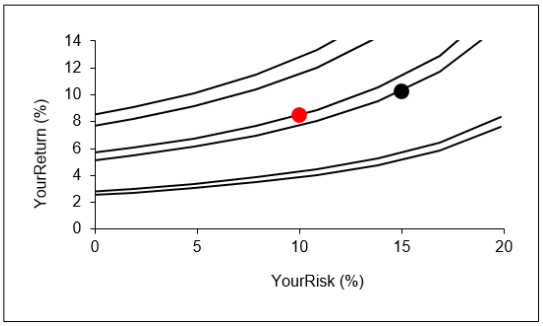


Figure 6.2: The investor's indifference map, $a = b = 1$.
Source: RiskReturn.xls!Preferences

The red dot is *live* in the sense that it depends on the value of B10. The chart displays the indifference curve that goes through the mix value in B10, along with an indifference curve and another below it.

A mix of 50% risky is better than 75% for this investor because utility went up. The red dot is on a higher indifference curve. Notice also that the MRS fell, getting closer to the slope of the budget line. That means the investor is getting closer to the optimal solution.

STEP Change B12 to 90.

Now the reverse is true. The red dot is on a lower indifference curve and the MRS is farther away from the slope.

STEP Change the exponent on *YourReturn* in B19 to 4 and click the Set button.

The indifference curves are now much flatter. What does this mean?

STEP Change B12 to 50 and 90.

We are getting different results than before? What is going on?

If $b > a$, the investor cares more about return than risk. The flat indifference curves (with low MRS) mean that they are willing to accept a lot of risk for a little more return. These preferences mean that this investor will find an optimal solution with a high risk, high return combination.

STEP Change B19 to 0.4 and click the Set button. Explore the satisfaction produced by mixes of 50% and 90%. What do you learn?

With a low b (lower than a), this investor is more concerned with risk. They are conservative and their optimal solution will lie on a low mix value. In fact, these preferences produce a corner solution, with the investor putting all \$10,000 into the risk-free asset.

Preferences are not right or wrong. If you are young and saving for retirement, it makes sense that $a < b$, but even then, if a person does not like risk, that is not a defect. An aggressive investor is not in any sense better than a conservative investor. Some people like risk and others do not in the same way that some people like broccoli or the color blue and others do not.

Preferences are not set in stone. They can be affected by the environment. A short time horizon, such as needing funds for college in a year, will rotate the indifference map, reflecting an investor who is more conservative. Likewise, retired people, typically, become more conservative and less willing to accept risk.

With the constraint and preferences modeled, we are ready to find the optimal solution.

STEP Proceed to the *OptimalChoice* sheet to see the numerical method in action.

The *OptimalChoice* sheet opens with an inefficient solution. The MRS is greater than the slope of the budget line so the indifference curve cuts the line. The agent should move down the line, accepting less return for less risk. This increases satisfaction. But how far down to travel?

STEP Run Solver to find the answer to this question.

At the optimal solution, the MRS equals the slope of the budget line and the agent is on the highest attainable indifference curve.

For this agent (with these attitudes toward risk and return) and the given market trade-off between risk and return (captured by the equation of the budget constraint), the optimal solution is found with a mix of about 39% of funds invested in the risky asset. Thus, the optimal risk to accept is $7\frac{6}{7}$ and the optimal return is $7\frac{3}{4}$.

Via analytical methods, we can use this Lagrangean to find optimal *YourRisk* (x_1) and *YourReturn* (x_2).

$$\max_{x_1,x_2,\lambda} L = (30 - x_1)x_2 + \lambda\left(x_2 - \left(\frac{r_m - r_f}{\sigma}\right)x_1 - r_f\right)$$

Try doing this problem and if you get stuck, the solution for a similar problem in the *Q&A* sheet is in the *Answers* folder.

Comparative Statics

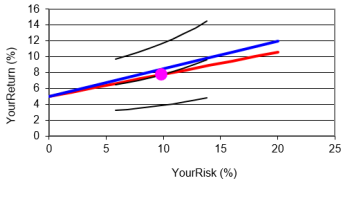
As usual, there are a number of comparative statics exercises to consider and they can be done via numerical or analytical methods. Let's explore the effect of an increase in sigma, the amount of risk the market forces you to bear in return for better performance.

STEP In the *OptimalChoice* sheet, increase σ from 20 to 25. What happens?

Figure 6.3 and your screen show a new, red budget line that has rotated clockwise and down.

Exogenous Variables		
Risk-free Return	5	r_f
Average Risky Return	12	r_m
Market Risk	25	σ
a	1	exponent on Risk
b	1	exponent on Return
Endogenous Variables		
YourMix	39.2857136	% of risky asset
YourRisk	9.8214284	
YourReturn	7.74999995	
Goal		
Utility	156.383929	

Risk	Return
0	5
2	5.56
4	6.12
6	6.68
8	7.24
10	7.8
12	8.36
14	8.92
16	9.48
18	10.04
20	10.6



slope	MRS
0.2800	0.3841

Figure 6.3: Increasing sigma, Solver yet to be run.
Source: RiskReturn.xls!OptimalChoice

The flatter slope is bad for the investor because consumption possibilities have been reduced. The market says that for a given amount of return, you must accept more risk. How will the investor respond to this shock?

STEP Run Solver to find out.

You will see that the agent chooses less risk and less return. What elasticity is under consideration here? There are several. There is the *sigma* elasticity of *YourRisk*, the *sigma* elasticity of *YourReturn*, and the *sigma* elasticity of *YourMix*.

Of course, these elasticities can also be computed at a point, using the derivative. One of the exercises asks you to do exactly that.

STEP Try your hand at computing the *sigma* elasticity of *YourRisk* from $\sigma = 20\%$ to 25%. Check your answer in the *CSsigma* sheet.

Of course, these elasticities can also be computed at a point, using the derivative. One of the exercises asks you to do exactly that.

Because the change in sigma is a change in the slope of the budget line, we can use the Slutsky decomposition approach to break down the total effect into income and substitution effects. This work is left for you as an exercise.

Asset Allocation is an Optimization Problem

Optimal Portfolio Theory is yet another application of the Theory of Consumer Behavior. The twist here is that one of the choices, risk, is a bad. The agent cannot ignore risk. She is forced to accept more risk to secure greater return.

The core concepts of the Theory of Consumer Behavior remain easily visible: a budget constraint describing consumption possibilities, preferences translated into an indifference map, maximization of utility given a budget constraint, and MRS equals slope of budget line at the optimal solution.

Perhaps most importantly, once we cast the problem as a choice, how to allocate assets among stocks, bonds, and other financial instruments, we are firmly in the land of Economics. This particular optimization problem is different from previous applications in that individuals are keenly interested

in getting the optimal solution right. There is often a lot of money at stake and mistakes can prove costly (for example, with a retirement portfolio).

As economists, we remain interested in comparative statics. Changing preferences are an important shock variable in this application. We do not shake our heads at the conservative investor who finds an optimal solution (given conservative preferences) at a low risk, low return point.

Exercises

1. Use the equation that follows to solve for *YourRisk** (x_1) and *YourReturn** (x_2) in terms of the exogenous variables. Show your work.

$$\max_{x_1,x_2,\lambda} L = (30 - x_1)x_2 + \lambda \left(x_2 - \left(\frac{r_m - r_f}{\sigma} \right) x_1 - r_f \right)$$

2. Use your reduced form solution to find the *sigma* elasticity of *YourRisk* at $\sigma = 20\%$ (and the values of the other exogenous variables from the initial position of the OptimalChoice sheet—click the Reset button if needed). Show your work.

3. Use Word's Drawing Tools to draw a well-labeled graph that depicts the total, income, and substitution effects for *YourRisk*. Make the substitution effect greater than an opposing income effect.

4. Compute the total, income, and substitution effects for *YourRisk* for the change in sigma from 20% to 25%. Show your work and describe your procedure.

References

The epigraph is from page 184 (9th edition) of a classic, excellent book on personal finance and the stock market. *A Random Walk Down Wall Street* by Burton Malkiel was originally published in 1973 by W. W. Norton & Company and the 12th edition came out in 2020. This is not one of those silly books with a scheme to beat the market. Malkiel is sober and reliable. On page 26, he says,

> Let me make it quite clear that this is not a book for speculators; I am not going to promise you overnight riches. I am not promising you stock-market miracles. Indeed, a subtitle for this book might well have been *The Get Rich Slowly but Surely Book*.

For a much deeper analysis of finance with an Excel-based presentation style, see *Principles of Finance with Excel* by Simon Benninga (New York: Oxford University Press, 2017. 3rd edition).

> Minivans have the lowest fraction of driver fatalities that are men under 26 years old (4 percent); sports cars have the highest (39 percent). So we suspect that differences in the behavior of their drivers account in large measure for why these two classes of vehicles pose such different risks to the people who operate them.
>
> Thomas P. Wenzel and Marc Ross

6.2 Automobile Safety Regulation

Cars are much, much safer today than in the past. Everyone knows that seat belts, airbags, and anti-lock brakes have made cars safer. The future holds great promise: guidance and avoidance systems, fly-by-wire technology that will eliminate steering columns, and much more; culminating in self-driving vehicles that communicate with each other.

But cars remain dangerous, both to vehicle occupants and others, such as cyclists and pedestrians. The United States uses the Fatal Accident Reporting System (FARS) to gather information about every motor vehicle crash in which someone dies. Such an event requires sending detailed information to FARS. Police record many variables, including time, weather conditions, demographic data, and whether drugs or alcohol were involved.

STEP To see the data, open the Excel workbook *SafetyRegulation.xls* and read the *Intro* sheet, then go to the *Data* sheet.

You can see that 36,650 people died in 2018 in a traffic accident. About half of the fatalities were drivers, almost 5,000 were motorcyclists, and 7,354 were non-motorists.

While FARS has data on the total number of deaths back to 1994 (36,254), simply comparing total fatalities over time is not a good way to measure driving safety. Under *Other National Statistics*, the data show that, year after year, there are many more people driving cars many more miles. So, we need to adjust the total number of fatalities to account for these increases.

We need a *fatality rate*, not the total number of fatalities. By dividing total deaths by the number of miles traveled, we get a measure of fatalities per mile traveled. This results in a tiny number so, to make it easier to read, the fatality rate is reported per 100 million miles traveled.

Adjusting with miles traveled is not the only way to create a fatality rate. The *Data* sheet shows rates based on population, registered vehicles, and licensed drivers. They all tell the same story.

Figure 6.4 shows the United States traffic fatality rate. The number of fatalities per 100 million miles traveled has fallen from 1.73 in 1994 to 1.17 in 2017, which is about a 30% decrease during this time period. That is welcome news.

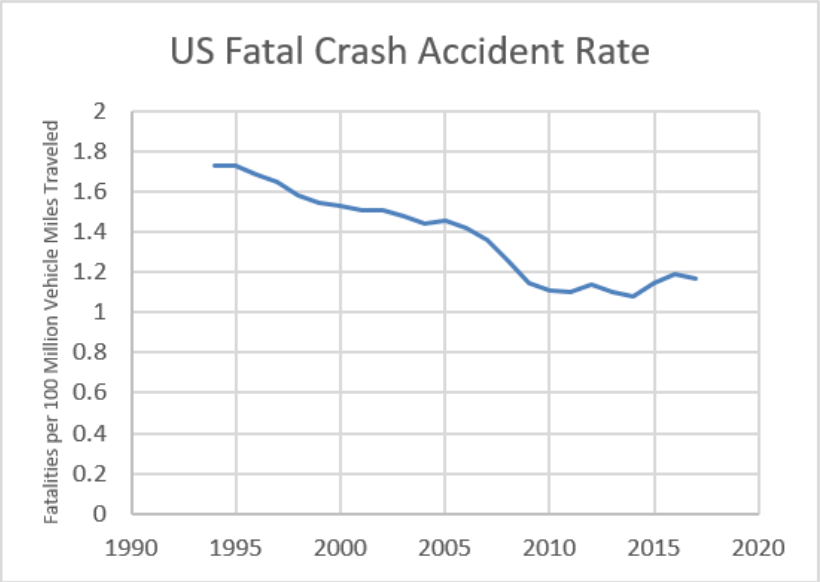


Figure 6.4: Traffic fatalities per 100 million miles traveled.
Source: SafetyRegulation.xls!Data
www-fars.nhtsa.dot.gov/

Less encouraging in Figure 6.4 is the leveling off since 2009 and the increase from 2014 to 2016. Distracted driving because of phone use and texting are suspected contributors.

The data in FARS only track fatalities and, thus, say nothing about nonfatal accidents. It turns out we are doing better here also—injury rates and severity of injury have also declined.

So, all is well? Actually, not exactly.

Although it may seem greedy, fatalities and injuries should have fallen by a lot more. We are doing better because fatal accident and injury rates have

fallen, but we should be doing much, much better. After all, the car you drive today is much, much safer than a car from 20 or 30 years ago. If the vehicle you drive today is much safer than vehicles from 20 or 30 years ago, then fatal accident and injury rates should have fallen more to reflect these improvements. So, what is going on?

Economics can help answer this question. We will apply the remarkably flexible Theory of Consumer Behavior to driving a car. Any problem that can be framed as a choice given a set of exogenous variables can be analyzed via the economic approach. There are certainly choices to be made while driving: what route to take, how fast to drive, and what car to drive are three of many choices drivers make. We will focus on a subset of choices that involve how carefully to drive.

Theoretical Intuition

The key article that spawned a great deal of further work in this area was written in 1975 by University of Chicago economist Sam Peltzman. The abstract for "The Effects of Automobile Safety Regulation" (p. 677) says,

> Technological studies imply that annual highway deaths would be 20 percent greater without legally mandated installation of various safety devices on automobiles. However, this literature ignores offsetting effects of nonregulatory demand for safety and driver response to the devices. This article indicates that these offsets are virtually complete, so that regulation has not decreased highway deaths. Time-series (but not cross-section) data imply some saving of auto occupants' lives at the expense of more pedestrian deaths and more nonfatal accidents, a pattern consistent with optimal driver response to regulation.

This requires some translation. By technological studies, Peltzman is referring to estimates by engineers that are based on extrapolation. Cars with seat belts, airbags, anti-lock brakes, and so on are assumed to be driven in exactly the same way as cars without these safety features. This will give maximum bang for our safety buck.

Economics, however, tells us that we won't get this maximum return on improved safety features because there is a driver response to being in a safer car. By offsetting effects, Peltzman means that the gains from the safety devices are countered, offset, by more aggressive driving.

Peltzman's key insight, which separates an economist from the way an engineer considers the problem, is to incorporate driver response. He says on page 681:

> The typical driver may thus be thought of as facing a choice, not unlike that between leisure and money income, involving the probability of death from accident and what for convenience I will call "driving intensity." More speed, thrills, etc., can be obtained only by forgoing some safety.

This claim sounds rather outrageous at first. Do I suddenly turn into an Indy 500 race car driver upon hearing that my car has airbags? No, but consider some practical examples in your own life:

- Do you drive differently in the rain or snow than on a clear day?
- Do speed bumps, if you can't swerve around them, lead you to reduce your speed?
- Would you drive faster on a road in Montana with no cars for miles around versus on the Dan Ryan Expressway in Chicago? In which case, Montana or Chicago (presuming you are actually moving on the Dan Ryan), would you pay more attention to the road and your driving?
- If your car had some magic repulsion system that prevented you from hitting another car (we almost have this), would you drive faster and more aggressively?

Economists believe that agents change their behavior to find a new optimal solution when conditions change. In fact, many believe this is the hallmark of economics as a discipline. Many non-economists either do not believe this or are not aware of how this affects us in many different ways.

If you do not believe that safer cars lead to more aggressive driving, consider the converse: Do more dangerous cars lead to more careful driving? Here is how Steven Landsburg puts it:

> If the seat belts were removed from your car, wouldn't you be more cautious in driving? Carrying this observation to the extreme, Armen Alchian of the University of California at Los Angeles has suggested a way to bring about a major reduction in the accident rate: Require every car to have a spear mounted on the steering wheel, pointing directly at the driver's heart. Alchian confidently predicts that we would see a lot less tailgating. (Landsburg, p. 5)

The idea at work here is only obvious once you are made aware of it. Consider the tax on cars over $30,000 passed by Congress in 1990. By adding a 10% tax to such luxury cars, staffers computed that the government would earn 10% of the sales revenue (price x quantity) generated by the number of luxury cars sold the year before the tax was imposed. They were sadly mistaken. Why?

People bought fewer luxury cars! This is a response to a changed environment. You cannot take for granted that everyone will keep doing the same thing when there is a shock.

This idea has far-reaching application. Consider, for example, its relevance to the field of macroeconomics. Robert Lucas won the Nobel Prize in Economics in 1995. His citation reads, "for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy." (See www.nobelprize.org/prizes/economic-sciences/1995/press-release/)

What exactly did Lucas do to win the Nobel? One key contribution was pointing out that if policy makers fail to take into account how people will respond to a proposed new policy, then the projections of what will happen will be wrong. This is called the *Lucas Critique*.

The Lucas Critique is exactly what is happening in the case of safety features on cars. Economists argue that you should not assume that drivers are going to continue to behave in exactly the same way before and after the advent of automobile safety improvements.

What we need is a model of how drivers decide how to drive. The Theory of Consumer Behavior gives us that model. You know what will happen next: we will figure out the constraint. And after that? Preferences. That will be followed by the initial solution and, then, comparative statics. We will find the effect of safer cars on accident risk. This is the economic approach.

The Initial Solution

The driver chooses how *intensively* to drive, which means how aggressively to drive. Faster starts, not coming to a complete stop, changing lanes, and passing slower cars are all more intensive types of driving, as are searching for a song or talking on your phone while driving. More intensive driving saves time and it is more fun. Driving intensity is a good and more is better.

Unfortunately, it isn't free. As you drive more intensively, your chances of having an accident rise. No one wants to crash, damaging property and injuring themselves or others. Your accident risk, the probability that you have an accident, is a function of how you drive.

The driver chooses a combination of two variables, *Driving Intensity* and *Accident Risk*, that maximize utility, subject to the constraint.

The equation of the constraint ties the two choice variables together in a simple way.

$$DrivingIntensity = SafetyFeatures * AccidentRisk$$

Safety Features represents the exogenous variable, safety technology, and provides a relative price at which the driver can trade risk for intensity.

On the Initial line in Figure 6.5, the driver is forced to accept a great deal of additional *Accident Risk* for a little more *Driving Intensity* because the line is so flat.

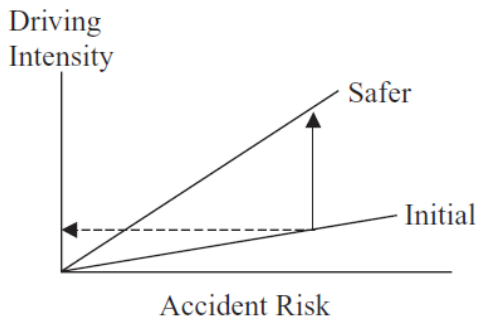


Figure 6.5: The driver's constraint.

When cars get safer, the constraint line gets steeper, rotating counterclockwise from the origin, as shown in Figure 6.5. There are two ways to understand the improvement made available by better safety technology. The horizontal, dashed arrow shows that you can get the same *Driving Intensity* at a much lower *Accident Risk*. You can also read the graph vertically. For a given *Accident Risk*, a safer car gives you a lot more *Driving Intensity* (follow the vertical, solid arrow).

Figure 6.5 shows that safer technology can be interpreted as a decrease in the price of *Driving Intensity*. It affects the graph just like a decrease in p_2 in the Standard Model.

The constraint is only half of the story. We need preferences to find out how a driver will decide to maximize satisfaction.

We use a Cobb-Douglas functional form to model the driver's preferences for *Accident Risk* (x_1) and *Driving Intensity* (x_2), subtracting *Accident Risk* from a constant so that increases in x_1 lead to less utility.

$$U(x_1, x_2) = (1 - x_1)^c x_2^d$$

Risk is measured between zero and 100 percent so $0 \leq x_1 \leq 1$. As x_1 increases in this interval, utility falls. The indifference curves will be upward sloping because x_1, *Accident Risk*, is a bad.

We can solve this model via numerical and analytical methods. We begin with Excel's Solver.

STEP Proceed to the *OptimalChoice* sheet.

The sheet shows the goal, endogenous variables, and exogenous variables. Initially, the driver is at 25%,0.25, which is a point on the budget line (because the constraint cell shows zero). We will use % notation for *Accident Risk* because it is a probability. The unrealistically high chances of an accident were chosen to maximize visibility on the graph. We use decimal points (such as 0.5) for the driving intensity variable, which we interpret as an index number on a scale from 0 to 1.

We know the opening point is feasible, but is it an optimal solution?

In previous Excel files, the graph is immediately displayed so you can instantly see if there is a tangency. The missing graph gives you a chance to exercise your analytical powers. Can you create a mental image of the chart even though it is not there? Remember, comparing the slope of the budget line to the MRS at any point tells us what is going on.

The slope is simply the *Safety Features* exogenous variable, which is +1. So now the graph looks like Figure 6.5 with a 45 degree line from the origin.

But what about the indifference curves? The MRS is minus the ratio of marginal utilities. With $c = d = 1$, we have

$$MRS = -\frac{\frac{dU}{dx_1}}{\frac{dU}{dx_2}} = -\frac{-x_2}{1-x_1} = \frac{x_2}{1-x_1}$$

We evaluate this expression at the chosen point, 25%, 0.25, and get

$$\frac{x_2}{1-x_1} = \frac{[0.25]}{1-[25\%]} = \frac{1}{3}$$

We immediately know the driver is not optimizing.

In addition, we know he can increase satisfaction by taking more risk and more intensity, traveling up the budget line because the indifference curve is flatter ($\frac{1}{3}$) than the budget line (+1) at the opening point of 25%,0.25.

Do you have a picture in your mind's eye of this situation? Think about it. Remember, the MRS is smaller than the slope so the indifference curve has to be flatter where it cuts the line.

STEP When you are ready (after you have formed the mental picture of the situation), click the Show Chart button to see what is going on at the 25%,0.25 point.

The canonical graph (with a bad) appears and the cells below the chart show the slope and MRS at the chosen point.

STEP Next, run Excel's Solver to find the optimal solution.

With $c = d = 1$ and a *Safety Features* value of 1, it is not surprising that the optimal solution is at 50%,0.50. Of course, at this point, the slope = MRS.

To implement the analytical approach, the Lagrangean looks like this:

$$\max_{x_1,x_2,\lambda} L = (1-x_1)x_2 + \lambda(x_2 - Sx_1)$$

An exercise asks you to find the reduced form solution.

Comparative Statics

Suppose we get safer cars so the terms of trade between *Driving Intensity* and *Accident Risk* improve. What happens to the optimal solution?

STEP Change cell B16 to 2.

How does the engineer view the problem? To her, the driver keeps acting the same way, driving just like before. There will be a great gain in safety with much lower risk of an accident. This is shown by the left-pointing arrow in Figure 6.6. Intensity stays the same and risk falls by a great deal.

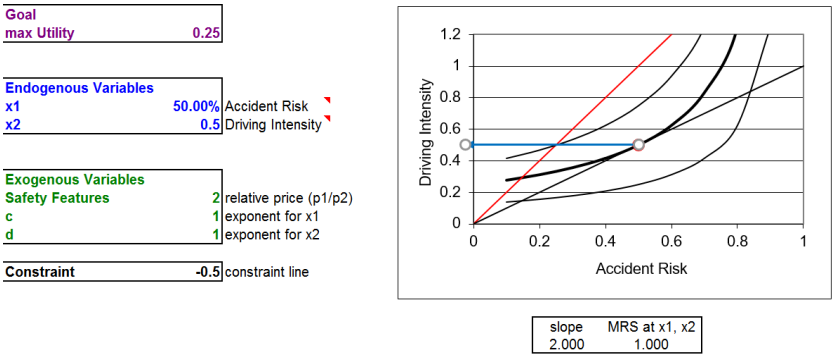


Figure 6.6: Improved safety features shock.
Source: SafetyRegulation.xls!OptimalChoice

For the engineer, because *Driving Intensity* remains constant, if it was 0.5, then improving Safety Features to 2 makes the accident risk fall to 25%. We simply travel horizontally along a given driving intensity to the new constraint.

The economist doesn't see it this way at all. She sees *Driving Intensity* as a choice variable and as the solution to an optimization problem. Change the parameters and you change the optimizing agent's behavior. It is clear from Figure 6.6 that the driver is not optimizing because the slope does not equal the MRS.

STEP With new safety technology rotating the constraint line, we must run Solver to find the new optimal solution.

The result is quite surprising. The *Accident Risk* has remained exactly the same! What is going on? In Peltzman's language, this is *completely offset-*

ting behavior. The optimal response to the safer car is to drive much more aggressively and this has completely offset the gain from the improved safety equipment.

How can this be? By decomposing the zero total effect on *Accident Risk* into its income and substitution effects, we can better understand this curious result.

Figure 6.7 shows what is happening. The improved safety features lower the price of driving intensity, so the driver buys more of it. On the y axis, the substitution and income effects work together to increase the driver's speed, lane changes, and other ways to drive more intensively. On the x axis, which measures risk taken while driving, the effects oppose each other, canceling each other out and leaving no gain in accident safety.

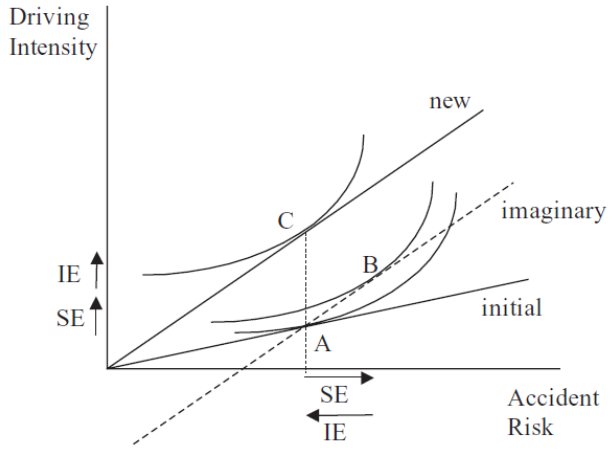


Figure 6.7: Income and substitution effects.

As driving intensity gets cheaper, the substitution effect (the move from A to B in Figure 6.7) leads the driver to choose more intensity and pay for it with more risk. The income effect leads the driver to buy yet more intensity and (because risk is a normal bad) less risk. The end result, for this utility function, is completely offsetting behavior.

Of course, this is not necessarily what we would see in the real world. We do not know how many drivers are represented by these preferences. The income effect for risk could outweigh the substitution effect, leaving point C to the left of A in Figure 6.7.

Theory alone cannot answer the question of what we will see in the real world. Empirical work in this area does confirm that offsetting behavior exists, but there is disagreement as to its extent.

An Economic Analysis of Driving

Choices abound when it comes to cars and driving. Should I take the highway or stay on a surface street? Change the oil now or wait a while longer? Pass this slow car or just take it easy and get there a few minutes later? Because there are choices, we can apply economics. This chapter focused on applying the Theory of Consumer Behavior to the choice of how intensively to drive. The agent is forced to trade off a bad (the risk of having an accident) for getting there faster and greater driving enjoyment.

Yes, teenagers make different choices than older drivers and everyone drives differently on a congested, icy road than on a sunny day with no traffic, but our comparative statics question focused on how improved automobile technology impacts the optimal way to drive.

Offsetting behavior is an application of the *Lucas Critique*: do not extrapolate. Instead, we should recognize that agents change their behavior when the environment changes. Theory cannot tell us how much offsetting behavior we will get. Only data and econometric analysis can tell us that.

Economists believe that we have not had as great a reduction in automobile fatalities and injuries as our much, much safer cars would enable because drivers have chosen to maximize satisfaction by trading some safety for driving intensity. Offsetting behavior explains why we aren't doing much, much better in traffic fatalities. But do not despair—we are maximizing satisfaction given our new technology.

Exercises

1. Use the equation that follows to solve for x_1^* and x_2^* in terms of S (safety features). Show your work.

$$\max_{x_1,x_2,\lambda} L = (1 - x_1)x_2 + \lambda(x_2 - Sx_1)$$

2. Use your reduced form solution to find the S elasticity of x_1^* at $S = 1$. Show your work.

3. If the utility function was such that *Driving Intensity* was a Giffen good, describe where point C would be located on Figure 6.7.

4. If the utility function was such that *Driving Intensity* was a Giffen good, would this raise or lower traffic fatalities? Explain.

References

The epigraph is from page 125 of Thomas P. Wenzel and Marc Ross, "Safer Vehicles for People and the Planet," *American Scientist*, Vol. 96, No. 2 (March–April, 2008), pp. 122–128, www.americanscientist.org/article/safer-vehicles-for-people-and-the-planet. They claim that the conventional wisdom that we need cars to be heavy to be safe is wrong. Heavier cars waste more fuel. How much more? "If a typical car could somehow drop 10% of its mass, its fuel economy would increase by anywhere from 3% to 8%. (The larger value applies if the size of the engine is also reduced to keep acceleration performance the same.)" (p. 124). The authors are not economists, but notice how they frame the result with percentage changes.

For an excellent review of empirical work on traffic safety, see *Traffic Safety* by Leonhard Evans, online at www.scienceservingsociety.com/.

This original idea is from Sam Peltzman, "The Effects of Automobile Safety Regulation," *The Journal of Political Economy*, Vol. 83, No. 4 (August, 1975), pp. 677–726, www.jstor.org/stable/1830396

For a simple (no math or graphs) explanation of the idea behind offsetting behavior, see "The Power of Incentives: How Seat Belts Kill," in Steven E. Landsburg, *The Armchair Economist* (New York: The Free Press, 1993).

Russell S. Sobel and Todd M. Nesbit point out that aggregated traffic fatality data is a poor way to test for a Peltzman effect. They find strong support for offsetting behavior from improved safety in professional auto racing in "Automobile Safety Regulation and the Incentive to Drive Recklessly: Evidence from NASCAR," *Southern Economic Journal*, Vol. 74, No. 1 (Jul., 2007), pp. 71-84, www.jstor.org/stable/20111953.

Tom Vanderbilt's *Traffic: Why We Drive the Way We Do (and What It Says About Us)* (New York: Alfred A. Knopf), 2008, touches on a variety of issues about cars and driving.

> In the past it was futile to double the wages of an agricultural worker in Silesia who mowed a certain tract of land on a contract, in the hope of inducing him to increase his exertion. He would simply have reduced by half the work expended.
>
> Max Weber

6.3 Labor Supply

We began the Theory of Consumer Behavior with the Standard Model where cash income (m) is given. The Endowment Model replaced given cash income with an initial endowment of two goods so the budget constraint became $p_1x_1 + p_2x_2 = p_1\omega_1 + p_2\omega_2$. We then focused on choices with bads—risky assets and accidents.

The application in this section is another example using a bad. As always, our eventual goal is comparative statics and elasticity. In this case, we will derive a supply curve for labor and concentrate on the wage elasticity of labor supply.

An innovation in this section is that the accompanying Excel workbook is less finished than usual. This enables you to practice implementing the model in Excel.

Setting Up the Problem

Instead of a mere consumer, the agent in this application is a consumer and worker.

Although an initial amount of non-labor income is assumed, total income can be increased by working. More hours at work means more income and greater consumption of goods and services. Consumption is good, but work is bad. Therein lies the problem.

Our consumer/worker can buy a single good, G, representing all consumer goods, at price p. Utility increases as she consumes more G.

The 24 hours in a day are divided into two types: work and leisure. The number of hours spent working in one day, H, is chosen by the agent. Earned

income is simply wH,where w is the wage rate in \$/hr. Although work generates income, our agent does not like to work. H is a bad in the utility function.

With this background, we are ready to organize the information into the three areas that comprise an optimization problem:

1. *Goal*: maximize utility, which is a function of goods consumed, G, and work, H, where H is a bad.

2. *Endogenous variables*: G, the amount of goods consumed, and H, the number of hours worked.

3. *Exogenous variables*: p, the price of the composite good; w, the wage rate; m, unearned, non-labor income; and parameters in the utility function.

The solution to this constrained optimization problem is depicted on a graph with a budget constraint and set of indifference curves. We consider each of these elements separately and then combine them.

Budget Constraint

The budget constraint is $m + wH \geq pG$. This equation says that total income is composed of unearned income (m) and earned income (wH). The inequality means that the consumer/worker cannot spend more on goods and services (pG) than the total income available.

Because no time elapses in this optimization problem, there is no reason for the agent to save (i.e., spend less than available) and we can make the constraint a strict equality, $m+wH = pG$. This allows us to use the Lagrangean method to solve the problem analytically.

In terms of a graph, it is easy to see that we can write the constraint as the equation of a line (with G on the y axis and H on the x axis) by dividing by p:

$$m + wH = pG$$

$$G = \frac{m}{p} + \frac{wH}{p}$$

Suppose $w = \$10/\text{hr}$, $m = \$40$, and $p = \$1/\text{unit}$. What would the constraint look like?

STEP Open the Excel workbook *LaborSupply.xls* and read the *Intro* sheet, then go to the *YourConstraint* sheet.

Your task is to fill in the G column and create a chart of the constraint. There are three steps.

STEP Click on B12 and enter a formula equal to the equation for G. The cells w, p, and m are not named so you should use absolute references ($ in front of column letters and row numbers) to enable easy filling down of the formula.

When finished, the formula in B12 should look like this: = $B$4/$B$3 + ($B$2/$B$3)*A12.

STEP The next step is to fill down the formula.

STEP Finally, create a chart with H and G as the source data. Be sure to label the axes of your chart.

The chart is based on hour intervals of work, but fractions of hours are possible. Thus, your chart should be a scatter chart with points connected by lines.

STEP Click the Reveal the Constraint button to see a finished version of the budget constraint.

The agent is free to choose any point on the constraint. The y intercept, 40 (equal to $\frac{m}{p}$), yields a small value of consumption, but the agent does not have to work. Movement up the line yields more G, but requires more H.

Points to the northwest of the line are unattainable. For example, the consumer/worker cannot afford the 10,200 combination. Working 10 hours adds $100 to the $40 non-labor income. This is not enough to buy $200 worth of goods.

What shock would enable our consumer/worker to buy the 10,200 combination?

There are three possibilities, one for each exogenous variable in the constraint.

STEP From the *Constraint* sheet (click the Reveal the Constraint button from the *YourConstraint* sheet if needed), change the wage to 16 in B2.

The constraint rotates up, counterclockwise, with a steeper slope and the same intercept, and the combination 10,200 is now feasible, which is easily confirmed by looking at the chart and row 22.

Changes in wages, ceteris paribus, rotate the constraint around the unearned income intercept.

STEP Return the wage to 10 in B2 (the constraint returns to its initial position when you hit the Enter key) and set p (in B3) to 0.7.

Instead of raising the wage, we have made the composite good cheaper. As with a wage increase, this is welcome news since there are more consumption possibilities.

The constraint appears to simply rotate up again, but look more carefully at the chart and underlying data. The slope is steeper, but the intercept has also changed. The \$40 of unearned income now buys a little more than 57 units of G. As before, it is easy to see that the combination 10,200 is now feasible.

Changes in price (p), ceteris paribus, rotate and shift the constraint.

STEP Return the price to 1 in B3 (the constraint returns to its initial position when you hit the Enter key) and set m (in B4) to 100.

This time, the constraint shifts vertically up. With \$100 of unearned and \$100 of earned income (from working 10 hours), the combination 10,200 is now feasible.

Changes in unearned income (m), ceteris paribus, shift the constraint.

Changes in w, p, and m affect the constraint. The initially unattainable combination of 10,200 can be made feasible by appropriately changing any of one of these three exogenous variables.

Preferences

In previous applications with bads, we used a Cobb-Douglas utility function and subtracted the bad from a constant. The same approach is adopted here.

Because the time period under consideration is a day, which has 24 hours, preferences can be represented by $U(H, G) = (24 - H)^c G^d$.

With $H = 0$, the agent gets the maximum value from the first term of the utility function, but remember that earned income will then be low and, therefore, G will be small.

Like the budget constraint, we need a visual representation of the utility function.

STEP Proceed to the *YourIndiffCurve* sheet to implement the utility function in Excel.

The sheet is unfinished. You need to fill in column B and draw a graph of the indifference curve. The indifference curve is initially based on $c = d = 1$ and a level of utility of 1960.

To fill in column B, you need to solve for the value of G that yields a utility level of 1960, given H. In other words, rewrite the utility function in terms of G, like this:

$$U(H, G) = (24 - H)^c G^d$$
$$G^d = \frac{U(H, G)}{(24 - H)^c}$$
$$G = \left[\frac{U(H, G)}{(24 - H)^c}\right]^{1/d}$$

STEP Use the expression above to enter a formula in B12 that computes the value of G necessary to produce a utility of 1960 when H = 2.

Your formula should look like this: = ($B$5/((24 - A12)^($B$3)))^(1/$B$4). It evaluates to a value of $G = 89.09$. This result makes sense because when $H = 2$, then $24 - 2 = 22$ and 22 x 89.09 (since $c = d = 1$) equals a utility value of 1960.

Notice again the use of absolute references.

STEP Fill down the formula and draw a chart with H and G as the source data. Label the axes.

Your chart is a graph of a single indifference curve. In fact, the entire quadrant is full of these upward sloping indifference curves and utility increases as you move in a northwesterly direction (taking less of the bad, H, and more of the good, G). This is the usual indifference map when we have a bad on the x axis.

Click the Reveal the Indiff Curve button to check your work or if you need help.

Finally, remember that changes in the exponents make the indifference curves flatter or steeper. A Q&A question explores this point.

Finding the Initial Optimal Solution

Having modeled the constraint and preferences, we are ready to find the initial solution.

The numerical approach is covered here; the analytical method is an exercise question.

STEP Proceed to the *YourOptimalChoice* sheet.

It is blank! You need to implement the problem in this sheet and run Solver to find the initial solution.

Organize the problem into the usual components: goal (maximize utility), endogenous variables (H and G), exogenous variables (w, p, m, c, and d), and a cell for the constraint.

The utility function is $U(H, G) = (24 - H)^c G^d$. The wage rate is \$10/hr, the price of G is \$1/unit, unearned income is \$40, and $c = d = 1$.

Click the Reveal the Optimal Choice button once you are finished or if you get stuck and need help.

Figure 6.8 shows the canonical graph of the initial optimal solution for the consumer/worker's constrained utility maximization problem.

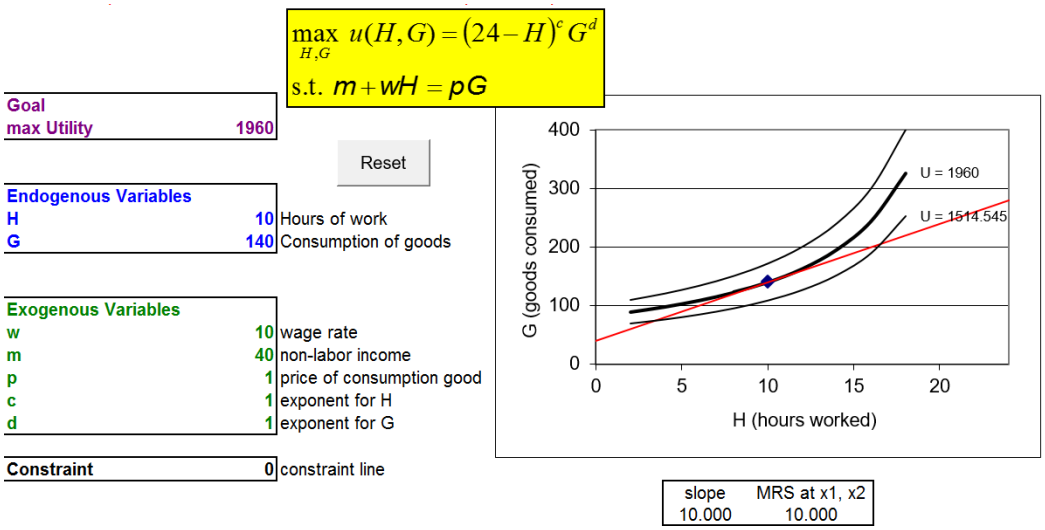


Figure 6.8: The initial solution.
Source: LaborSupply.xls!OptimalChoice

This consumer/worker maximizes utility by working 10 hours, thereby earning \$100 and then buying 140 units of G. There is no better solution. Traveling up or down the budget constraint is guaranteed to lower utility because the indifference curve is just touching the constraint at 10,140. The mathematical way of saying this is that the MRS $= \frac{w}{p}$ at 10,140.

Comparative Statics: Deriving Labor Supply

How does H^* respond as the wage rate changes, ceteris paribus? This comparative statics question yields the labor supply curve.

We concentrate on the numerical approach and leave the analytical method for an exercise question.

STEP Proceed to the *OptimalChoice* sheet (in the *YourOptimalChoice* sheet, click the Reveal the Optimal Choice button if needed). Use the Comparative Statics Wizard to pick a few points off of the labor supply curve. Make the size of the change in the wage rate 10 and apply the default five shocks.

Use the CSWiz data to compute the wage elasticity of hours worked from $w = \$10$ to \$20/hr. Create a graph the supply and inverse supply of labor curves.

STEP Proceed to the *CS1* sheet and scroll down (if needed) to check your work.

Notice the labor supply and inverse labor supply curves (scroll down if needed). The shape of the curve is intriguing. As wage rises, optimal H seems to level off—it continues to increase, but ever more slowly.

Notice also that the computed wage elasticity of labor supply from $w = 10$ to 20 in E14 is quite small at 0.1. This means that hours worked is unresponsive to changes in wages.

Labor supply has been extensively studied and extremely small elasticities with respect to wage are commonly found (see McClelland and Mok (2012) for a review of the literature). Income and substitution effects explain this result.

STEP Return to the *OptimalChoice* sheet and click the Reset button, then change the wage rate (in B16) from 10 to 20.

The budget constraint rotates up (counterclockwise) in the chart—a welcome change in consumption possibilities. The initial optimal solution, 10,140, is no longer optimal. The consumer/worker needs to re-optimize.

STEP Run Solver (with $w = 20$).

The new optimal solution is at $H = 11$. A 100% increase in the wage (from 10 to 20) has produced a total effect of a 1 hour, or 10%, increase in hours worked.

We can decompose this total effect into income and substitution effects by shifting down the budget line to cancel out the increased purchasing power of the wage increase. In other words, we need to draw in an imaginary, dashed line that goes through the initial solution, with a steeper slope caused by the higher wage.

We can use a modified version of the Income Adjuster Equation to determine the amount of income we need to take away. Recall that we determine how

much income to change via $\Delta m = x_1 \Delta p_1$. In the labor supply model, x_1 is obviously H, and the price is now the wage, but we also need a sign change. An increase in the wage increases consumption possibilities in the labor supply model so we need a minus sign to show that wage increases must be offset by income decreases. Below is our modified Income Adjuster Equation with values substituted in:

$$\Delta m = (\Delta H^*)(-\Delta w)$$

$$\Delta m = (10)(-10) = -100$$

This says that we must lower unearned income by \$100 to cancel out the increased purchasing power from the \$10/hr wage increase.

STEP Confirm that $w = 20$ (in B16) and change m to -60 (in B17).

Notice that the budget line goes through the initial combination, 10,140. The line is not dashed, but it should be. Remember that this budget line does not actually exist. No one is going to take \$100 from the agent. We are doing this to decompose the total effect of the wage increase into the income and substitution effects.

STEP Run Solver with $w = 20$ and $m = -60$.

$H^* = 13.5$ hours of work and Figure 6.9 shows the three effects.

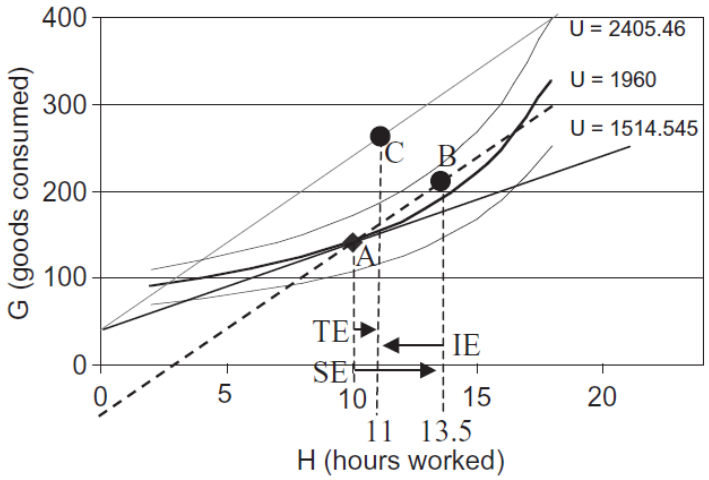


Figure 6.9: Total, income, and substitution effects.
Source: LaborSupply.xls!OptimalChoice

The substitution effect is +3.5, the movement from $H = 10$ (the initial optimal solution) to 13.5 (the optimal solution with the higher wage, but lower m). It is the horizontal movement from point A to B.

The income effect is -2.5, the movement from $H = 13.5$ (point B) to $H = 11$ (point C). The negative sign is important. It says that when income rises, the agent buys less of the bad.

The total effect is, of course, the observed movement from point A to point C, a 1-hour increase in hours worked. This is what would actually be observed as the wage rose from \$10/hr to \$20/hr.

Figure 6.9 makes clear why the response of hours worked to a wage increase is inelastic—the income and substitution effects are working against each other. The fact that the relative price of goods for an hour of work is cheaper drives the agent to work and consume more (this is the substitution effect, from A to B). But the increase in purchasing power encourages the agent to work less (from B to C, the income effect). The total effect on hours worked is small when the two effects are added together.

In fact, the income and substitution effects can explain an even more curious phenomenon that has been observed in the real world—hours worked actually falling as wage rises. Figure 6.10 shows the underlying graph and derived labor supply curve for an unknown utility function. Unlike the labor supply derived from the Cobb-Douglas utility function, which was always positively sloped, the labor supply curve in Figure 6.10 is said to be *backward bending*. At low wages, increases in wage lead to more hours worked (such as from point 1 to 2), but the supply curve becomes negatively sloped when wages rise from point 2 to 3.

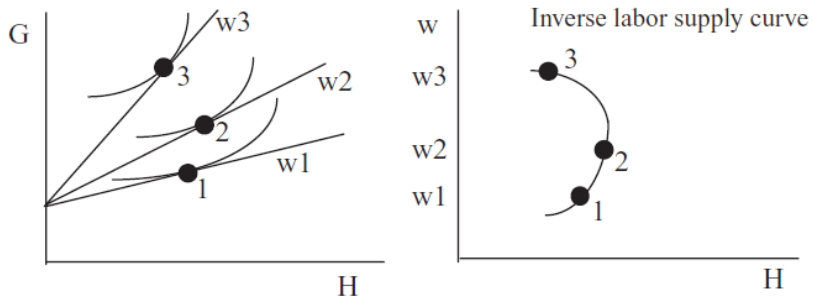


Figure 6.10: A backward bending supply curve.

We have already seen that the small wage elasticity from point 1 to 2 is caused by the income effect's working against the substitution effect. The same explanation underlies the negative response in hours worked as wages rise from point 2 to 3. In this case, not only does the income effect oppose the substitution effect, it actually swamps it.

Figure 6.11 shows what happens when we are on the backward bending portion of the labor supply curve. The substitution effect always induces more hours worked as wages rise. This is the movement from A to B. The income effect, however, counters some of this increase in hours worked. We can afford to work less (from B to C) because the wage is higher. When we are on the backward bending portion of the labor supply curve, the income effect actually overcomes the substitution effect so that the total effect (A to C) is a reduction in hours worked as the wage rises. In Figure 6.11, any point C to the left of A yields a point on the backward bending portion of the labor supply curve.

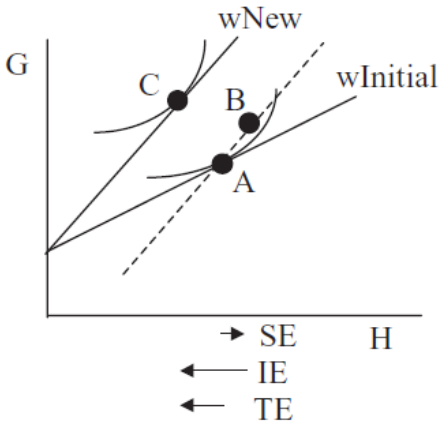


Figure 6.11: Income and substitution effects when H^* falls as w rises.

Wage rises and I work less sounds just about as weird as price rises and I buy more. Is this Giffen behavior?

No because the wage change is not an own price effect. Figure 6.12 shows p_1 and p_2 changes in the Standard Model where two goods are purchased given fixed income. On the left, the change in p_1 produces an own effect on x_1 and a cross effect on x_2.If x_1 rises as p_1 rises, then x_1 is Giffen. If x_2 rises as p_1 rises (notice the cross effect), however, that does not make x_2 a Giffen good. We use the cross effect to say that the goods are substitutes

(instead of complements). To determine whether x_2 is Giffen, we have to use the graph on the right of Figure 6.12. If x_2 rises as p_2 rises (notice the own effect), then x_2 is Giffen. In other words, we need an own price change to determine Giffenness.

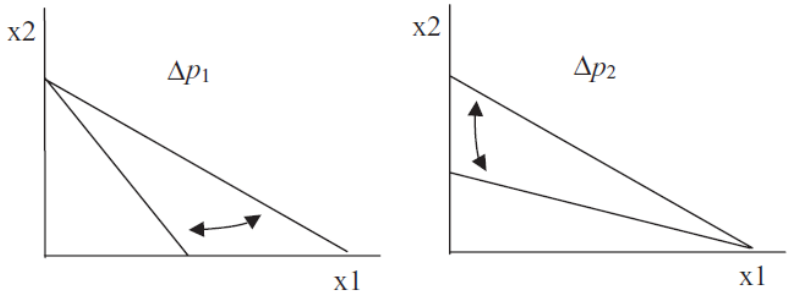


Figure 6.12: Understanding own and cross effects.

Figure 6.12 makes clear that a change in the wage in the labor supply optimization problem is like a change in the price of x_2 in the Standard Model. The wage change is like the graph on the right, with an upward sloping budget constraint. The rotation is around a fixed value—the x intercept in the Standard Model and unearned income in the labor supply model. Thus, the change in wage is an own price effect for G (on the y axis) and a cross price effect for H (on the x axis).

Because a change in the wage exerts a cross effect on hours worked, we cannot say anything about Giffenness for hours worked. We could, however, say that G was Giffen if it fell when wage rose. That would really be weird. Look at the figures of income and substitution effects in this chapter and you will never find a final point C that lies below an initial point A. In fact, leisure (work's counterpart) is usually treated as a normal good: higher income leads to more leisure (and less work).

Deriving the Labor Supply Curve

Labor Economics is a major field within Economics. As a course, it is usually offered as an upper-level elective, with Intermediate Microeconomics as a prerequisite. Labor supply and demand are fundamental concepts. The former is based on a model in which work is a bad (the opposite of leisure, which is a good) and a consumer/worker maximizes satisfaction subject to a budget constraint.

By changing the wage, ceteris paribus, we can derive a labor supply curve. Economists are well aware that labor supply is often quite insensitive to changes in wages. This is explained by the opposing substitution and income effects. The backward bending portion of the labor supply curve is observed when the income effect swamps the substitution effect. This is not Giffen behavior, however, because we are dealing with a cross (not own) price effect.

Exercises

1. Use the Lagrangean method to solve this consumer/worker's constrained optimization problem:

$$\max_{H,G} U = (24 - H)G$$
$$\text{s.t. } 40 + wH = G$$

   Show all of your work.

2. Do your results for H^* and G^* agree with the numerical approach in the text? Is this surprising?

3. Using the Comparative Statics Wizard, the wage elasticity of labor supply from \$10/hr to \$20/hr is 0.1. Use your reduced form solution for H^* to find the wage elasticity of labor supply at $w = \$10/\text{hr}$. Show your work.

4. Does your point wage elasticity from the previous question equal 0.1 (the wage elasticity based on a \$10 wage increase)? Why or why not?

5. Whether the labor supply curve is upward sloping or backward bending has nothing to do with the Giffenness of work. If labor supply is positively sloped, G and H are substitutes or complements, but which one? Draw a graph that helps you explain your answer.

References

The epigraph comes from page 355 of Max Weber's classic, *General Economic History*, originally published in German in 1923 and translated to English by Frank H. Knight in 1927. If you are unfamiliar with Weber (pronounced vay-ber), he was interested in the way capitalism changed people's minds and values, especially how it made people more rational and calculating.

With respect to labor supply, the consumer/worker's goals and attitudes are a critical issue. In this chapter, labor supply was derived as the solution to an optimization problem. The agent, however, might not be an optimizer, but a target earner, working only enough hours to make a certain amount of money. If wages double, hours worked are cut in half. If everyone was a target earner, the typical way to attract more workers—pay more—would not work.

Consider this abstract from Henry Farber's 2003 NBER working paper, "Is Tomorrow Another Day? The Labor Supply of New York Cab Drivers":

> I model the labor supply of taxi drivers as the result of optimization based on an inter-temporal utility function. Since income effects in response to temporary fluctuations in daily earnings opportunities are likely to be small, cumulative hours will be much more important than cumulative income in the decision to stop work on a given day. However, if these income effects are large due to very high discount and interest rates, then labor supply functions could be backward bending, and, in the extreme case where the wage elasticity of daily labor supply is minus one, drivers could be target earners. Indeed, Camerer, Babcock, Lowenstein, and Thaler (1997) and Chou (2000) find that the daily wage elasticity of labor supply of New York City cab drivers is substantially negative and conclude that it is likely that cab drivers are target earners. I conclude from my empirical analysis, based on new data, of the stopping behavior of New York City cab drivers that, when accounting for earnings opportunities in a reduced form with measures of clock hours, day of the week, weather, and geographic location, cumulative hours worked on the shift is a primary determinant of the likelihood of stopping work while cumulative income earned on the shift is weakly related, at best, to the likelihood of stopping work. This is consistent with there being inter-temporal substitution and inconsistent with the hypothesis that taxi drivers are target earners.

See http://www.nber.org/papers/w9706.

Google Scholar has tens of thousands of papers on Uber and how drivers decide how many hours to work.

Robert McClelland and Shannon Mok's 2012 working paper that summarizes the wage elasticity literature, "A Review of Recent Research on Labor Supply Elasticities," is freely available from the Congressional Budget Office at

www.cbo.gov/publication/43675. A remarkable finding is that men's much larger substitution effect than women's has all but disappeared so that men and women today respond similarly to wage shocks.

Chapter 7

Search Theory

Fixed Sample Search

Sequential Search

> Price dispersion is a manifestation—and, indeed, it is the measure—of ignorance in the market.
>
> George Stigler

7.1 Fixed Sample Search

The Theory of Consumer Behavior is based on the idea that buyers choose how much to buy based on preferences, income, and given prices. We know, however, that buyers do not face a single price—there is a distribution of prices and sellers change their prices frequently.

You would think consumers would be unable to choose in such an environment. After all, how can they know the budget constraint without prices? The answer is that they search or, in other words, they go shopping, and then use the lowest prices found to solve their constrained utility maximization problem.

Search Theory is an application of the economic approach to the problem of how long to shop in a world of many prices. Search is a productive activity because it enables one to find lower prices, but it is costly. One can search too little, ending up paying a high price, or search too much—spending hours to find a price that is a few pennies lower does not make much sense.

This chapter introduces the consumer's search optimization problem and is based on the idea that consumers decide in advance how many price quotes to obtain, according to an optimal search rule. This type of search procedure is known as a *fixed sample search*.

Describing the Search Optimization Problem

We assume that consumers do not know the prices charged by each firm. We simplify the problem by assuming that the product in different stores is identical (i.e., homogeneous) so the consumer just wants to buy at the lowest price. Unfortunately, finding that lowest price is costly so the buyer has to decide how long to search.

STEP Open the *FixedSampleSearch.xls* workbook and read the *Intro* sheet, then proceed to the *Setup* sheet.

The first task is to create the distribution of prices faced by the consumer. We assume that prices remain fixed during the search process.

STEP Click the Create the Population button.

You will be asked a series of questions that will establish the prices charged by all of the sellers. This is the population. The idea is that consumer will sample (draw) from the population. This is shopping.

STEP Hit OK when asked the number of stores selling the product to accept the default number of 1000 (no comma separator when entering numbers in Excel). Choose *Uniform* for the distribution and then press OK to accept 5 when prompted for the number of stores. Accept the default values of 0 and 1 for the minimum and maximum prices.

After you hit Enter, you will see a column of red numbers in column A that represent the prices charged by each of the 1,000 stores selling the product. The consumer knows that stores charge different prices, but cannot immediately see each individual store price. They cannot see the lowest and highest price stores in cells B2 and B3.

STEP Scroll down to see the prices charged at each store and confirm that the minimum price store, displayed in cell F2, is correct.

It is difficult to see by simply scrolling down and looking at the prices, but the uniform distribution you used means that prices are scattered equally from zero to one. The normal distribution, on the other hand, would concentrate prices near the average, with fewer low and high prices (like a bell-shaped curve). The log-normal is the most realistic of the three—prices have a long right-hand tail (with a few stores charging very high prices). The primary advantage of the uniform distribution is that it is the easiest to work with analytically.

Figure 7.1 shows a histogram of 1,000 prices from U[0,1]. This notation means that we include the endpoints so we have a uniform distribution with a zero minimum and a maximum of one (giving an average of 0.5 and an SD of 0.2887).

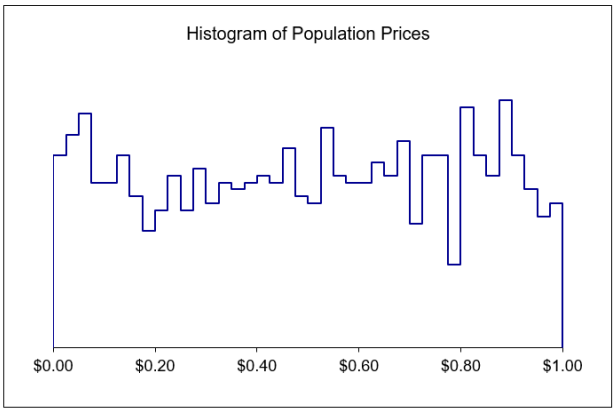


Figure 7.1: An example uniform distribution of prices.

The prices are not exactly evenly distributed on the interval from zero to one. They are drawn from a uniform distribution on the interval 0 to 1, but each realization of 1,000 prices deviates from a purely rectangular distribution due to randomness in sampling from the uniform distribution. The more stores you include in the population, the closer Figure 7.1 will get to a smooth, rectangular distribution. You can see a histogram of your population prices by scrolling over to column AA of the *Setup* sheet.

Consumers know the distribution of prices, but they do not know which firm is charging which price, so they cannot immediately go to the firm that has the lowest price. Instead, the fixed sample search model says that the consumer chooses a number of prices to sample (which you set as 5) and then chooses the lowest of the observed prices.

STEP Click the Draw a Sample One Price at a Time button. A price will appear in the sample column, and a pop-up box tells you where that price came from. Hit OK each time the display comes up. You will hit OK five times because you chose to sample from five stores.

The consumer chooses among the 1,000 stores randomly and ends up with five observed prices. Column L reports the sample average price, the SD of the sampled prices, and the minimum price in the sample (in cell L7). The consumer will purchase the product at the minimum price observed in the sample.

Why doesn't the consumer visit every store and then pick the lowest price? Because it is costly to obtain price information, as shown in cell L11. Each

shopping trip (to collect a price) costs 4 cents. To sample all 1,000 stores would cost the consumer an exorbitant $40. On average, the consumer would pay $0.54 (the average of the price distribution plus the cost of obtaining one price) by buying the product at the very first store visited. Clearly, it is better to buy immediately, $n = 1$, than to sample every store, $n = 1{,}000$, but what about other fixed sample sizes? How much will the consumer pay, on average, when sampling five stores?

STEP Hit the Draw a Full Sample button repeatedly to draw more samples of size five. Keep your eye on the total price paid in cell L22.

Every time you get a new sample, you get a new total price (composed of the minimum price in sample plus 20 cents). There is no doubt about it—the total price the consumer ends up paying is a random variable. This makes this problem difficult because we need to figure out what the consumer can expect to pay usually or typically. We want to know the average total price. The next section shows how.

Monte Carlo Simulation

The plan is to alter the spreadsheet so a new sample can be drawn simply by recalculating the sheet, which is done by hitting the F9 key. We can then install the Monte Carlo simulation add-in and use it to repeatedly draw new samples, tracking the lowest price in each sample.

STEP Select cell range J2:J6. You should have five cells highlighted. In the formula bar, enter the following formula:

=DRAWSAMPLEARRAY()

and then press *Ctrl + Shift + Enter* (hold down and continuing holding down the *Ctrl* key, then hold down and continue holding down the *Shift* key, and then hit the *Enter* key). Your sample of five prices will appear in the sample column.

After you select the cells, do not simply hit the *Enter* key. This will put the formula only in the first cell. You want the formula in all five cells that you selected. You have to press *Ctrl + Shift + Enter* simultaneously.

You have used an *array function* (built into the workbook) that spans the five cells you selected. You cannot individually edit the cells. If you mistakenly try to do so and get stuck, hit the *esc* (escape) key to return to the spreadsheet.

When using this array function, it may display *#VALUE*. Simply hit the F9 key when this happens to refresh the function. If that does not work, recreate the population.

When using the DRAWSAMPLEARRAY() function, you must be sure to set the number of draws in cell C15 to correspond to the number of cells selected and used by the function. If there is a discrepancy, a warning will be displayed.

STEP Hit F9 a few times and keep your eye on cells L7, the minimum price, and L22, the total price paid.

These cells update each time you hit F9. A new sample of five prices is drawn and the minimum price and total price paid are recalculated for the new sample.

The DRAWSAMPLEARRAY() function enables Excel to display the minimum (best) price random variable, but we need to figure out the average minimum price when five price quotes are obtained. This can be done by repeatedly resampling and tracking each outcome – this is called Monte Carlo simulation.

STEP Install the Monte Carlo simulation Excel add-in, *MCSim.xla*, available freely from www3.wabash.edu/econometrics and the *MicroExcel* archive (in the same folder as the Excel workbook for this section). Full documentation is available at this web site. This powerful add-in enables sophisticated simulations with the click of a button.

Remember that installing an add-in requires use of the Add-ins Manager. Do not simply open the *MCSim.xla* file.

Once installed, you can use the add-in to determine the average minimum price and total price paid for the product when five prices are sampled.

STEP Run the Monte Carlo simulation add-in on cells L7 and L22 with 10,000 repetitions.

Your MCSim add-in dialog box should look like Figure 7.2. Click the Proceed button to run the simulation.

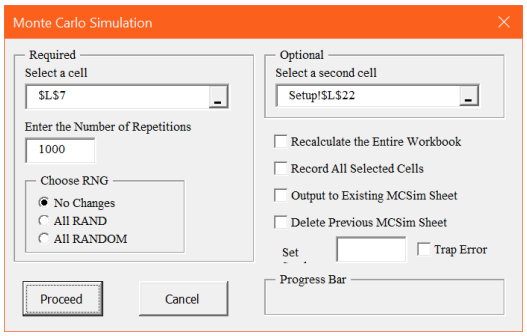


Figure 7.2: Configuring the MCSim dialog box.

Your simulation results will look something like Figure 7.3, but your results will be slightly different. The average of the minimum price distribution should be near 0.17 (1/6). Thus, the consumer will usually pay around \$0.37 (adding the 20 cents in search cost) for the product. The total price paid is a shifted version of the best price.

BestPrice		TotalPrice		Notes
Average	0.171	Average	0.371	
SD	0.1410	SD	0.1410	
Max	0.827	Max	1.027	
Min	0.001	Min	0.201	

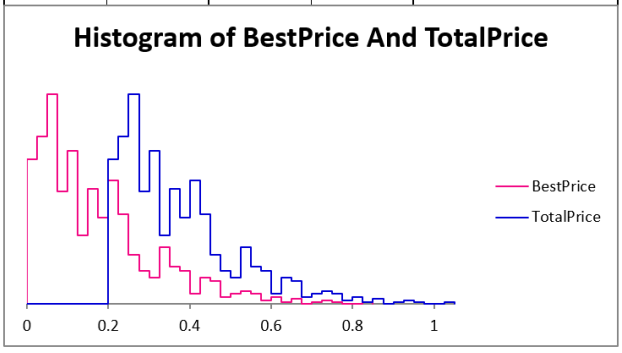


Figure 7.3: Monte Carlo simulation results with $n = 5$.
Source: FixedSampleSearch.xls!MCSim

So now we know that the consumer can expect to pay about \$0.37 when searching five stores. This is better than buying at the first store visited, which was \$0.54. Compared to the buying at the first store, the expected

marginal gain of shopping at five stores, in terms of a lower expected minimum price, is \$0.50 − \$0.17 = \$0.33. The additional cost of searching for five prices instead of one is \$0.16. The additional benefit is greater than the additional cost is another way to know that five stores is better than one store.

But we want to know more than just that searching five stores is better than buying at the first store; we want to find the best sample size—the one that gives the lowest total price paid.

STEP Hit the Clear the Sample button. Change the number of draws in cell C15 to 10. Select cell range J2:J11 and then type in the formula bar: =DRAWSAMPLEARRAY(). Then press the *Ctrl + Shift + Enter* combination to input the array formula. Your sample of 10 prices will appear in column J.

Hit F9 a few times and watch what happens to cell L7, the minimum price. It bounces, but with 10 prices instead of five, it bounces around a different, lower mean.

STEP To find the typical price the consumer can expect to pay, run a Monte Carlo simulation of the minimum and total price when 10 stores are visited.

Figure 7.4 shows the exact average best price and average total price as a function of the sample size for the U[0,1] price distribution. Your simulation results for the best price for $n = 10$ should be close to \$0.0909.

Sample Size	Average Best Price	Search Cost	Total Price Paid
1	\$ 0.5000	\$ 0.04	\$ 0.54
2	\$ 0.3333	\$ 0.08	\$ 0.41
3	\$ 0.2500	\$ 0.12	\$ 0.37
4	\$ 0.2000	\$ 0.16	\$ 0.36
5	\$ 0.1667	\$ 0.20	\$ 0.37
6	\$ 0.1429	\$ 0.24	\$ 0.38
7	\$ 0.1250	\$ 0.28	\$ 0.41
8	\$ 0.1111	\$ 0.32	\$ 0.43
9	\$ 0.1000	\$ 0.36	\$ 0.46
10	\$ 0.0909	\$ 0.40	\$ 0.49

Figure 7.4: Optimal Search with a Uniform Distribution on the interval [0,1].
Source: FixedSampleSearch.xls!Summary

The typical \$0.0909 best price when 10 prices are obtained is lower than when we shopped at five stores, but notice that it isn't worth it. The cost of obtaining 10 prices (\$0.40) is so high that the total price paid is higher than getting just five prices. In fact, getting four prices is the optimal sample size.

Analytical Methods

The optimal search optimization problem can be solved via analytical methods. For the uniform price distribution on the interval from zero to one, the average minimum price in the consumers' hands after visiting n firms is

$$AverageP_{min} = \frac{1}{n+1}$$

The equation for the average minimum price shows that it is decreasing as n rises and it does so at a decreasing rate. In other words, there are diminishing returns to searching for low prices.

The consumer's optimization problem is to minimize the expected total cost of acquiring the product, where $P(n)$ represents the expected minimum price that we know is a function how many prices are collected:

$$\min_n TC = P(n)q + cn$$

We also know that for U[0,1], $P(n) = \frac{1}{n+1}$ so we have:

$$\min_n TC = \frac{1}{n+1}q + cn$$

To find optimal n, take the derivative with respect to n and set it equal to zero:

$$\frac{dTC}{dn} = -\frac{1}{(n+1)^2}q + c = 0$$

$$\frac{1}{(n+1)^2}q = c$$

This equimarginal condition says that the optimal sample size is found where marginal savings from additional search equals marginal cost. As long as the savings from searching an additional store exceeds the cost of collecting one more price, the consumer will continue to search. The marginal savings is just the drop in the expected price, times the number of units that the consumer wants to purchase.

From the equimarginal condition, we can solve for optimal n to get a reduced form solution.

$$\frac{1}{(n+1)^2}q = c \rightarrow q = c(n+1)^2 \rightarrow \sqrt{\frac{q}{c}} = n+1 \rightarrow \sqrt{\frac{q}{c}} - 1 = n^*$$

With $q = 1$ and $c = \$0.04$, we have the same solution we found earlier:

$$n^* = \sqrt{\frac{q}{c}} - 1 = \sqrt{\frac{1}{0.04}} - 1 = 4$$

Comparative Statics

The reduced form expression makes comparative statics analysis straightforward. It is obvious that higher c, search cost, leads to lower optimal sample size, as shown in Figure 7.5.

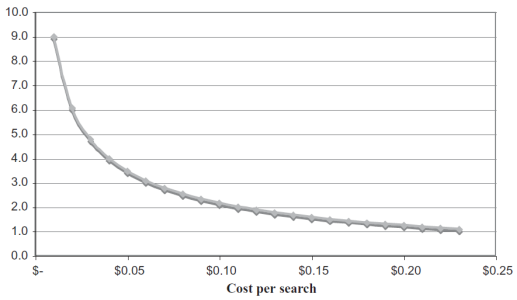


Figure 7.5: Optimal search with changing search cost for $q = 1$.

Search cost is not the same for each consumer. Time is an important element of search cost. Those with more valuable time and, therefore, higher search cost will optimize by obtaining fewer price quotes.

The availability of information is another component of search cost. Informational advertising is how firms let consumers know where they are and what prices are being charged. We can model this type of advertising as a decrease in search costs—today, all the consumer has to do is go online to see what prices are being offered. Search costs are still positive (consumers do not know, for example, whether all firms advertise or just some), but lower than without advertising. Consumers obtain the product for a lower total price when advertising lowers search costs.

If we allow for multiple purchases, that is, a value of $q > 1$, then the returns to search increase and, other things equal, the optimal number of searches

increases. The effect of increasing q on the relationship between the cost of search and the optimal number of searches is shown in Figure 7.6.

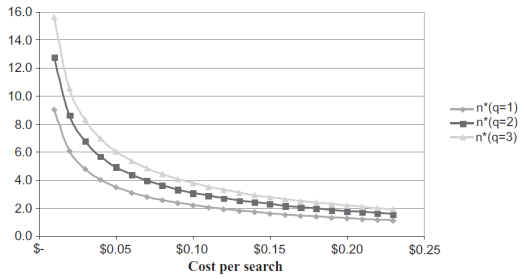


Figure 7.6: Optimal search with changing n and q.

For example, the driver of an 18-wheel truck that carries two 200-gallon diesel tanks is going to search more than someone looking to fill her car with gas. But this example leads to the next chapter, where we introduce a different search model.

Results of Fixed Sample Search

Incomplete price information leads to an entirely new optimization problem. Because consumers will not search every store, since that is too expensive, we see price dispersion. This is a major result of search theory and it deserves further explanation.

You would think that competition would tend to make prices of the same product equal. This is known as the *Law of One Price*. But this only applies to a world where consumers can costlessly gather prices.

In other words, the Law of One Price will fail to hold whenever it is costly to collect price data. This is true in the real world, where some consumers will end up paying higher prices than others because the minimum price in their particular information set is different than the minimum price in another consumer's set.

Because lower search costs induce more search, a reduction in search costs would have the effect of reducing (but not eliminating) price dispersion. Because optimizing consumers will choose not to canvass every store for prices as long as search is costly, price dispersion will exist. This is the key result of the fixed sample search model.

Economists have been interested in search theory for decades. The internet promised a big decrease in search cost and it may well have delivered that, but more recently, technology has really upended search theory. Today, your online search behavior is monitored and your clicks influence the prices you see.

The next level search models do not treat the population of prices as given and do not allow the consumer to randomly sample without changing the price distribution. Consumers still have an optimization problem to solve, but so do firms.

Exercises

Suppose the price distribution of 1,000 firms is uniform, with an average price of \$50 and an SD of \$28.87. Search cost, c, is \$1 per price.

1. On what interval (from the minimum to the maximum) are prices equally likely to fall?

2. Implement this problem in the *Setup* sheet and run a Monte Carlo simulation with a sample size of 20. Take a picture of your results (like Figure 7.3) and paste it in a Word document. What is good about obtaining 20 prices? What is bad?

3. Use the equation for the average minimum price as a function of n for this distribution, $AverageP_{min} = \frac{100}{n+1}$, to find the optimal sample size. Show your work.

4. Find the c elasticity of n at $q = c = 1$. Show your work.

References

The epigraph is from page 214 of George J. Stigler, "The Economics of Information," The *Journal of Political Economy*, Vol. 69, No. 3 (June, 1961), pp. 213–225, www.jstor.org/stable/1829263. This paper is recognized as the beginning of the economics of search.

Stigler was trying to explain price dispersion, but search theory has expanded far beyond this and is especially important in Labor economics. A consumer shopping for a low price product is the same as a worker looking for a high wage job or a firm seeking a high quality employee. See Richard Rogerson,

Robert Shimer, and Randall Wright, "Search-Theoretic Models of the Labor Market: A Survey," *Journal of Economic Literature*, Vol. 43, No. 4 (December, 2005), pp. 959–988, www.jstor.org/stable/4129380.

> Job offers are independent random selections from the distribution of wages. These offers occur periodically and are either accepted or rejected. Under these conditions it is easy to show that the optimal policy for the job searcher is to reject all offers below a single critical number and to accept any offer above this critical number.
>
> J. J. McCall

7.2 Sequential Search

We introduced Search Theory with a Fixed Sample Search Model. A consumer samples from the population of stores and gets a list of n prices for a product, then chooses the minimum price. The bigger n, the lower the minimum price in the list, but the price paid to obtain the price quotes increases as n rises. The consumer has to decide how many prices to obtain.

This section explores the properties of a different situation that is known as the Sequential Search Model. Unlike fixed sample search, where the consumer obtains a set of price quotes and then picks the lowest price, sequential search proceeds one at a time. The consumer samples from the population and gets a single price, then decides whether or not to accept it. If she rejects it, she cannot go back. As the epigraph shows, the sequential search model is easily applied to job offers, but it will be applied in this chapter to another common search problem—buying gas.

Setting Up the Model

Imagine you are driving down the road and you need fuel. As you drive, there are gas stations (say $N = 100$) to the left and right (taking a left does not bother you too much) and you can easily read the price per gallon as you drive up to each station. If you drive past a station, turning around is out of the question (there is traffic and you have a weird phobia about U-turns).

There is a lowest price station and the stations can be ranked from 1 (lowest, best price) to 100 (highest, worst price). You do not know the prices coming up because the stations are randomly distributed on the road. The lowest price station might be 18^{th} or 72^{nd} or even the very first one. Figure 7.7 sums it all up.

Figure 7.7: Deciding where to buy gas.

Suppose you focus on the following question: How do you maximize the chances of finding the cheapest station?

You might argue that you should drive by all of the stations, and then just pick the best one. This is a terrible idea because you cannot go back (remember, no U-turns). Once you pass a station, you cannot return to it. So, this strategy will only work if the cheapest station is the very last one. The chances of that are 1 in a 100.

A strategy for choosing a station goes like this: Pick some number $K < N$ where you reject (drive by) stations 1 to K, then choose the first station that has a price lower than the lowest of the K stations that you rejected.

Perhaps $K = 50$ is the right answer? That is, drive by stations 1 to 50, then look at the next (51^{st}) station and if it is better than the lowest of the 50 you drove by, pull in. If not, pass it up and consider the 52^{nd} station. If it is cheaper than the previous 51 (or 1 to 50 since we know the 51^{st} station isn't cheaper than the cheapest of the first 50), get gas there.

Continue this process until you get gas somewhere, pulling into the last (100^{th}) station if you get to it (it will have a sign that says, "Last chance gas station").

This strategy will fail if the lowest price is in the group of the K stations you drove by, so you might want to choose K to be small. But if you choose K too small, you will get only a few prices and the first station with a price lower than the lowest of the K stations is unlikely to give you the lowest price.

So, $K = 3$ is probably not going to work well because you probably won't get a super low price in a set of just three so you probably won't end up choosing the lowest price. For example, say the first three stations are ranked 41, 27,

and 90. Then as soon as you see a station better than 27, you will pull in there. That might be 1, but with 26 possibilities, that's not likely.

On the other hand, a high value of K, say 98, suffers from the fact that the lowest price station is probably in that group and you've already rejected it! Yes, this problem is certainly tricky.

The Sequential Search Model can be used for much more than buying gas—it has extremely wide applicability and, in math, it is known as *optimal stopping*. In hiring, it is called the *secretary problem*. A firm picks the first K applicants, interviews and rejects them, then picks the next applicant that is better than the best of the K applicants. It also applies to many other areas, including marriage—search online for Kepler optimal stopping to see how the famous astronomer chose his spouse.

STEP Open the Excel workbook *SequentialSearch.xls* and read the *Intro* sheet, then proceed to the *Setup* sheet.

Column A has the 100 stations ranked from 1 to 100. The lowest priced station is 1, and the highest priced station is 100.

STEP Click the Randomly Assign button. It shuffles the stations, randomly distributing them along the road you are traveling in column D.

Cell B7 reports where the lowest priced station (#1) is located. Columns C and D report the location of each station. Column D changes every time you click the Randomly Assign button because the stations are shuffled.

Cell F2 sets the value of K. This is the choice variable in this problem. Our goal is to determine the value of K that maximizes the probability that we get the lowest priced station.

On opening, $K = 10$. We pass up stations 1 to 10, then take the next station that is better than the best of the 10 stations we rejected.

STEP Click the How Did I Do? button. This reshuffles the stations and draws a border in column D for the cell at the K^{th} station.

Cell F5 reports the best of the K stations (that were rejected). Cell F7 displays the station you ended up at.

STEP Scroll down to see why you ended up at that station and read the text on the sheet.

Cell F7 always displays the first station that is better (lower) than the best of the K stations in cell F5.

STEP Repeatedly click the How Did I Do? button. After every click, see how you did. Is 10 a good choice for K?

The definition of a good choice in this case is one that has a high probability of giving us the cheapest station. Our goal is to maximize the chances of getting the cheapest station. We could have a different objective, for example, minimize the average price paid, but this would be a different optimization problem. For the classic version of the optimal stopping problem, we count success only when we find the cheapest station.

STEP Change K to 60 (in F2) and repeatedly click the How Did I Do? button. Is 60 better than 10?

This is difficult to answer with the *Setup* sheet. You would have to repeatedly hit the How Did I Do? button and keep track of the percentage of the time that you got the cheapest station. That would require a lot of patience and time tediously clicking and recording the outcome. Fortunately, there is a better way.

Solving the Problem via Monte Carlo Simulation

The *Setup* sheet is a good way to understand the problem, but it is not helpful for figuring out the optimal value of K. We need a way to quickly, repeatedly sample and record the result. That is what the *MCSim* sheet does.

STEP Proceed to the *MCSim* sheet and look it over.

With $N = 100$ (we can change this parameter later), we set the value of K (in cell D7) and run a Monte Carlo simulation to get the approximate chances of getting the best station (reported in cell H7).

Unlike the MCSim add-in used in the previous section, this Monte Carlo simulation is hard wired into this workbook. Thus, it is extremely fast.

STEP With $N = 100$ and $K = 10$, click the Run Monte Carlo Simulation button. The default number of repetitions is 50,000, which seems high, but a computer can do hundreds of thousands of repetitions in a matter of seconds.

Figure 7.8 shows results. Choosing $K = 10$ gives us the best station about 23.4% of the time. Your results will be slightly different.

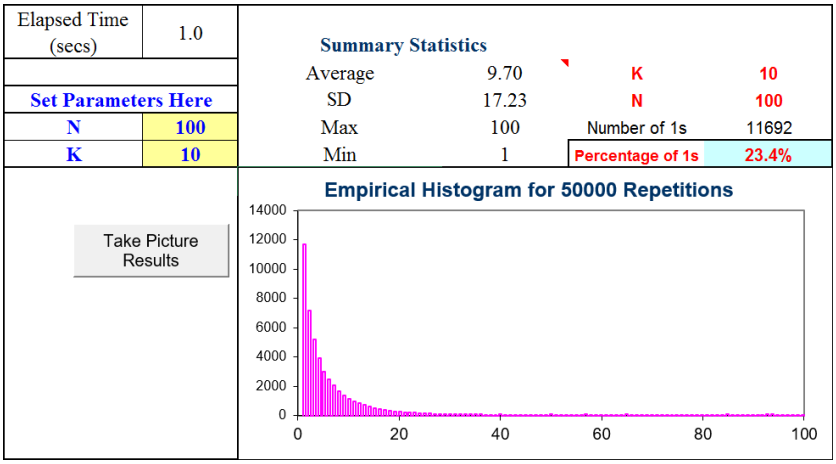


Figure 7.8: Monte Carlo simulation results.
Source: SequentialSearch.xls!MCSim

Notice that we are using Monte Carlo simulation to approximate the exact answer. Monte Carlo simulation cannot give us the exact answer. By increasing the number of repetitions, we improve the approximation, getting closer and closer, but we can never get the exact truth with simulation. The answer it gives depends on the actual outcomes in that particular run. The only way simulation would give the exact answer is if it was based on an infinite number of repetitions.

Can we do better than getting the best station about 23% of the time?

We can answer this question by exploring how the chances of getting the lowest price varies with K. By changing the value of K and running a Monte Carlo simulation, we can evaluate the performance of different values of K.

STEP Explore different values of K and fill in the table in cells J3:M10.

As soon as you do the first entry in the table, $K = 20$, you see that it beats $K = 10$.

STEP Use the data in the filled in table to create a chart of the chances of getting the lowest price station as a function of K. Use the button under the table to check your work.

What do you conclude from this analysis?

One problem with Monte Carlo simulation is the variability in the results. Each run gives different answers since each run is an approximation to the exact answer based on the outcomes realized. Thus, it seems pretty clear that the optimal value of K is between 30 and 40, but using simulation to find the exact answer is difficult.

Figure 7.9 displays results of series of Monte Carlo experiments. Notice that we doubled the number of repetitions to increase the resolution. The best value of K appears to be 36, but the noisiness in the simulation results makes it impossible to determine the answer.

Number of Repetitions	Value of K	Percentage of 1s	Average
100,000	30	35.84%	17.06
100,000	31	36.54%	17.41
100,000	32	36.59%	17.76
100,000	33	36.81%	18.24
100,000	34	36.60%	18.99
100,000	35	36.61%	19.30
100,000	36	36.93%	19.58
100,000	37	36.73%	20.21
100,000	38	36.62%	20.56
100,000	39	36.55%	21.38
100,000	40	36.52%	21.57

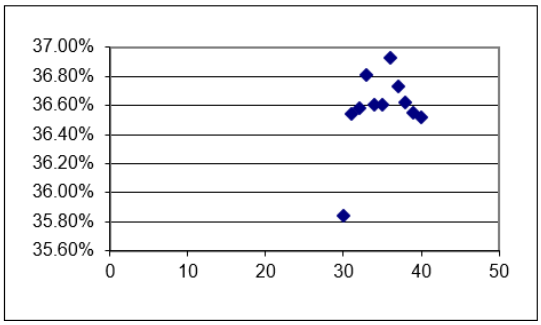


Figure 7.9: Zooming in on the value of optimal K.
Source: SequentialSearch.xls!Answers

With Monte Carlo simulation, we can continue to increase the number of repetitions to improve the approximation.

STEP Proceed to the *Answers* sheet to see more simulation results.

The *Answers* sheet shows that even 1,000,000 repetitions are not enough to definitively give us the correct answer. Simulation is having a difficult time distinguishing between a stopping K value of 36 or 37.

An Exact Solution

This problem can be solved analytically. The solution is implemented in Excel. For the details, see the Ferguson citation at the end of this chapter.

STEP Proceed to the *Analytical* sheet to see the exact probability of getting the cheapest station for a given K-sized sample from N stations from 5 to 100.

For example, cell G10 displays 32.74%. This means you have a 32.74% probability of getting the cheapest station out of 10 stations if you drive by the first six stations and then choose the next station that has a price lower than the cheapest of the K stations you drove by.

For $N = 10$, is $K = 6$ the best solution?

No. The probability of choosing the cheapest station rises if you choose $K = 5$. The 3 and 4 choices are close, but clearly, optimal $K = 3$ (with a 39.87% likelihood of getting the cheapest station) is the best choice.

In the example we have been working on, we had $N = 100$. Monte Carlo simulations showed optimal K around 36 or 37, but we were having trouble locating the exact right answer.

STEP Scroll down to see the probabilities for $N = 100$. Click on cells AL100 and AM100 to see the exact values. The display has been rounded to two decimal (percentage) places, but the computation is precise to more decimal places.

$K^* = 37$ just barely beats out $K = 36$. The fact that they almost give the exact same chances of getting the lowest price explains why we were having so much trouble zooming in on the right answer with Monte Carlo simulation.

It can be shown (see the Ferguson source in the References section) that optimal K is $\frac{N}{e}$, giving a probability of finding the cheapest station of $\frac{1}{e}$. For $N = 100$, $\frac{N}{e} \approx 36.7879$.

If K was a continuous endogenous variable, $\frac{N}{e}$ would be the optimal solution. But it is not, so the exact, correct answer is to pass on the first 37 stations and then take the first one with a lower price than the lowest price of stations 1 to 37.

It is a mystery why the transcendental number e, the base of natural logarithms, plays a role in the solution.

Figure 7.10 shows that as N rises, so does optimal K. What elasticity is under consideration here?

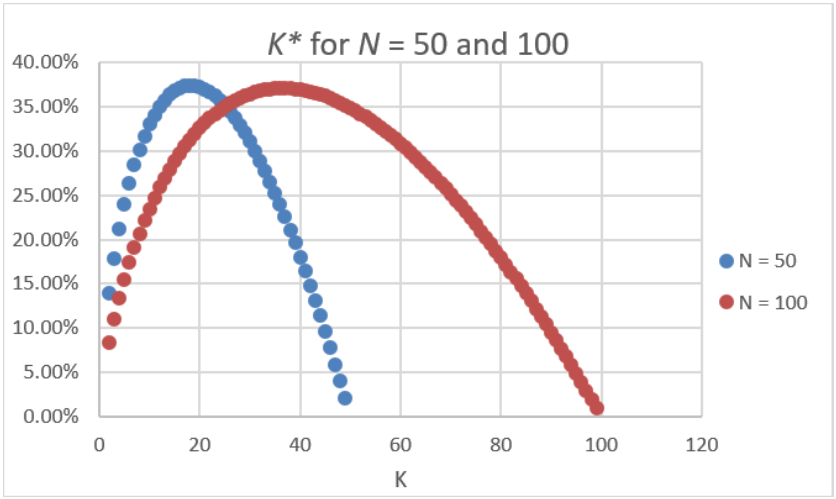


Figure 7.10: Exact probabilities of finding the cheapest station.
Source: SequentialSearch.xls!Analytical

The answer is the N elasticity of K. From $N = 50$ to 100 is a 100% increase. What happens to optimal K? It goes from 18 to 37, so a little more than a 100%. The elasticity is slightly over one. If you use the continuous version of K, then K exactly also doubles and the N elasticity of K is exactly one.

Sequential Search Lessons

Unlike the Fixed Sample Search Model (where you obtain a set of prices and choose the best one), the Sequential Search Model says that you draw sample observations one after the other. This could apply to a decision to choose a gas station. As you drive down the road, you decide whether to turn in and get gas at Station X or pass up that station and proceed to Station Y.

Faced with price dispersion, a driver deciding where to get gas can be modeled as solving a Sequential Search Model. Although there can be other objectives (such as getting lowest average price), the goal could be to maximize the chances of getting the lowest price. We found that as N rises, so

does optimal K. The more stations, the more driving you should do before picking a station.

Like the Fixed Sample Search Model, the Sequential Search Model does not have any interaction between firms and consumers. Price dispersion is given and the model is used to analyze how consumers react in the given environment.

In the pre-internet and smartphone days, deciding where to get gas was quite the challenge. A driver passing signs with prices (like Figure 7.7) was a pretty accurate representation of the environment. There was no Google maps or apps displaying prices all around you. Notice, however, that the Law of One Price does not yet apply to gas prices.

Ferguson points out that our Sequential Search Model (which mathematicians call the secretary problem) is part of a class of finite-horizon problems. "There is a large literature on this problem, and one book, *Problems of Best Selection* (in Russian) by Berezovskiy and Gnedin (1984) devoted solely to it" (Ferguson, Chapter 2).

Fixed Sample and Sequential Search Models are merely the tip of the iceberg. There is a vast literature and many applications in the economics of search, economics of information, and economics of uncertainty.

Exercises

1. Use the results in the *Analytical* sheet to compute the N elasticity of K^* from $N = 10$ to 11. Show your work.

2. Use the results in the *Analytical* sheet to draw a chart of K^* as a function of N. Copy and paste your graph in a Word document.

3. Run a Monte Carlo simulation that supports one of the N-K^* combinations in the *Analytical* sheet. Take a picture of your simulation results and paste it in a Word document.

4. Explain why the Monte Carlo simulation was unable to exactly replicate the percentage of times the lowest priced station was found.

References

The epigraph is from pages 115 and 116 of J. J. McCall, "Economics of Information and Job Search," *The Quarterly Journal of Economics*, Vol. 84, No. 1 (February, 1970), pp. 113–126, www.jstor.org/stable/1879403. This paper shows that sequential search (with recall) dominates fixed sample search. For more on this point, see Robert M. Feinberg and William R. Johnson, "The Superiority of Sequential Search: A Calculation," *Southern Economic Journal*, Vol. 43, No. 4 (April, 1977), pp. 1594–1598, www.jstor.org/stable/i243526.

Thomas Ferguson, *Optimal Stopping and Applications* is freely available online at www.math.ucla.edu/ tom/Stopping/Contents.html. Ferguson offers a technical, mathematical presentation of search theory.

C. J. McKenna, *The Economics of Uncertainty* (New York: Oxford University Press, 1986), is a concise, nontechnical introduction to imperfect information models.

John Allen Paulos, *Beyond Numeracy* (New York: Alfred A. Knopf, 1991), p. 64, discusses the optimal interview problem with an easy, intuitive style.

This course surveys research which incorporates psychological evidence into economics. Topics include: prospect theory, biases in probabilistic judgment, self-control and mental accounting with implications for consumption and savings, fairness, altruism, and public goods contributions, financial market anomalies and theories, impact of markets, learning, and incentives, and memory, attention, categorization, and the thinking process.

MITOpenCourseware

Chapter 8

Behavioral Economics

The field of Behavioral Economics (and Behavioral Finance) is a growing research area that focuses on how decisions are actually made. It is closely tied to psychology and neuroscience. Behavioral economists reject the idea of utility maximization as an assumed black box. Both experimental methods and sophisticated procedures (such as MRI brain scans) are used to examine how real-world problems are actually solved. A number of results have emerged that challenge the conventional wisdom in mainstream economics.

One area of long-standing interest in psychology involves repeated choice problems. This chapter focuses on a particular kind of repeated choice in which the satisfaction obtained currently depends on past decisions. This is called distributed choice.

Suppose you are deciding whether to watch TV or play a video game. You face this choice repeatedly. The satisfaction from watching TV or playing a video game depends on how often that choice has been made before. What is the best combination of TV and video games over a period of time and, more importantly, how well do people handle this kind of repeated choice?

Instead of explaining why the repeated choice optimization problem is difficult and presenting results from human trials, it is more fun (and you will learn more) to let you first participate in an experiment.

The Choice Game

STEP Open the Excel workbook *Melioration.xls* and read the Intro sheet, then go to the *Choice Game* sheet to play this simple game.

Your goal is to click the A or B buttons as many times as possible in 10 minutes. When you make a choice, by clicking on one of the buttons, you are forced to wait. Waiting is costly because you cannot click (make another choice) while waiting.

STEP Click the Practice option button (near the top left corner of the screen) to see how the game works.

You get up to 100 practice trials. In practice mode, time is not kept. You can take as long as you want between button clicks. Practice now.

There is definitely something going on that you are trying to figure out and there is an optimal strategy. You can click the same button over and over or switch back and forth.

Are you ready to play? Unlike practice, when you play, a timer will be running. You will not use the buttons on the sheet like you did in practice mode. The buttons will be on a dialog box, right next to each other. You will have 10 minutes to make as many choices as possible. The time remaining will be displayed as you play.

Ten minutes might be too long for you to play so click the Exit Now button if you want to stop playing. As long as you start play and make a few choices, you will be able to continue working and learning about melioration.

STEP Click the Play option button. Good luck!

After you finish the game, a message box displays your score and a *Results* sheet shows a record of your picks. It reports results based on a full ten minutes of play, so if you stopped prematurely, you can ignore your results.

Let's deconstruct this game and see how it works. Figure 8.1 shows the first 10 choices made by a player. The player started with A, then switched to B with his 7^{th} choice, but switched back to A, then ended with B.

	A	B	C	D
1	Choice Number	Pause Time A	Pause Time B	Choice Made
2	1	2.00		A
3	2	2.40		A
4	3	2.80		A
5	4	3.20		A
6	5	3.60		A
7	6	4.00		A
8	7		8.00	B
9	8	4.40		A
10	9	4.80		A
11	10		7.60	B

Figure 8.1: Ten plays of the game.

STEP You can see the full record of yet another player by clicking the Show Example button (near cell G9 in the *Results* sheet, which was revealed when you finished playing the choice game).

This player tried streaks of A and B. Notice how the time paused changed.

These results sheets also compare the number of choices made to the maximum possible and computes a score as a percentage of the maximum. Let's find out how the maximum can be attained and why people are usually so bad at playing this game.

Actual Results

Experimental trials with this game were conducted by Herrnstein and Prelec (1991) and you can compare how you did to the average result (and to the player in the *MoreResults* sheet).

STEP Click the Show Data button in the *MoreResults* sheet.

The *Data* sheet shows how 17 subjects played the choice game that you just played. Each dot in the chart, reproduced in Figure 8.2, shows the fraction of times that a player chose A (on the x axis) and the corresponding average delay endured by that player (on the y axis). The player with the shortest

delay, the first one in the table, also has the most choices (number of choices = 600/average delay) and is the winner in this set of players. How did you do?

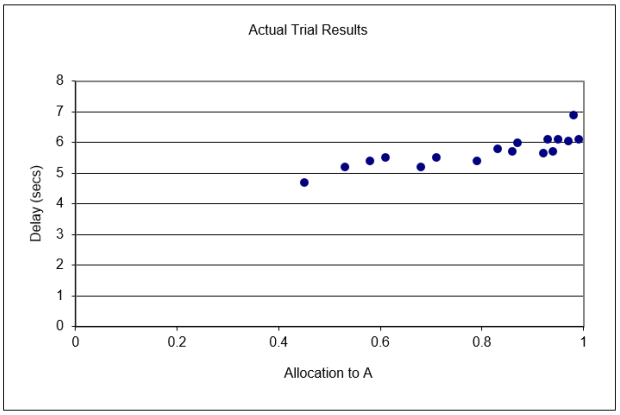


Figure 8.2: Actual results from a single session of the choice game. *Source: Melioration.xls!Data, revealed after game played.*

STEP To add your result to the chart in Excel, copy your results from cells J2 and K2 of the *Results* sheet, select cell A23 in the Data sheet, and Paste Special (Values) (or simply type in the two numbers). A red dot will appear in the chart. This shows how you did.

Did you beat the best player out of the 17 in the chart? We know you could have because even the best player in that group of 17 failed to optimize. The explanation for this failure requires that we understand the delay function for each choice.

The heart of the choice game is the wait time between choices. The duration of the pause is a function of the previous 10 choices (including the current choice). For choice A, the wait time, in seconds, is 2 + 0.4 x Proportion of A Choices in the last 10 choices. So, if the last 10 choices had been B, then A would have a very short and satisfying pause time of just 2 seconds. As you click on A, however, the pause time for choice A rises by 0.4 seconds until it hits a maximum of 6 seconds.

Choice B's wait time is determined by 8 - 0.4 x Proportion of B Choices in the last 10 choices. As you click on B, the duration of the pause gets lower and lower until reaching a minimum of 4 seconds.

STEP Confirm that the wait times were determined as described by returning to the three results sheets and examining the pause times in columns B and C.

You can see that the first clicks of A and B had pause times of 2 and 8 seconds, respectively. You can also check that each pause time is following the functions described above. The *MoreResults* sheet with the streaky A and B strategy makes it easy to see the mechanics of the choice game.

Choice A exhibits increasing marginal cost—every time you click on A, you are penalized and forced to wait longer. Choice B rewards you with a decrease in wait time when it is clicked, but the wait time starts very high so you have to be persistent and stick to it. Plus, choice A is always 2 seconds lower than choice B so you are constantly being lured toward choice A.

Most people play this game by being attracted to A's short wait time, until it gets unbearable and they switch to B. But they can't stay with B long because it is painful to wait at first and they do not have the patience and self-discipline to stick with B. Sound familiar? B could be exercise or dieting or studying—you know you should and it gets easier if you stick to it, but it can be hard to start.

Now that you know the rules of the game, how do you actually optimize with this game? Simple—start with choice B and never deviate.

STEP To see this optimal strategy in action, go to the *Solution* sheet by clicking the Show Solution button in the *Data* sheet (below the chart).

Column B shows what happens when you exclusively choose A. It starts well, but you end up with many 6 second pauses.

STEP Scroll down to see that you make 103 choices in 600 seconds, yielding an average delay of 5.8 seconds. This is a poor outcome.

Column F displays what happens when B is exclusively chosen. The first few wait times are long, but each choice of B lowers the wait time until the minimum, 4 seconds, is reached.

STEP Scroll down to see that clicking choice B every time lets you make 144 choices (with an average delay of 4.167 seconds).

The strategy of choosing B exclusively cannot be beat (except for an endgame correction, which is one of the exercise questions). If the player switches from B to A, the temporary gain is swamped by higher wait times when the inevitable switch back to B occurs.

To be sure that this point is clear, consider switching after having reached the 4 second minimum pause time for choice B. What would happen?

STEP Change cell K15 (in the *Solution* sheet) to A.

Five consecutive A choices are made and each one has a pause time less than or equal to four seconds, as shown in column L. Thus, we have saved time. But when we switch back to B (since we know A's pause time will continue to rise and we can get to 4 seconds with B), we have to suffer higher pause times. The trade-off is not worth it. We end up making fewer choices (142 instead of 144) and suffering a longer average delay.

The *Solution* sheet makes clear the following key point: The optimal strategy is to choose B exclusively and never deviate. If you failed to do this, do not worry; you have plenty of company. Very few humans figure this out.

Melioration Explained

Herrnstein and Prelec (1991) designed the experiment to test for the presence of something called melioration (pronounced mee-lee-uh-RAY-shun). To *meliorate* (or *ameliorate*) means to make better or more tolerable. Melioration says that we are drawn to choices that *immediately* reduce pain or give immediate satisfaction. We do a poor job of maximizing when there is a trade-off between short- and long-run returns. We are shortsighted and look to make immediate improvements. In fact, melioration has been found in other animals besides humans.

The attraction of switching to A and having the pause time fall is melioration at work. The immediate pain of waiting is lessened and, thus, players are drawn toward choice A.

In addition to the actual choices from the 17 players, Figure 8.3 shows wait times for choices A and B given the proportion of A choices in the previous 10. It is easy to see, once again, that the optimal solution is to choose B exclusively because that lets you travel down the solid line to the intercept

at 4 seconds. If you ever jump on the A train, you are swept upwards toward a 6-second wait time.

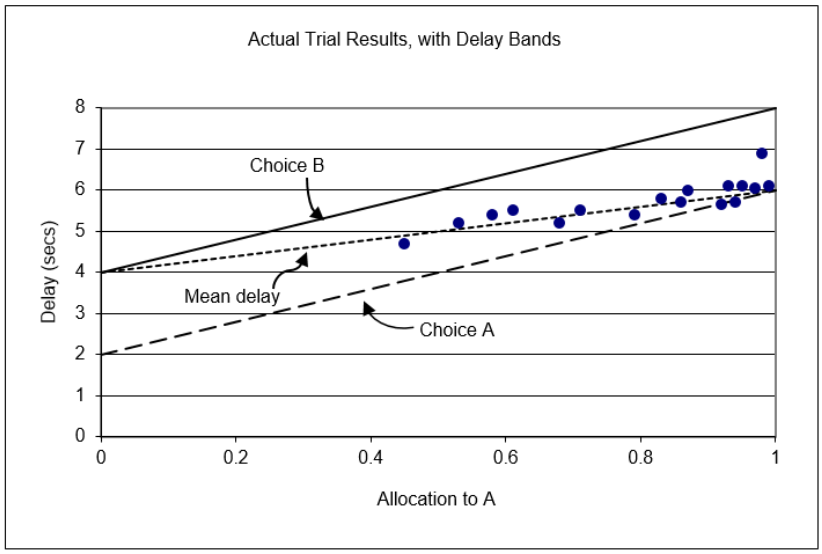


Figure 8.3: Understanding melioration.
Source: Melioration.xls!Data, revealed after game played.

Figure 8.3 shows that if the last 10 choices were B and then A was chosen, the player would immediately gain a reduction in wait time from 4 to 2 seconds (jumping from the higher to the lower line). For a few choices, the player would be better off, but after the 5^{th} consecutive A choice, the wait time would be greater than 4 seconds. The player would be forced to endure longer wait times than would have been obtained by sticking with B.

Furthermore, it is hard to switch to B because wait time immediately jumps by 2 seconds. The player will have to suffer through the ride down the B line, with choice A promising a 2-second decrease with every click. The immediate attraction of the 2-second decrease is the core of the melioration process that guides subjects to choose A.

Figure 8.3 makes clear that the 17 human subjects who played the choice game failed to optimize. The fraction of allocation to A should be zero, but most players do not do this. This begs the question, so what?

Herrnstein and Prelec (1991) argue that the lack of optimization is a big deal. For them, choice is often not a single, isolated decision, but a series of many

decisions, distributed over time. Frequency of athletic exercise, buying lottery tickets, choices of restaurants, and rate of work in freelance occupations are some of the examples offered.

For all of these distributed choice problems, melioration is common and this means people systematically fail to optimize. "This would imply that preferences as revealed by the marketplace may be a distortion of the true underlying preferences" (Herrnstein and Prelec, 1991, p. 137). Melioration helps explain complaints about one's own behavior (such as exercising too little), which is part of a growing literature on self-control. It also may contribute to the study of impulsiveness and addiction.

Of course, this presumes that the laboratory findings carry over to real-world settings. This is often an Achilles' heel of experimental economics. Results are often criticized as having little external validity because they are based on fake scenarios played by college students. Herrnstein and Prelec (1991) acknowledge that little money was at stake (they paid their players based on performance), but they rely on two other motivating factors. "First, delays are genuinely annoying and the difference between two and four seconds is not trivial, as any computer user will appreciate. Second, the 'puzzle' nature of the experiment presents a challenge that is presumably satisfying to solve" (Herrnstein and Prelec, 1991, p. 144).

Others have tried to nail down exactly what causes melioration and how it can be overcome. Neth, Sims, and Gray (2005, p. 357) were surprised:

> We hypothesized that frequent and informative feedback about optimal performance might be the key to enable people to overcome the documented tendency to meliorate when choices are rewarded probabilistically. Much to our surprise, this intuition turned out to be mistaken. Instead of maximizing, 19 out of 22 participants demonstrated a clear bias towards melioration, regardless of feedback condition.

The Future of Behavioral Economics

With faculty, courses, conferences, and specialized journals, there is no doubt that Behavioral Economics is here to stay. In 2002, the Nobel Prize in Economic Sciences was awarded to Daniel Kahneman and Vernon Smith for

work incorporating psychology and laboratory methods in the study of decision making. Richard Thaler won the Nobel in 2017 for his contributions to behavioral economics.

Unlike conventional economics, which simply assumes optimizing behavior and rationality, behavioral economists seek to determine under what conditions agents struggle to optimize. They work with psychologists and neuroscientists to devise tests and laboratory experiments. The key result is that they find persistently sub-optimizing behavior.

Melioration is but one simple example of work in this area. Melioration means that decision makers fail to optimize because they focus on the small (immediate, single choice) instead of the large (future, many choices). This can be applied any time that incremental steps lead to an undesirable place:

> A person does not normally make a once-and-for-all decision to become an exercise junkie, a miser, a glutton, a profligate, or a gambler; rather, he slips into the pattern through a myriad of innocent, or almost innocent choices, each of which carries little weight. Indeed, he may be the last one to recognize "how far he has slipped," and may take corrective action only when prompted by others. (Herrnstein and Prelec, 1991, p. 149)

According to the behavioral economists, the list of examples where humans struggle to optimize is actually quite long. Evaluating probabilities (such as risk), choice over time, and misperception of reality are all areas being actively studied.

It remains unclear whether the results being generated by behavioral economists are merely a series of peculiar puzzles that will extend the boundaries of economics or more serious anomalies that will one day bring down the paradigm of rationality and optimizing behavior that is the hallmark of modern, mainstream economics.

Exercises

If you did the Q&A problems and changed the parameters, set them back to the original values (2 and 0.4 for A and 8 and -0.4 for B).

1. With your observation included, copy and paste the chart titled *Actual Trial Results* in a Word document. Comment briefly on how you did.

2. What endgame correction could be implemented to increase the total number of choices? What is the true, exact maximum number of choices? Explain.

   Herrnstein and Prelec (1991), p. 142, point out that, "In fact, the subjects showed no evidence of having been influenced by the endgame contingency."

3. With columns Q:U in the *Solution* sheet, use Solver to find the optimal solution to the choice game. Notice how the choice variables have been constrained. How does Solver do? Explain.

4. Training someone to touch type does not guarantee continued touch typing in the workplace. How would melioration explain this result?

References

The epigraph is from a course available freely at ocw.mit.edu. The course description in the epigraph was from the Spring 2004 version of Behavioral Economics and Finance (see ocw.mit.edu/courses/economics/14-127-behavioral-economics-and-finance-spring-2004/). The readings for this course include introductory and more advanced work.

The repeated choice problem in this chapter is based on two papers: (1) Richard J. Herrnstein and Drazen Prelec, "Melioration: A Theory of Distributed Choice," *The Journal of Economic Perspectives*, Vol. 5, No. 3 (Summer, 1991), pp. 137–156, www.jstor.org/stable/1942800 and (2) Herrnstein and Prelec's "Melioration," pages 235–263 in *Choice Over Time*, edited by George Loewenstein and Jon Elster (1992).

Herrnstein, a psychologist, teamed up with Charles Murray, a political scientist, to write a controversial book titled *The Bell Curve: Intelligence and Class Structure in American Life* (1994). The book argued that nature (IQ) is more important than nurture (socioeconomic status) in explaining a wide range of outcomes.

Another paper specifically focused on melioration is Hansjörg Neth, Chris R. Sims, and Wayne D. Gray, "Melioration Despite More Information: The

Role of Feedback Frequency in Stable Suboptimal Performance," *Proceedings of the Human Factors and Ergonomics Society* 49th Annual Meeting, 2005, doi.org/10.1177/154193120504900330.

There are many books on behavioral economics and finance. A classic is from Nobel Prize winner Richard Thaler, *The Winner's Curse: Paradoxes and Anomalies of Economic Life* (1994). This is a good place to start learning about behavioral economics. Other good reads include the following:

Dan Ariely, *Predictably Irrational: The Hidden Forces that Shape Our Decisions* (2008).

Daniel Kahneman, *Thinking Fast and Slow* (2011).

Michael Lewis, *The Undoing Project: A Friendship that Changed Our Minds* (2016).

Richard Thaler, *Misbehaving: The Making of Behavioral Economics* (2015).

Richard Thaler and Cass Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (2008).

How use doth breed a habit in a man!

William Shakespeare

Chapter 9

Rational Addiction

This chapter is different. It does not have steps that you follow as you work in Excel. It does not have any exercise questions. There is an Excel file that you will open and work on, but it is entirely self-contained. Just open the file and start reading.

Before you begin, however, consider a little of the science behind learning. Once we know how we learn, then we can optimize!

The Neuroscience of Learning

Suppose you want to improve your free throw shooting and you really cared about this so you decided to practice for one hour per day for two weeks. Most people think that standing at the free throw line and shooting free throws would be the best use of your time, but this is wrong. A much better use of your one hour per day is to shoot from all over the court—spend 10 minutes in one spot, then move to another spot, varying distance from say 10 to 20 feet (the free throw line is 15 feet from the basket). This is *interleaved practice* and it works also for learning and studying.

Interleaved practice is counter-intuitive and paradoxical. Many coaches refuse to believe it, but careful controlled experiments in a variety of applications reveal it is a fundamental principle (Brown, et al., 2014). It works for physical skills (don't throw 100 curve balls, interleave with other pitches), memorization (don't repeat one thing, interleave items), and higher learning—reflect on how this book has repeated concepts like elasticity in a variety of applications.

In addition to interleaving, below is a list of best-practice learning strategies that you can apply to every course you take:

1. Interleaved Practice (switching)
2. Spaced Practice (avoid cramming)
3. Elaboration (invent your own how and why questions)
4. Concrete Examples (the more specific, the better)
5. Dual Coding (words and visuals)
6. Retrieval Practice (repeatedly recall what you know)

Unbeknownst to you, this book has been using all of these strategies to help you learn.

To get more information on these six science-based ways to learn more efficiently, visit these two web sites:

- www.learningscientists.org/posters
- www.youtube.com/watch?v=CPxSzxylRCI

And one more thing that you believe about learning that is wrong: you think your ability to learn economics (or math or music) is preordained. Your brain either has a knack for economics or it does not and, if not, you cannot learn economics (or math or music). This is wrong.

Neuroscience makes clear that your brain is *plastic*. It is moldable and flexible. You have already learned a great deal of economics, math, and Excel. Yes, some details are fuzzy and you have not mastered every single thing, but keep trying. As you see more examples and applications, it gets easier to grasp and your understanding deepens.

Rational Addiction

As you work on the Excel file, you will be reviewing concepts and feel comfortable with Solver, charts, and Excel itself. This will reinforce basic material that you already know, but you will also be exposed to some new ideas as you continue to master the economic way of thinking.

This application is controversial and generates passionate debate. Non-economists, especially, find it outrageous. After you finish, you can make up your own mind on what you think about it.

Open *RationalAddiction.xlsm* to begin.

References

The epigraph is from Shakespeare's *Two Gentlemen of Verona.* The *Conversation* sheet in *RationalAddiction.xlsm* explains what it means.

The full story behind the puzzling interleaved practice phenomenon and much more about how we learn is in Peter Brown, Henry Roediger III, and Mark McDaniel (2014), *Make it Stick: The Science of Successful Learning*.

Part II

The Theory of the Firm

> For Friedman, lack of realism of assumptions is not a virtue. It is a necessary evil: to base theories on absolutely realistic assumptions is like drawing a map on a one-to-one scale.
>
> Mark Blaug

Overview

Consumer Theory focuses on the buyer. It models a consumer's optimization problem and emphasizes deriving a demand curve as the most important result.

The Theory of the Firm is about the seller. Firm decisions about inputs and outputs are modeled as optimization problems. The key result will be deriving a supply curve.

The chapters are organized as shown in Figure II.1. Notice that the production function is the first idea presented. It plays a role in each of the three optimization problems faced by the firm.

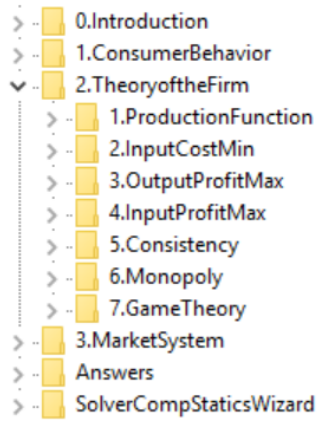


Figure II.1: Content map with focus on the theory of the firm.

Figure II.1 also provides a broad overview of the entire landscape. We have completed the Theory of Consumer Behavior and, once we finish our work in the Theory of the Firm, we will be ready to analyze the behavior of consumers and firms together in part III, the Market System.

Unlike the Theory of Consumer Behavior, the Theory of the Firm is made up of three interrelated optimization problems.

1. Input cost minimization: Choose inputs to minimize the cost of producing a given level of output. Derive the cost function by changing q and tracking the minimum total cost.

2. Output profit maximization: Choose output to maximize profits. Derive the supply curve by changing the price and tracking the optimal output.

3. Input profit maximization: Choose inputs to maximize profits. Derive an input demand curve by changing an input price and tracking optimal input use.

Of course, optimization and comparative statics play a prominent role, but watch out for these three crucial innovations in the Theory of the Firm.

1. Market structure: The Theory of the Firm includes the market environment as an important consideration in the model. The firm can be a *price taker*, a perfectly competitive firm, or a *price maker*, a monopolist. There are many other market structures, for example, oligopoly (where there are a few firms) and monopolistic competition.

2. Time period: The Theory of the Firm distinguishes between long run and short run decision making horizons. In the *long run*, all factors are freely variable and firms may enter or exit the industry. In the *short run*, at least one input (usually capital) is fixed and the firm may cease production (shut down), but it must pay fixed costs whether it produces or not.

3. Output is cardinally measurable: Unlike utility, the output produced by a firm and the resulting revenues, costs, and profits can be directly observed and measured on a cardinal scale. Thus, we will be able to use and interpret the Lagrangean multiplier.

Methodology

The assumptions underlying the Theory of Consumer Behavior are never seen in reality and the Theory of the Firm doubles down on this strategy by making even more outlandish assumptions. The time has come to explain why economists do this.

Each discipline has its own rules for determining truth and acceptable procedures for producing knowledge. These rules and norms are known as the *methodology*. For example, economics utilizes highly abstract models. The assumptions of these models are plainly unrealistic and false. Real-world human beings do not behave like perfectly rational, calculating machines. Then why do economists assume they do?

The classic defense of unrealistic assumptions is "The Methodology of Positive Economics," the first chapter in Milton Friedman's (1953) *Essays in Positive Economics*. Friedman's argument was initially controversial, but it became conventional thinking in economics. Friedman urged economists to ignore how unrealistic the assumptions were and focus on the predictive power of a model. If you want to predict how billiard balls will move when hit by an expert pool player, vectors and complicated mathematics are involved.

> It seems not at all unreasonable that excellent predictions would be yielded by the hypothesis that the billiard player made his shots as if he knew the complicated mathematical formulas that would give the optimum directions of travel, could estimate accurately by eye the angles, etc., describing the location of the balls, could make lightning calculations from the formulas, and could then make the balls travel in the direction indicated by the formulas. (Friedman, 1953, p. 21)

The Theory of the Firm (like the Theory of Consumer Behavior) is built on the idea of decision makers acting as if they were rationally calculating and optimizing agents. This is plainly unreal, but the point is not to describe how consumers or firms actually make decisions. Instead, we want a model that makes predictions about changes in output, for example, as product price changes (this is a supply curve).

We know firms do not take price as given and there is no such thing as perfect competition in the real world, but we assume this because we are not trying to build an accurate representation of an actual firm. Instead, we want to able to predict how a firm responds to a price shock—just like we want to predict how a billiard ball will move when struck.

It is quite easy to forget the methodology of economics and find oneself wondering how economists can believe such a ridiculously unreal and abstract model of a firm. Remember, economists do not test theories via the assumptions—it is the implications that matter.

The usefulness of abstract models and unrealistic assumptions often drives opinion or evaluation of the work. For example, the economic theory of rational addiction says addicts rationally "choose" harmful addiction or dangerous activities. To some, this is so obviously untrue that they cannot engage with the theory.

The term *homo economicus*, a non-existent version of *homo sapiens*, is used to mock the narrow, calculating humans that inhabit the made-up world of economics.

In France, over a thousand Economics graduate students signed a letter in 2000 attacking the abstract, unrealistic, mathematical training they were receiving. This launched the Post-Autistic Economics movement.

Friedman's view came to dominate mainstream economics, but it did not end the argument in philosophical circles and heterodox economics. The debate about methodology rages on while most economists continue to build and work with highly abstract, completely unrealistic models.

Your Role—A Reminder

As before, mastery of the Theory of the Firm requires your effort, energy, and engagement. Be sure to **experiment**, changing cells and asking "what if" questions as you proceed through the Excel workbooks. Focus on the repeated patterns and continue to add to your stock of knowledge.

Remember that economics has a core logic that has been referred to as "the economic way of thinking" or "the economic approach." Learning to see and think like an economist should be your ultimate goal.

References

The epigraph is from page 703 of the third edition (1978) of Mark Blaug's *Economic Theory in Retrospect* (originally published in 1962). Blaug's concluding chapter, "A Methodological Postscript," is a good review of how theories develop and knowledge grows.

Methodology is part of the *philosophy of science*. Economists pay little attention to methodology, but that does not mean it is unimportant.

Let us choose that function $P' = bL^kC^{k-1}$ and find such numerical values of b and k that P will "best" approximate P [output] in the sense of the Theory of Least Squares. Then relative to the indices and the period we have the norm $P = 1.01L^{\frac{3}{4}}C^{\frac{1}{4}}$.

Charles W. Cobb and Paul H. Douglas

Chapter 10

Production Function

The production function is the backbone of the Theory of the Firm. It describes the current state of technology and how input can be transformed into output.

The production function can be displayed in a variety of ways, including product curves and isoquants. In every optimization problem faced by the firm, the production function is included.

Key Definitions and Assumptions

Inputs, also known as factors of production, are used to make output, sometimes called product. As shown in Figure 10.1, the firm is a highly abstract entity—a black box—that transforms inputs into output.

Figure 10.1: The black box nature of the firm.

The specific details of how the firm is organized and how it actually combines the inputs to make goods and services is ignored by the theory, hidden in the black box.

Inputs are often broken down into large categories, such as land, labor, raw materials, and capital. We will simplify even further by collapsing everything that is not labor into the capital category.

Labor, L, is human toil and effort. It is measured in units of time, usually hours.

Capital has a confusing history in economics. As a factor of production, *capital*, K, means things that produce other things, such as machinery, tools, or equipment. That is different from financial or venture capital that is a fund of money. The title of Karl Marx's famous book, *Das Kapital*, uses capital in the sense of wealth, denominated in money. The Theory of the Firm's K is measured in numbers of machines.

Like labor, capital is rented. The firm does not own any of its machines or buildings. This is extremely unrealistic, but allows us to avoid complicated issues involving depreciation, financing of machinery purchases (debt versus equity, for example), and so on.

Another extreme simplifying assumption is that there is no time involved. Like the consumer maximizing utility subject to a budget constraint, the firm exists only for a nanosecond. It makes decisions about how much to produce to maximize profits with no worries about inventories or the trajectory of future sales. It produces the output in an instant.

We avoid complications arising from the production of more than one good or service by assuming that the firm produces only one product. That makes revenues simply price times quantity sold of the one product.

Without going into detail again about unrealistic assumptions, it seems helpful to point out that we are not trying to build an accurate model of a real-world firm. Our primary goal is to derive a supply curve. We want to know how a firm responds to a change in price, ceteris paribus. By assuming away many real-world complications, we can model the firm's maximization problem, solve it, and do comparative statics to get the supply curve.

Mathematical Representation

Just like the Theory of Consumer Behavior, which uses a utility function to model tastes and preferences, the Theory of the Firm uses a production function to capture the ability of firm's to transform inputs into outputs. Unlike utility, production is objective and observable. We can count how much output is made from a given number of hours of labor and machines.

The *production set* describes all of the technologically feasible outputs from a given amount of inputs. The *production function* describes the maximum output possible from a given amount of inputs. Notice how the production function assumes the inputs are being used in the best way possible.

The most abstract, general notation for a production function is $y = f(L, K)$. The $f()$ represents the technology available to the firm. A specific, concrete example of a production function is the Cobb-Douglas functional form: $y = AL^{\alpha}K^{\beta}$. Let's see what it looks like in Excel.

STEP Open the Excel workbook *ProductionFunction.xls*, read the *Intro* sheet, then go to the *Technology* sheet to see an example of the production function.

In Figure 10.2, the production set is the surface of the 3D object and everything inside; the production function is just the surface.

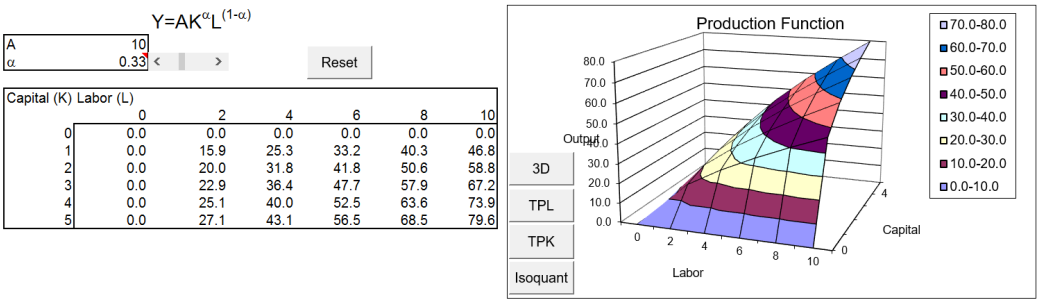


Figure 10.2: The production function.
Source: ProductionFunction.xls!Technology.

The production function implicitly includes an already solved engineering optimization problem—it gives the maximum output from any given combination of inputs. In other words, we are assuming that the inputs are organized in their most productive configuration and nothing is wasted.

Notice that the Cobb-Douglas function on the Technology sheet has been set up so it can be controlled by a single parameter, α (alpha), by making the exponents α and $(1 - \alpha)$. Use the scroll bar to change alpha and notice how the shape of the production function surface changes. Alpha is a parameter that takes values between zero and one.

STEP Click the Reset button to return the sheet to its default, initial position.

Product Curves

In addition to the 3D view, the production function can be displayed in other ways. To graph the production function in two dimensions, we need to suppress an axis. If we keep output and suppress one of the input axes we get a *total product curve*. If we suppress output and keep the two inputs, we get an *isoquant*.

Product and output mean the same thing. The total product curve is the number of units of output produced as one input is varied, holding the other constant.

STEP Click the TPL and TPK buttons to see the product curves for labor and capital.

In addition to the total product curves, there are average and marginal product curves. The *average product* is simply output per unit of input. Thus, the average product of labor is Y/L and the average product of capital is Y/K.

The *marginal product curves* tell us the *additional* output that is produced as input is increased, holding the other input constant. Marginal product can be computed based on finite-size changes in an input or via the derivative.

Via calculus, the marginal product is simply the derivative of the production function with respect to the input. For the Cobb-Douglas function in the *Technology* sheet, the marginal products are found by taking the partial derivatives with respect to L and K:

$$MP_L = \frac{\partial Y}{\partial L} = (1-\alpha)AK^{\alpha}L^{(1-\alpha)-1} = (1-\alpha)AK^{\alpha}L^{-\alpha}$$

$$MP_K = \frac{\partial Y}{\partial K} = \alpha AK^{\alpha-1}L^{1-\alpha}$$

STEP Scroll down and click on cell C52 to see that the marginal product is computed via the change in output from an increase of 2 hours of labor, with $K = 4$.

This computes the marginal product of labor as the rise over the run from $L = 0$ to $L = 2$ on the total product curve.

STEP Click the Show MPL button and then click on cell C58 to reveal the marginal product computed via the derivative.

Since the total product is a curve, the slope of the tangent line at $L = 2$ is not the same as the rise over the run from one point to another.

STEP Now look at the total, marginal, and average product curves.

Notice how the product curves are drawn based on a given amount of capital. If the amount of capital changes, then the product curves shift.

Marginal and average product can be graphed together because they share a common y axis scale, output per unit of input. The total product curve can never be graphed with the marginal and average product curves because the total product curve uses output as its y axis scale.

The graphs demonstrate that when total product increases at a decreasing rate, marginal product is decreasing. When total output increases at a decreasing rate as more input is applied, ceteris paribus, we are obeying the *Law of Diminishing Returns*. As long as alpha is between zero and one, our Cobb-Douglas production function exhibits diminishing returns.

The Law of Diminishing Returns does not deny that there can be ranges of input use where output increases at an increasing rate. It says that, eventually, continued application of more input along with a fixed factor of production must lead to diminishing returns in the sense that output will increase, but not as fast as before. Thus, the Law of Diminishing Returns is simply a statement that marginal productivity must, eventually, be falling.

As with utility, the Cobb-Douglas functional form is convenient, but there are many, many other functional forms available.

STEP Proceed to the *Polynomial* sheet to see a different functional form. The charts are strikingly different than before.

Unlike the Cobb-Douglas functional form, which always shows diminishing returns, the polynomial production function exhibits all three different phases of returns: increasing, diminishing, and negative returns.

At low levels of labor use, output is increasing at an increasing rate so the total product curve is curved upward and marginal product is increasing. In this range, as long as marginal product is rising and output is increasing at an increasing rate, output rockets upward, growing faster and faster.

When the marginal product curve reaches its peak, the total product curve is at an inflection point. From here, additional labor leads to increases in output, but at a decreasing rate, leveling off as L increases. We say that diminishing returns have set in.

The *Polynomial* sheet is color coded so it is easy to see where the total product curve changes character. Cells with yellow backgrounds signal the range of labor use where diminishing returns apply.

As more and more labor is used, total product reaches its maximum point (where marginal product is zero). Beyond this point, we are in a range of negative returns. This is a theoretical possibility, but not a practical one. No profit-maximizing firm would ever operate in this region because you can get the same amount of output with fewer workers.

It is worth remembering that the Law of Diminishing Returns does not say that we always have diminishing returns for every level of labor use. Instead, the law says that, eventually, diminishing returns will set in. It is also important to understand the difference between diminishing and negative returns. The former says output is rising, but slower and slower, while the latter says output is actually falling.

Notice the relationship between the marginal and average product curves. It is no coincidence that the marginal product curve intersects the average product curve at the maximum value of the average product. There is a guaranteed relationship between marginal and average curves: Whenever the marginal is greater than the average, the average must be rising and whenever the marginal is less than the average, the average must be falling. Thus, the only time the two curves meet is when the marginal and average are equal.

STEP Change the parameter for the b coefficient from 30 to 40.

Notice that the S shape becomes much more linear. The range of increasing returns is larger and we do not hit negative returns over the observed range of L from 0 to 25.

STEP Set the parameter for the b coefficient to 80.

Over the observed range of L from 0 to 25, we see only increasing returns.

STEP Change the δL parameter from 1 to 2. This makes L go up by two and the range goes from 0 to 50.

Diminishing returns do kick in; it just takes more labor for the Law of Diminishing Returns to be observed when the b coefficient is set to 80.

Diminishing versus Decreasing Returns

One extremely confusing thing about the Law of Diminishing Returns has to do with another concept called *returns to scale*. Unlike the Law of Diminishing Returns—which is based on applying more and more of a particular input while holding other inputs constant—returns to scale focuses on the effect on output of changing *all* of the inputs by the same proportion.

There is no law for returns to scale. A production process may exhibit increasing, decreasing, or constant returns to scale, across all values of input use. For example, the Cobb-Douglas function in the *Technology* sheet has constant returns to scale because if you double L and K, you are guaranteed to double output.

You can see this is true by comparing the points 2,2 and 4,4 in the table in the *Technology* sheet. A more complete demonstration uses a little algebra. We begin with the production function:

$$AK^{\alpha}L^{1-\alpha}$$

Next, we double both L and K:

$$A(2K)^{\alpha}(2L)^{1-\alpha}$$

We expand the terms with exponents:

$$A(2^{\alpha})(K^{\alpha})(2^{1-\alpha})(L^{1-\alpha})$$

We collect the "2" terms:

$$A(2^{\alpha+(1-\alpha)}(K^{\alpha})(L^{1-\alpha})$$

The alphas add to zero ($\alpha - alpha = 0$) so we get:

$$A2K^{\alpha}L^{1-\alpha}$$

Thus, we have shown that doubling the inputs from any input levels leads to doubling the output, and this is called constant returns to scale. If the exponents in the Cobb-Douglas function do not sum to 1, then the function does not exhibit this property.

The Cobb-Douglas function in the *Technology* sheet obeys the Law of Diminishing Returns for each input (with $0 < \alpha < 1$), yet it has constant returns to scale. Do diminishing returns imply decreasing returns to scale? No, absolutely not. The two concepts are independent. They ask different questions. The Law of Diminishing Returns is about what happens to output when a single input is increased, ceteris paribus, and decreasing returns to scale says that output will less than double when all inputs are doubled.

Isoquants

In addition to product curves, another way to represent the production function uses the isoquant. The prefix *iso*, meaning equal or the same (as in isosceles triangle), is combined with *quant* (referring to the quantity of output) to convey the idea that the *isoquant* displays the combinations of L and K that yield the same output.

STEP Return to the top of the *Technology* sheet and click the Isoquant button (near cell H28) to see the isoquant map, as displayed in Figure 10.3.

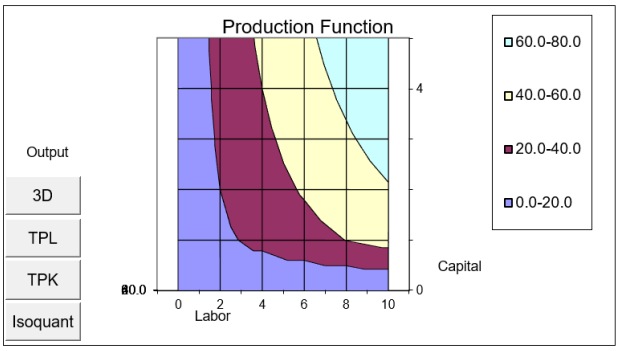


Figure 10.3: Isoquants for a Cobb-Douglas technology.
Source: ProductionFunction.xls!Technology.

An isoquant is simply a 2D, top down view of the 3D surface. Unlike the product curves, which give a view from the side, the isoquant shows L and K on the x and y axes, respectively, and suppresses output.

Notice that Excel cannot correctly draw the isoquant map, putting garbled characters in the bottom left-hand corner of the chart and producing a squiggly, jagged display at the bottom.

You might be thinking that it looks a lot like an indifference map. There are definitely strong parallels between isoquants and indifference curves. Both are top-down views of a 3D object and, therefore, both are level curves or contour plots. Both are used to find and display the solution to an optimization problem.

However, there is one critical difference: unlike an indifference curve, each isoquant is, in principle, directly observable and the isoquants can be compared on a cardinal scale. With indifference curves, the utility function is a convenient fiction and the numerical values merely reflect rankings. No one cares that a particular indifference curve yields 28 utils of satisfaction. This is not the case for isoquants because the suppressed axis, output, is measurable. You can certainly say that one isoquant gives twice the output as another or that one isoquant gives 17 more units of output than another.

One way in which indifference curves and isoquants are the same is that we can compute the slope between two points or the instantaneous rate of change at a point on an isoquant. To avoid confusion with MRS, we call this slope the *technical rate of substitution*, TRS. With labor on the x axis and capital on the y axis, the TRS tells us how much capital we can save if one more unit of labor is used to produce the same level of output.

From one point to another, the TRS can be computed as the rise over the run, $\frac{\delta K}{\delta L}$. At a point, we compute the TRS as the ratio of the derivatives with respect to L and K from the production function:

$$TRS = -\frac{MP_L}{MP_K} = -\frac{\frac{\partial f(L,K)}{\partial L}}{\frac{\partial f(L,K)}{\partial K}}$$

Whereas MRS is universally used for the slope of an indifference curve, MRTS (marginal rate of technical substitution) is sometimes used for the slope of the isoquant. MRTS and TRS are perfect synonyms. We will use TRS.

The TRS (like the MRS) is a number that expresses the substitutability of labor for capital at a point on an isoquant. So, the TRS of two different L and K combinations on the same isoquant might be -100 and -2. The TRS $= -100$ value says that the firm can replace 100 units of capital with 1 unit of labor and still produce the same output. The isoquant would be steep at this point. If a point has a TRS $= -2$, 1 unit of labor can replace 2 units of capital to get the same output. The isoquant at this point would be much flatter than the point with the TRS $= -100$.

Just like the MRS, the TRS tells us how steep the isoquant is at a point. The steeper the isoquant, the more capital can be replaced by labor and still produce the same output.

Technological Progress

Over time, technology—our ability to transform inputs into output—improves. Electric power and computers are examples of technological progress that enables more output to be produced from the same input.

There are two kinds of technological change. The Cobb-Douglas functional form can be used to illustrate each type.

Suppose increased education improves the productivity of labor. This would be modeled as an increase in the exponent for labor in the Cobb-Douglas production function. Small changes, say from 0.75 to 0.751, lead to large responses (e.g., in output or labor use) because we are working with an exponent. This is known as labor-augmenting technological change.

We could also have a situation where the coefficient A in the function $AK^{\alpha}L^{\beta}$ increased over time. As A rises, the same number of inputs can make more output. This technological progress is said to be neutral (in terms of the utilization of L and K) because TRS does not depend on A.

We can show this by walking through the steps needed to find the TRS. First, we compute the marginal products of L and K from the function, $Y = AL^{\alpha}K^{\beta}$:

$$MP_L = \frac{\partial Y}{\partial L} = \alpha A L^{\alpha-1} K^{\beta}$$

$$MP_K = \frac{\partial Y}{\partial K} = \beta A L^{\alpha} K^{\beta-1}$$

The TRS is minus the ratio of the marginal products:

$$TRS = -\frac{MP_L}{MP_K} = -\frac{\alpha A L^{\alpha-1} K^{\beta}}{\beta A L^{\alpha} K^{\beta-1}} = -\frac{\alpha K}{\beta L}$$

The A terms cancel out, which means that the ratio of the marginal productivities of each input depends only on each input's exponent and the amount of the input used.

The Firm as a Production Function

The production function is the starting point for the Theory of the Firm. As with utility, many, many functional forms can be used to represent real-world production processes.

Economists represent the production function not as a 3D object, but in two dimensions. We get product curves (total, marginal, and average product curves) by focusing on output as a function of a single input, holding all other inputs constant. An isoquant suppresses the output and shows the different combinations of L and K that produce a given level of output.

The TRS is similar to the MRS, and it will play an important role in the understanding the firm's cost minimizing input choice.

Remember to keep straight the difference between the Law of Diminishing Returns and idea of returns to scale. The former applies more and more of a single input, holding all other inputs constant; the latter reports what happens to output when all inputs are changed by the same proportion. Those are two different things.

Exercises

1. Starting from a blank workbook, with $K = 100$, draw total, marginal, and average product curves for $L = 1$ to 100 by 1 for the Cobb-Douglas production function, $Q = L^{\alpha}K^{\beta}$, where $\alpha = 3/4$ and $\beta = 1/2$. Use the derivative to compute the marginal product of labor.

   Hint: Label cells in a row in columns A, B, C, and D as *L, Q, MPL,* and *APL*. For L, create a list of numbers from 1 to 100. For the other

three columns, enter the appropriate formula and fill down. For *MPL*, do not use the change in Q divided by the change in L; instead enter a formula for the derivative for the *MPL* at a point.

2. For what range of L does the Cobb-Douglas function in question 1 exhibit the Law of Diminishing Returns? Put your answer in a text box in your workbook.

3. Determine whether this function has increasing, decreasing, or constant returns to scale. Use the workbook for computations and include your answer in a text box.

4. From your work in question 3 and the comment in the text that you cannot have constant returns to scale "if the exponents in the Cobb-Douglas function do not sum to 1," provide a rule to determine the returns to scale for a Cobb-Douglas functional form.

5. Is it possible for a production function to exhibit the Law of Diminishing Returns and increasing returns to scale at the same time? If so, give an example. Put your answer in a text box in your workbook.

6. Draw an isoquant for 50 units of output for the Cobb-Douglas function in question 1.

   Hint: Use algebra to find an equation that tells you the K needed to produce 50 units given L. Create a column for K that uses this equation based on L ranging from 20 to 40 by 1 and then create a chart of the L and K data.

7. Compute the TRS of the Cobb-Douglas function at $L = 23$, $K = 312.5$. Show your work on the spreadsheet.

References

The epigraph comes from page 152 of "A Theory of Production" by Charles W. Cobb and Paul H. Douglas, *The American Economic Review*, Vol. 18, No. 1, Supplement, Papers and Proceedings of the Fortieth Annual Meeting of the American Economic Association (March, 1928), pp. 139–165, www.jstor.org/stable/1811556.

Douglas, an accomplished professor and US Senator from Illinois, explained how he and Cobb used the functional form that would be named after them:

> I was then temporarily lecturing at Amherst College, and consulted with my friend and colleague, Charles W. Cobb, a mathematician. At the latter's suggestion, the formula $P = bL_kC_{k-1}$ was adopted, a form that had also been used by Wicksteed and Wicksell.

See p. 904 in Paul H. Douglas, "The Cobb-Douglas Production Function Once Again: Its History, Its Testing, and Some New Empirical Values," The *Journal of Political Economy*, Vol. 84, No. 5 (October, 1976), pp. 903–916, www.jstor.org/stable/1830435

Chapter 11

Input Cost Minimization

Initial Solution

The Enfield Arsenal

Deriving the Cost Function

Cost Curves

The term "isoquant" was introduced by R. Frisch but originally for a different concept, for which it should have been reserved.

Joseph Schumpeter

11.1 Initial Solution

Input cost minimization is one of the three optimization problems faced by the firm. It revolves around the question of choosing the best combination of inputs, L and K, to produce a given level of output, q.

The best combination is defined as the cheapest one. The idea is that many combinations of L and K can produce a given q. We want to know the amounts of labor and capital that should be used to produce a given amount of output as cheaply as possible.

Of course, we answer this question by setting up and solving an optimization problem; then we do comparative statics. Because there is a constraint (we must produce the given q), we will use the Lagrangean method.

Setting up the Problem

The economic approach organizes optimization problems by answering three questions:

1. What is the goal?
2. What are the choice variables?
3. What are the given variables?

The goal is to minimize *total cost*, TC, which is simply the sum of the amount paid to the workers, wL, and the amount spent on renting machines, rK.

The endogenous variables are L and K. Labor is measured in hours and capital is the number of machines. The firm can decide to produce the given output by being labor intensive (using lots of labor and little capital), or roughly equal amounts of both, or by renting a lot of machinery and using little labor.

The exogenous variables are the input prices, *wage rate* (w), and the *rental price of capital* (r). The wage rate, or wage for short, is measured in \$/hour and the rental price of capital is \$/machine. We assume that the firm is a price taker in the markets for labor and capital so it can rent as much L and K as it wants at the given w and r. The amount to produce, q, is also an exogenous variable in this problem. We are not considering how much should be produced, but what is the best way to produce any given amount of output. Finally, the firm's technology, the production function, $f(L, K)$, is also given.

Because the firm has to produce a given amount of output, we know this is a constrained optimization problem. Our work in the Theory of Consumer Behavior has made us expert at solving this kind of problem. As you will see, the analysis is similar, but there are some striking differences.

One thing that does not change is our framework. We first explore the constraint to determine our options, then focus on the goal (to minimize TC), and, finally, we will combine the two to find the initial optimal solution.

The Constraint

The menu of options available to the firm is given by the isoquant. It serves as the constraint because the firm is free to choose L and K on the condition that it must produce the assigned level of output. Mathematically, the equation for the constraint is simply the production function, $q = f(L, K)$.

STEP Open the Excel workbook *InputCostMin.xls*, read the *Intro* sheet, then go to the *Isoquant* sheet to see the isoquant displayed in Figure 11.1.

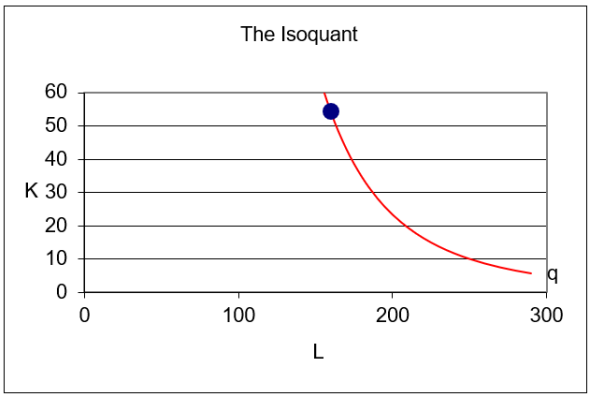


Figure 11.1: An isoquant from a Cobb-Douglas production function. *Source: InputCostMin.xls!Isoquant.*

Like the budget constraint in the Theory of Consumer Behavior gives us consumption possibilities, the isoquant gives the firm its feasible input options. All combinations below and left of the isoquant are ruled out. For example, there is no way to produce 100 units of output, holding quality and everything else constant, with the L,K combination of 100,20. The technology is simply not advanced or powerful enough to make 100 units of output with 100 hours of work and 20 machines.

The points above and to the right of the isoquant are feasible, but they are clearly wasteful. In other words, the firm could produce 100 units of output with an L,K combination of 250,50, but the isoquant tells the firm it does not need that much labor and capital to make 100 units. At 250,50, it could travel straight down to $K = 10$ and still produce $q = 100$ or straight left (on the horizontal line at $K = 50$) until it hit the isoquant and use a lot less labor. The firm could also travel in a diagonal, southwest direction until it hit the isoquant to economize on both inputs.

Points off the isoquant to the northeast (such as 250,50) are said to be *technically inefficient*. The *inefficient* part tells us that the firm is not minimizing its total cost at that point; *technical* describes the fact that the firm is not organizing its inputs so as to maximize output. In other words, the firm is not correctly solving the engineering optimization problem represented by the production function. Making 100 units of output with 250 hours of labor and 50 machines means that you are not getting the most out of your labor and capital. Economists call this situation technically inefficient.

Since the firm cannot choose a combination below the isoquant and it is wasteful to choose a combination above the isoquant, we know the answer has to lie on the isoquant.

STEP Use the scroll bar next to cell B11 to see the input mixes the firm might choose. As you change cell B11, the cell below changes also. It has a formula that computes the amount of K needed to produce the required output when you choose a value for L.

The idea is quite clear: The firm will roll around the isoquant in search of the best combination. Rolling is a good word choice and image to remember—the firm is free to choose a point high up or roll down to the bottom right. Because we do not have the input prices, we cannot find the optimal solution with the isoquant alone.

STEP Change the exogenous variables to see how the isoquant is affected. Increases in A, c, and d pull the isoquant down. That makes sense given that these shocks are all productivity enhancing and the firm will need less L and K to make the given $q = 100$.

Lowering q has the same effect, but this is not a productivity shock. You are simply telling the firm it does not have to produce as much as before so it makes sense that it can use less labor and capital.

Notice how the constraint for this input cost minimization problem is a curve, not a line like it was for the utility maximization problem. Mathematically, that does not matter much, but it will impact the graph we draw to show the initial solution.

Goal

With the constraint in hand, we are ready to model the goal. In this problem, the goal is represented by a series of *isocost* (equal cost) lines.

Total cost is $TC = wL + rK$. If we solve this equation for K (in order to graph it in L-K space), we get the equation of a line:

$$TC = wL + rK$$

$$rK = TC - wL$$

$$K = \frac{TC}{r} - \frac{w}{r}L$$

The K (or y axis) intercept is $\frac{TC}{r}$ and the slope is $-\frac{w}{r}$.

Isocosts are a little tricky at first because you are used to seeing a linear constraint and a set of indifference curves. Input cost minimization has a curved constraint and a set of linear isocosts. In the equation of the line above, TC can take on any value. Thus, there is an isocost for $TC = \$500$ and another for $TC = \$500.01$ and an isocost for every single dollar amount. Every L,K point is on an isocost and the L,K points that have the same TC are on the same isocost.

STEP Proceed to the *Isocost* sheet to see how the isocost lines are used to find the optimal solution.

Each point on a particular isocost line has the exact same total cost. So, the point on Figure 11.2 (and on your screen) has a cost of \$500 (since 2 x 190 + 3 x 40 = 500).

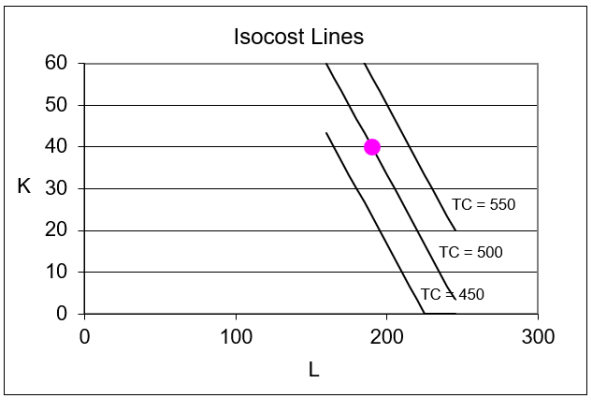


Figure 11.2: Three representative isocost lines.
Source: InputCostMin.xls!Isocost.

STEP Click the Show Goal to see how the firm's cost minimization goal is represented on this graph.

The firm can move to a new point by choosing a different combination of L and K. If the new point has the same TC of \$500 as the initial point, then it will be on the same \$500 isocost.

STEP Increase L by 30 and decrease K by 20 so you will be at another point on the same isocost line of \$500.

Now you know that all points on the TC = \$500 isocost line share the same total cost of \$500. It is also obvious that the slope of each isocost line is $-\frac{2}{3}$ since $w = 2$ and $r = 3$.

Because the firm can choose the input mix, it can choose any combination of L and K, provided that the chosen combination can produce the given amount of output. The firm wants to hire as few inputs as it can (to save on costs), but it has to meet the production target. How can it solve this problem?

The Initial Optimal Solution

We have the constraint (the isoquant) and the goal (get to the lowest isocost possible), so now we combine the two to find the optimal solution.

STEP Proceed to the *OptimalChoice* sheet.

The starting position shows an L,K combination that costs \$482.81. You can confirm this number both in cell B7 and on the chart (the middle label for the middle line).

The idea is to be on the lowest isocost line (i.e., the one with the smallest intercept) that is just touching the isoquant because that means the firm will be minimizing the total cost of producing the given level of output.

Clearly, the starting position is not optimal. You can see that the isocost is intersecting the isoquant. This information is also revealed by the slope and TRS information below the chart. The TRS, which is the slope of the isoquant at a point, is greater (in absolute value) than the slope of the isocost line at that point.

At the opening position, the firm is said to suffer from *allocative inefficiency* because it is on the isoquant, but it fails to choose the cost minimizing input mix. Because it is on the isoquant, we know it is not technically inefficient—it is using the opening combination of L and K to get the maximum output. The problem is that it is using the wrong combination of inputs in the sense that there is a cheaper way to produce the given output.

We know there are two ways to solve optimization problems: analytically and numerically. Because we have Excel and the problem implemented on the sheet, we begin with the numerical approach.

STEP Run Solver. The optimal solution is depicted by the canonical graph displayed in Figure 11.3.

Solver's answer, which is correct, has the firm choose an L,K combination whose isocost just touches the isoquant. There is no cheaper combination that can produce 100 units with the existing technology (given by the production function). If the firm went to an isocost that was one cent lower, it could not rent enough L and K to make 100 units of output.

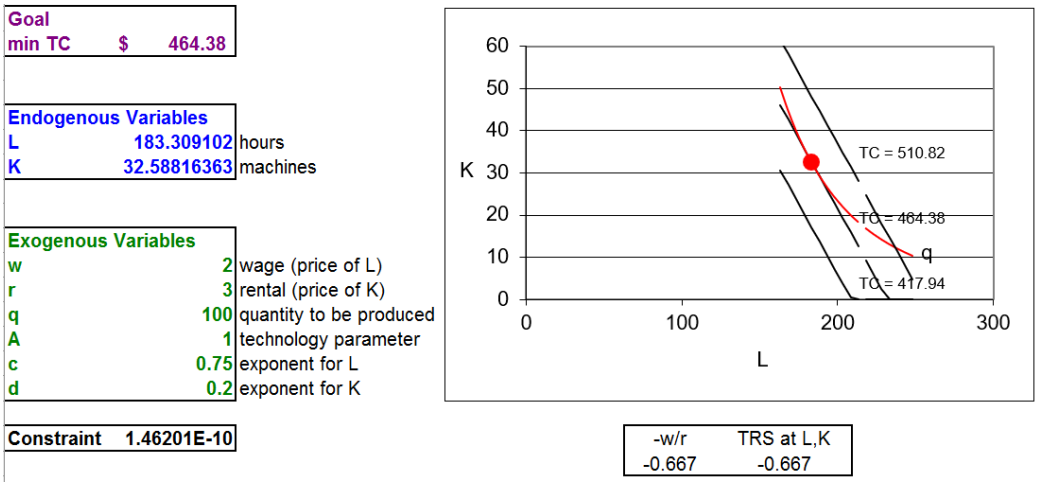


Figure 11.3: The initial optimal solution.
Source: InputCostMin.xls!OptimalChoice, after running Solver.

We can confirm Solver's result by applying the Lagrangean method to solve this constrained optimization problem.

We start by writing down the problem, using the parameter values from the *OptimalChoice* sheet.

$$\min_{L,K} TC = 2L + 3K$$
$$\text{s.t. } 100 = L^{0.75}K^{0.2}$$

The first step is to rewrite the constraint so that it is equal to zero.

$$100 - L^{0.75}K^{0.2} = 0$$

The second step is to form the Lagrangean by adding lambda, λ, times the rewritten constraint to the original objective function. We use an extra-large L for the Lagrangean function that is not at all related to the L for labor.

$$\min_{L,K,\lambda} L = 2L + 3K + \lambda(100 - L^{0.75}K^{0.2})$$

The third step to finding the optimal solution is to take the derivative of the Lagrangean with respect to each endogenous variable and set each derivative to zero (giving us the first-order conditions).

$$\frac{\partial L}{\partial L} = 2 - 0.75\lambda L^{-0.25} K^{0.2} = 0$$
$$\frac{\partial L}{\partial K} = 3 - 0.2\lambda L^{0.75} K^{-0.8} = 0$$
$$\frac{\partial L}{\partial \lambda} = 100 - L^{0.75} K^{0.2} = 0$$

The fourth, and last, step is to solve this system of equations for L^*, K^*, and λ^*. The system of three equations contains the answer, that is, the values of L and K that minimize TC. Our task is to use the equations to find these values that satisfy the three equations.

There are many ways to solve the system, but we will use the same approach that we used in the Theory of Consumer Behavior. We will reduce the system from 3 to 2 to 1 equation and unknown.

We move the terms with lambda in the first two equations to the right-hand side and then divide the first equation by the second. The Cobb-Douglas production function is easy to work with because the exponents of L and K sum to -1 and 1, respectively, when you apply the $\frac{x^a}{x^b} = x^{a-b}$ rule.

$$\frac{2}{3} = \frac{0.75\lambda L^{-0.25} K^{0.2}}{0.2\lambda L^{0.75} K^{-0.8}}$$
$$\frac{2}{3} = \frac{0.75 L^{-0.25} K^{0.2}}{0.2 L^{0.75} K^{-0.8}}$$
$$\frac{2}{3} = \frac{3.75K}{L}$$
$$L = 5.625K$$

As you can see above, this strategy cancels the lambdas and gives an expression for $L = f(K)$, which, in conjunction with the third first-order condition, reduces the system to two equations with two unknowns.

$$L = 5.625K$$
$$100 - L^{0.75} K^{0.2}$$

We substitute the expression for L into the constraint (the third first-order condition) and solve for K^*.

$$
\begin{aligned}
100 - [5.625K]^{0.75} K^{0.2} &= 0 \\
100 &= 3.6525 K^{0.75} K^{0.2} \\
27.3784^{\frac{1}{0.95}} &= (K^{0.95})^{\frac{1}{0.95}} \\
27.3784^{\frac{1}{0.95}} &= \left(K^{0.95}\right)^{\frac{1}{0.95}} \\
K^* &= 32.588
\end{aligned}
$$

Then, substituting K^* back into the expression for $L = f(K)$, we get L^*.

$$
\begin{aligned}
L &= 5.625K \\
L &= 5.625[32.588] \\
L^* &= 183.31
\end{aligned}
$$

Substituting L^* and K^* into the original objective function, we can compute the minimum cost of producing 100 units.

$$
\begin{aligned}
TC &= 2L + 3K \\
TC &= 2[183.1] + 3[32.588] \\
TC^* &= \$464.38
\end{aligned}
$$

The analytical solution agrees with Solver's answer.

The work we did in dividing the first equation by the second yields an equimarginal condition that is similar to the MRS $= \frac{p_1}{p_2}$ rule from constrained utility maximization. At the optimal solution, we have

$$\frac{2}{3} = \frac{3.75K}{L}$$

The left-hand side is the input price ratio and the right-hand side is the TRS. Thus, at the optimal solution we know that input price ratio must equal the TRS. This is a mathematical statement of the tangency we see in Figure 11.3.

If this equimarginal condition is not met, but the firm is on the isoquant (i.e., it is technically efficient), then we have allocative inefficiency. If $|TRS| > \frac{w}{r}$, then the isocost is cutting the isoquant and the firm can lower total costs by rolling down the isoquant. The reverse, of course, applies if $|TRS| < \frac{w}{r}$.

STEP If you have not done so already, double-click inside the box around cell J25 and use the scroll bar to show how the isocost and isoquant graph matches up with the TRS $= \frac{w}{r}$ equimarginal condition.

Comparing Consumer and Firm

Figure 11.3 bears a striking resemblance to the canonical graph used in the Theory of Consumer Behavior and the analytical work also contains strong similarities, but there are some critical differences between the consumer and firm optimization problems. Figure 11.4 presents a side-by-side comparison to highlight the contrasts between them.

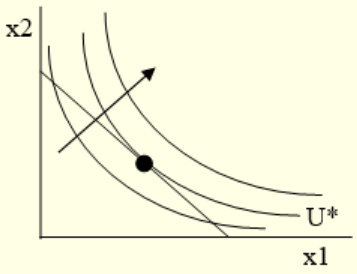
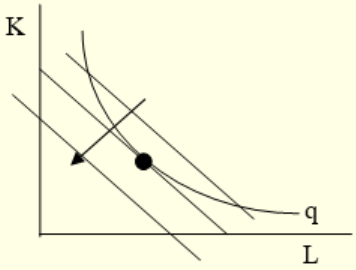
Characteristic	Theory of Consumer Behavior	Isoquant Side of the Theory of the Firm
Goal	maximize utility ($U = f(x_1, x_2)$)	minimize total cost ($TC = wL + rK$)
Canonical Graph of Initial Solution	 The line is the constraint. The curves are the goal. One line and several (representative) curves.	 The curve is the constraint. The lines are the goal. One curve and several (representative) lines.
Function Properties	Utility is a fiction that represents preferences. The actual value of the utility function has no meaning.	The production function gives q as a measurable, cardinal quantity.
Maximum Value Function	The numerical value of maximum utility (U^*) is not important.	The numerical value of minimum total cost (TC^*), measured in dollars, is the *highest* priority.
Key Comparative Statics Exercise	Demand Curve $x_1^* = f(p_1)$, ceteris paribus	Cost Function $TC^* = f(q)$, ceteris paribus
Interpreting λ^*	No real economic meaning because utility is merely ordinal.	λ^* is marginal cost, the additional cost of producing more output.

Figure 11.4: Comparing consumer and firm optimization problems.

It makes sense to use the knowledge and skills learned from the Theory of Consumer Behavior, but do not fall into a false sense of security. The input cost minimization problem has its own characteristics and terminology.

Cost Minimization is One of Three Problems

The Theory of the Firm is actually a set of three interrelated optimization problems. The initial solution to the firm's cost minimization problem focuses attention on the cheapest combination of inputs to produce a given level of output.

We can apply the same techniques we used to solve the consumer's utility maximization problem. The canonical graph is similar to the standard graph from the Theory of Consumer Behavior, but as Figure 11.4 shows, there are substantial differences between utility maximization and cost minimization.

One important similarity is the continued use of the comparison of a price ratio to the slope of a curve to determine whether the optimal solution has been found. In the case of the constrained cost minimization problem, the firm will choose that combination of inputs where TRS $= \frac{w}{r}$. If this condition is not met, the direction of the inequality (¿ or ¡) tells us which way the firm should move to find the minimum total cost.

Now that we understand the firm's cost minimization problem and have found the initial solution, we are prepared to take the next step—comparative statics analysis. The economic approach is unrelenting and monotonous. We apply the same framework to every problem. Through practice and repetition, you will learn to think like an economist.

Exercises

1. The *Q&A* sheet asks you to change r to 30 and use Solver to find the initial solution. Find the initial solution to this same problem via analytical methods and compare the two results. Are they the same? Show your work.

2. The fixed proportions production function, $q = min\{\alpha L, \beta K\}$ is analogous to the perfect complements utility functional form. Suppose $\alpha = \beta = 1$, $w = 10$, $r = 50$, and $q = 100$. Find L^*, K^*, and TC^*. Show your work. Use Word's Drawing Tools to draw a graph of the optimal solution.

3. Given the quasilinear production function, $q = \sqrt{L} + K$, and input prices $r = 2$, and $w = 5$, find the cheapest way to produce 1000 units of output. Use analytical methods and show your work.

4. Set up the problem in question 3 in Excel and use Solver to find the optimal solution. Take a screen shot of the solution on your spreadsheet and paste it into a Word document.

5. Can isoquants intersect? Explain why or why not.

References

The epigraph is from page 1044 of Joseph Schumpeter's *History of Economic Analysis* (published in 1954, shortly after his death). This classic traces the intellectual history of economics from Aristotle to the 20th century. Schumpeter reviews the theories and visions of giants like Adam Smith and Karl Marx, but also an incredible number of philosophers and economists that will not be familiar to you.

Ragnar Frisch, credited by Schumpeter with inventing the term *isoquant*, had a knack for inventing words, e.g., macroeconomics and econometrics. Luckily, "substitumal cost flexibility" did not catch on. A Norwegian, Frisch was part of an exceptionally strong quantitative and empirical tradition in Scandinavian economics that remains alive to this day.

> Several hundred years ago, an unknown inventor combined charcoal, sulfur and saltpeter and lit it afire. When the dust settled the world was changed forever.
>
> *The Story of the Gun*

11.2 The Enfield Arsenal

This chapter departs from the usual presentation style employed in this book. There is no Excel workbook associated with this application. Instead, you will be given the opportunity to answer questions and the answers are provided at the end of the chapter. Each question is highlighted by the usual *Step* marker. Try to work out each question on your own before looking at the answers.

There are four goals:

1. To understand cost minimization with isoquants and isocosts.

2. To provide an example of how theory can be applied to real-world problems.

3. To illustrate how economics can help us understand what we observe.

4. To see that economics has wide and varied application.

The inspiration and source of this application of cost minimization is from Edward Ames and Nathan Rosenberg, "The Enfield Arsenal in Theory and History," *The Economic Journal* (Vol. 78, No. 312, December, 1968), pp. 827–842, www.jstor.org/stable/2229180.

Ames and Rosenberg were economic historians and that immediately leads to a puzzler: how are economic historians different from regular historians? The answer has to do with the economic approach. Once trained as an economist, the methods and way of thinking can be applied to events and outcomes from the past. This is what Ames and Rosenberg did with the Enfield Arsenal. But before we get to that, we need to understand what rifling is all about.

Rifling

Rifles are a relatively recent innovation in firearms. Figure 11.5 shows an early version of the famous Enfield rifle with labels for the three main parts: the lock, stock, and barrel.

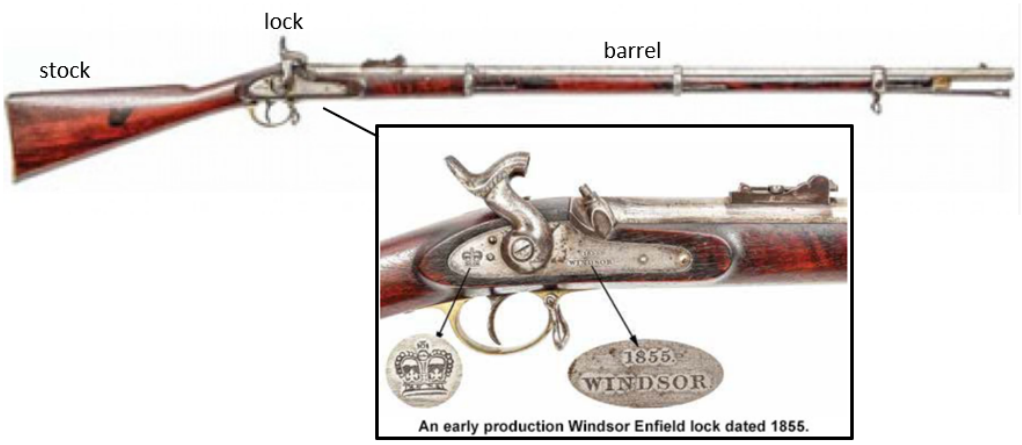


Figure 11.5: The Enfield rifle that was stored in the Enfield Arsenal. *Source:* collegehillarsenal.com/index.php?route=information/blogger&blogger_id=9

It is the barrel that distinguishes rifles from smooth-bore muskets. The barrel of a rifle has a striated pattern that spins the bullet, increasing velocity and accuracy compared with a ball from a musket.

STEP Watch this short video on rifling from *The Story of the Gun*: vimeo.com/25200729.

But the Enfield rifle was important not because it rifled, but because of how it was made.

The American System of Manufacturing

Ames and Rosenberg (p. 827) explain what the Enfield Arsenal was in the introduction to their paper:

> This paper analyses a particular historical event, the establishment of the Enfield Arsenal, in the context of the literature cited. The British Government committed itself to the construction of the Enfield Arsenal in 1854 because it wished to be able to make

> large numbers of rifles for an impending war with Russia (now known as the Crimean War). The event is important because it marked the beginning of the movement of mass-production techniques from the United States to Europe. Technical changes in gunmaking in the nineteenth century were a major source of new machine techniques; and industrialisation in the nineteenth century is overwhelmingly the history of the spread of machine making and machine using.

So an arsenal is an armory, a warehouse of guns and ammunition. Enfield is a place in England and the Enfield Arsenal is literally a building constructed by the British government in 1854 that would be used to store rifles made with mass-production techniques.

The Enfield Arsenal was special because it was the first time the British would use mass-production techniques to make weapons. Up to this point, the British had made guns the old-fashioned way—by hand in the small shops of thousands of skilled artisans in the area around Birmingham. The stock was carefully carved by an experienced craftsman who fitted the stock with the lock and barrel. It was like a tailor making a bespoke suit—each rifle was one of a kind. A work of art.

Ames and Rosenberg (p. 832, footnotes omitted) point out that making the stock by hand was especially slow and expensive to do:

> The gunstock was one of the most serious bottlenecks in firearms production. In England, at the time of the Parliamentary hearings, out of about 7,300 workmen in the Birmingham gun trade, the number of men employed in making gunstocks totalled perhaps as many as 2,000. Its highly irregular shape for long seemed to defy mechanical assistance, and the hand-shaping of the stock was a very tedious operation. Furthermore, the fitting and recessing of the stock so that it would properly accommodate the lock and barrel were extremely time-consuming processes, the proper performance of which required considerable experience. With Birmingham methods, it required 75 men to produce 100 stocks per day. Using the early (1818) version of the Blanchard lathe, 17 men could produce 100 stocks per day.

This quotation requires some explanation. First, the reason for the Parliamentary hearings was that British politicians were angry with the Birmingham gunsmiths for not adopting fast, efficient mass-production techniques.

There was an investigation and testimony was given. How could upstart Americans have better technology than the British, a nation that dominated the entire globe? It was a national embarrassment!

Second, the quotation mentions the *Blanchard lathe*. This is a machine that cuts and shapes wood (and other materials like metal), but it is easier to understand if you see it.

STEP Watch this video, vimeo.com/25200825, to understand how a lathe works and how the production of precision parts makes Diderot's dream come true.

The video explained that the new country of the United States of America needed weapons so the Springfield Armory was built in 1794 in Springfield, Massachusetts. At first, stocks were made by hand, just like in Birmingham. They were then individually fitted to each rifle.

But in 1818, the Blanchard lathe burst on the scene. The narrator, echoing the British Parliamentary hearings, says, "Prior to the Blanchard Lathe, it took one to two days to make a rifle stock by hand. Now, a twelve-year-old boy could turn out a dozen stocks in a single day."

The Blanchard lathe enabled a reorganization of the production process. In factories in the northeast, the United States began to use mass-production techniques to make rifles and pistols (and then bicycles, sewing machines, typewriters, and so on). This is the American system of manufacturing. A key element is that a machine can make a precision part so many almost identical parts can be made and then the product is assembled.

The video points out that the history of gun-making is closely tied to the rise of mass-production techniques and precision manufacturing. In the video, William Ruger cites an idea from French philosopher Denis Diderot (1713–1784). Ruger says Diderot's theory at that time was that "It would be possible to make all of the individual parts alike and then at the last minute assemble them, rather than fitting them together as you went, which was the customary thing up to that time."

Adam Smith (1723 - 1790) was a contemporary of Diderot. For Smith, *the division of labor* explained the explosion in productivity that he saw all around him as the Industrial Revolution began.

Breaking production into a series of steps and then assembling the parts enables many more units of output to be produced. This is called the division of labor. Smith emphasized several reasons for the greater productivity enabled by the specialization of labor:

1. Practice makes perfect: focusing on a single task makes you very good at it.

2. Saves time: no need to set things up when you move to a new task.

3. Innovation: adjustments are made by workers who are expert in a particular task.

Machines such as the Blanchard lathe feed into the division of labor by enabling much finer specialization. For rifles, production with a lathe meant that they were no longer one of a kind. They were all alike and could be easily connected to the lock and barrel to make a rifle.

By applying Diderot's theory of assembling perfectly fitted parts and Smith's division of labor, the Springfield Armory was able to enjoy a huge increase in productivity compared with Birmingham methods.

So now you know exactly what a lathe is and how mass production played a key role in the exponential increases in productivity during the Industrial Revolution, but there is one more important advantage to mass production. Let's see if you can figure it out.

STEP What are the tremendous advantages of interchangeable parts in a rifle (or anything else for that matter) for the end user?

The answer is at the end of this section, but take a few minutes to think about the question. What advantage would soldiers using rifles that were all alike have over enemies using individually made rifles?

Two Big Questions

The key date in this story is 1854. Until this time, the British used Birmingham methods, which means an experienced craftsman made each entire gun by hand. They shaped the stock, then attached it to the lock and the barrel. Each part was slightly different and could not be easily replaced if damaged.

Beginning in 1854, rifles produced for the Enfield Arsenal, however, were made with interchangeable parts (including stocks made on lathes) that could be put together in an assembly line. Once in use, broken parts could be removed and new ones snapped on.

Ames and Rosenberg (pp. 839–840, footnotes omitted) sum up the situation:

> As of 1785, neither the British nor the Americans could make guns with interchangeable parts. As of 1815, Americans could make guns with interchangeable metal parts, but could not make interchangeable gunstocks. As of 1820, they could make interchangeable gunstocks. At any date, presumably, they could use not only current methods but earlier methods which these had displaced.

The United States had been mass-producing guns with interchangeable parts since 1815. The British waited until 1854 to use the superior, mass-production techniques. This gives rise to two big questions:

1. Why did the British wait so long to use mass-production techniques to make rifles with interchangeable parts?
2. Why did the British switch to mass-production techniques in 1854?

1. Why Did the British Wait so Long?

A possible answer to the puzzle of why British gunsmiths did not adopt the new technology is that the British did not know about the Blanchard lathe so that is why they did not use it?

STEP Is lack of knowledge about American technology a good answer? Why or why not?

Another possible answer is poor management. Maybe British rifle manufacturers were lazy, stupid, and careless? The right answer—adopt mass-production techniques—was staring them in the face and they ignored it.

STEP Is managerial failure a good answer? Why or why not?

Economic historians give a third answer to why the British did not adopt the Blanchard lathe. They use the economic way of thinking. They look for differences in the environment that would lead to different optimal solutions.

In other words, Ames and Rosenberg stop searching for why the British made a mistake and accept the fact that their refusal to adopt mass-production techniques was actually smart and correct. They look for reasons that justify the British decision to reject the Blanchard lathe.

This is crazy, right? It is obvious that mass production is better. Well, it turns out that there are two critical differences between the United States and Britain in the first half of the 19th century that play an important role in deciding how to make rifles.

First, the two countries had quite different labor forces. The British had a cohort of skilled rifle craftsmen and the United States did not. As the Parliamentary hearings noted, there were several thousand skilled craftsmen in Birmingham making stocks and rifles. The United States was a young country with mostly unskilled, male workers. Few skilled craftsmen would emigrate to the United States since they had good, high paying jobs at home.

These supply and demand differences meant that, in the United States, wages for skilled craftsmen were much higher than in Britain, and wages for unskilled labor were lower.

Second, wood was plentiful and cheap in the United States, but it was much more expensive in Britain. Ames and Rosenberg offer the following footnote (p. 831) to help explain why wood plays a critical role:

> Report of the Small Arms Committee, op. cit., Q. 7273-81 and Q. 7520-7521; G. L. Molesworth, "On the Conversion of Wood by Machinery," Proceedings of the Institution of Civil Engineers, Vol. XVII, pp. 22, 45-6. In the discussion which followed Mr. Moleworth's paper Mr. Worssam, a prominent English dealer in woodworking machinery, made some interesting comparative observations which were summarised as follows: "He had seen American machines in operation, and he found that, although they might be adapted for the description of work required in that country, they were not so suitable for English work, in which latter high finish and economy of material were of greatest importance. In America the saws were much thicker than those used in the English saw-mills, so that they consumed more power, wasted more material, and did not cut so clean, or so true, though there was less care required in working them" (ibid., pp. 45-6).

A key point in this long quotation is that American saws (and, of course, lathes) "wasted more material." A British skilled craftsman making stocks from lumber would be careful to "economize" on the material. In America, a 12-year old boy working with a lathe (a dangerous job) would not care at all about wasting wood.

The different endowments of wood in the two countries meant that the Blanchard lathe was much more expensive to operate in Britain than in America.

Now that we know how the United States and Britain differed with respect to (1) wages for skilled and unskilled labor and (2) operating costs for the Blanchard lathe, we are ready to make the case for the economic explanation for why the British waited so long to adopt mass-production techniques.

As is typical in economics, the exposition will rely on graphs. But instead of just reading the explanation, you will try to do it yourself first. The idea is to apply the input cost minimization problem to this scenario. You can, of course, simply jump to the end of the section to see the answers, but you will learn much more if you try to do it yourself first. Follow the instructions and hints offered below and see how close you get. Make sure you understand where you made a mistake or in what ways you were confused.

STEP Draw graphs that show how the different resource endowments and input prices affected the optimal input mix. Use the detailed instructions that follow as a guide. How do the graphs explain why the British waited so long to adopt mass-production techniques?

We will use two sets of two graphs. The first set of two graphs will be for the labor force difference between the United States and Britain. The second set shows the effect of the different endowments of wood.

Begin by drawing a graph representing the British situation in 1820 with respect to using skilled and unskilled labor to make, for example, an order of 1,000 rifles. It should have skilled labor on the y axis and unskilled labor on the x axis. Draw in an isoquant (representing the combinations of skilled and unskilled labor that would make the requested 1,000 rifles).

Draw another graph, next to the first one, that is exactly the same. Your second graph represents the United States' options for making 1,000 rifles in 1820. The fact that both isoquants are the same means that the two countries had access to the same technology and are making the same product.

Next, you need to draw the isocost lines. This is where the difference in labor force comes into play. We know the British have skilled labor and the United States does not—immigrants to the United States were not typically experienced, educated workers, but young, unskilled males. That means the price of skilled labor is much higher in the United States. How is that reflected in the isocosts for your two graphs?

The second set of graphs uses L and K as the inputs. As before, draw a pair of graphs side by side, one for the British and the other for the United States, with machinery on the y axis and labor on the x axis. Include the isoquants. Once again, the isoquants are the same, meaning that the British were aware of and could have used American methods.

The key to the economic explanation for why the British did not do what the Americans were doing lies in the isocosts. Remember that early versions of the Blanchard lathe used a lot of wood and this increases the price of machinery. If r is much higher in Britain than in the United States, how does this affect the isocosts?

Take a moment to look at your two sets of graphs. How can they be used to explain why the British rejected mass production before 1854?

Proceed to the end of this section to check your graphs and answers.

2. Why Did the British Switch in 1854?

The second big question revolves around the British decision to switch in 1854 and mass produce the Enfield rifle. Why did they do this? Why did they abandon their decades-old system of production centered in Birmingham, with a network of many small artisans and smiths that made firearms to individual order or in small batches?

Our first possible answer matches up with the lack of knowledge answer to the first big question. Maybe, in 1854, the British heard that mass-production techniques utilizing the Blanchard lathe were available and immediately moved to adopt the new production methods?

STEP Is sudden awareness of new American technology a good answer? Why or why not?

The second possible answer, like before, relies on management. Maybe they wised up? What if British firearms manufacturers recovered from their slumber and moved quickly to modernize their industry?

STEP Is managerial improvement a good answer for the switch? Why or why not?

You probably got the first two right, but the third one is harder. It might be easy in general terms, but getting the details can be complicated.

The third answer is based on economic reasoning. This means that when we see changes in behavior, we look for changes in the environment. We do not search for events or causes that changed a mistake into the right answer. Instead, we accept that the answer to not use mass production was correct for, say, 1830, but the new optimal solution, in 1854, was to switch to the American system.

This is a key aspect of the economic approach, and it can be challenging to grasp. Our instinct when we see something change is to think of correction or improvement. Economists do not think this way. We see optimization everywhere so if something changes, it was optimal before and it has moved to a new optimal solution because of an exogenous shock.

The search is on for shocks that switch the correct answer from "reject" to "accept" interchangeable parts.

There are two ways in which Britain before 1854 differed from Britain after 1854 and these two ways impacted wages and the operating cost of machinery. These changes act as shocks on the input cost minimization problem and produce a new optimal solution. We first have to figure out the shocks, then we can see how they affect the optimal solution.

STEP Answer these two questions:

1. What happened to the British labor force?
2. What happened to the Blanchard lathe?

You may not be an expert on British labor in the 19^{th} century or know anything about the Blanchard lathe, but you can think about what might have happened. Try to come up with a hypothesis. Think of recent changes in the

labor force that you have heard about, especially those driven by technology (e.g., driverless cars and trucks). Think about how machines, computers, and technology in general have changed over time.

After checking your answer at the end of this section, so now you know what happened, you are ready to draw graphs that illustrate the economic historian's explanation for the British switch in production technique.

STEP Draw graphs that show how the changes mentioned affected the optimal input mix. How do the graphs explain why the British switched to mass-production techniques in 1854?

Draw two pairs of graphs just like before (unskilled and skilled labor on one and machinery and labor on the other), but this time we are comparing environments before and after 1854 in Britain (the United States has nothing to do with this). First, compare the optimal mix of unskilled and skilled labor for Britain in 1820 versus 1854. Remember that the skilled craftsmen died and were not replaced so the skilled wage rate rose. How does this make the 1854 graph different from the 1820 graph?

In the second set of two graphs, with machinery and labor on the axes, we know that machinery got better and better (wasting less and less wood) over time so r fell. What will this shock do to the isocost lines?

Check out the suggested answers at the end when you are finished. Take the time to debug any mistakes. Make sure you understand how the isocost lines shift and how the comparison of two graphs yields answers to the questions.

Evaluating this Application

At the beginning, we had four goals:

1. To understand cost minimization with isoquants and isocosts.

2. To provide an example of how theory can be applied to real-world problems.

3. To illustrate how economics can help us understand what we observe.

4. To see that economics has wide and varied application.

You decide to what extent the goals were met. At the very least, you learned a little about American manufacturing in the 19$^{\text{th}}$ century and rifles (including where the phrase "lock, stock, and barrel" comes from).

The application should help you understand the conventional isoquant–isocost graph and the firm's input cost minimization problem. Remember that the higher the price of the input on the x axis or lower the price of the input on the y axis, the steeper the isocosts.

But the real deep learning and big picture idea concerns how economists view the world. This is called economic reasoning or the economic approach. We did "an economic analysis of the Enfield Arsenal."

The idea is that economics is not a discipline organized around content (the stock market or money, for example), but a way of thinking. Economists often interpret observed behaviors as optimal solutions to optimization problems and they see change as driven by a shock that takes us from one optimal solution to another.

Thinking like an economist is difficult and sometimes counter-intuitive, but it can provide an interesting perspective on the world. Certainly, Ames and Rosenberg gave us a novel view of the issues surrounding the Enfield Arsenal.

Exercises

1. Explain why the endowment of wood affects the price of machinery used in producing rifles in the 19th century.

2. What could have caused the British to switch to mass-production techniques before 1854? Give a concrete example.

3. If the British had used the Blanchard lathe in 1820, then that would have been allocatively inefficient. Draw a graph that shows this and explain what it means.

4. Ames and Rosenberg (p. 836) include additional differences between America and Britain, such as the fact that the British consumer liked fancier gunstocks:

   > American machine processes could not produce guns of the kind favoured by English civilians. The Blanchard lathe produced stocks of a standard size, whereas English buyers did

> not want standard gunstocks. The English methods were suited to catering to the idiosyncratic needs of individual users.

How would this information change the comparison of the isoquant–isocost graph in the two countries?

References

The epigraph is on the description of *The Story of the Gun's* DVD set, available online and in many libraries. This entertaining video mixes the history of firearms with military history and technological change.

Ames and Rosenberg's article is an excellent example of economic history. The Cliometric Society is online at www.eh.net/Clio. In Greek mythology, Clio is the muse of history. Cliometricians use economic theory and econometrics to analyze economic history.

Denis Diderot's encyclopedia is available online. The Catholic Church banned it because it was too skeptical about Biblical miracles.

Adam Smith's discussion of the division of labor and his famous pin factory example is in the first book of *The Wealth of Nations*, at www.econlib.org/. Although it seems obvious today, specialization as a way to increase productivity was not so clear and Smith argued it drove the mighty economic engine he saw springing to life.

Appendix: Suggested Answers

STEP What are the tremendous advantages of interchangeable parts in a rifle (or anything else for that matter) for the end user?

Fixing broken rifles! You can quickly repair a mass-produced rifle if one of its pieces (lock, stock, or barrel) breaks. A rifle built by hand is useless once one of its individual parts fails. You would need a skilled craftsman to fix it.

On a battlefield, you could cobble together parts from different broken units to create operating weapons. And anyone could do this—they would not have to be a skilled craftsman.

In general, with precision parts, if the product breaks, you can buy a replacement part to repair the product. With bespoke items, you need an expert to adjust and refit to repair it.

STEP Is lack of knowledge about American technology a good answer? Why or why not?

This is a ridiculous explanation. Granted there is an ocean, but given the common language and communication, this answer makes no sense. In fact, there is lots of evidence that the British knew all about the American methods. They simply chose not to use them.

STEP Is managerial failure a good answer? Why or why not?

Like lack of knowledge, this is not a very satisfying answer. There is no reason to believe these specific people were especially poor managers. Economists are wary of this type of answer. Self-interested agents who respond to incentives are unlikely to make bad decisions, especially with great sums of money and lives at stake.

There is a subtle point to be made here that separates economists and non-economists. The latter are much more likely to accept mistake and stupidity to explain an observed decision or behavior that turned out badly. Economists tend to stick with rational, optimizing agents and explain bad choices as a result of lack of information or differing objectives.

STEP Draw graphs that show how the different resource endowments and input prices affected the optimal input mix. Use the detailed instructions that follow as a guide. How do the graphs answer explain why the British waited so long to adopt mass-production techniques?

The isoquant is exactly the same in each graph in Figure 11.6. US skilled labor wages were very high because there were few experienced craftsmen migrating to the United States, but lots of young, unskilled workers. The slope of each isocost is the input price ratio, $-\frac{w_{Unskilled}}{w_{Skilled}}$. Thus, the US isocost lines are flatter than Britain's. This leads to a different cost-minimizing input mix.

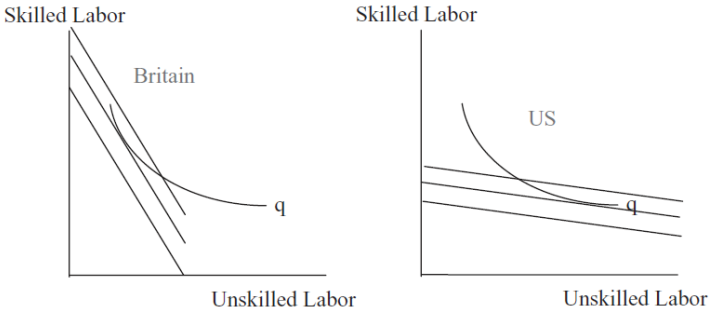


Figure 11.6: The effect of different wages for labor.

The price of machinery includes the cost of wood use just like a car's operating cost includes the cost of gasoline. The early versions of the Blanchard Lathe were quite wasteful, but this did not matter in the heavily forested United States. In Britain, however, wood was expensive. The British Isles were mostly deforested by then. This makes the isocost lines steeper in Figure 11.7 for Britain. Once again, factor prices help determine the input mix.

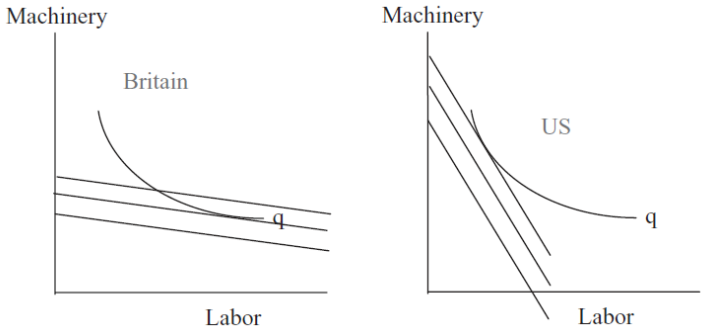


Figure 11.7: The effect of different prices for machinery.

So how do these graphs explain how economists view this historical episode? Varying resource endowments mean that each country faces its own set of input prices, which in turn lead to different cost-minimizing solutions. For the United States, unskilled labor with the Blanchard lathe was the cheapest way to make rifles. Not so for the British. At that time and place, with the skilled craftsmen and lack of cheap wood, rejecting mass production was the optimal decision.

In fact, the economic approach says something even more outlandish. Had the British used mass production for rifles before 1854, that would have been a mistake! Take the US tangency point and transfer it to the British graph in Figures 11.6 and 11.7. Producing with the US input mix is allocatively inefficient for Britain—that is, the British would not be minimizing cost.

Economists have no problem with agents making different choices. This does not mean that one is right and the other is wrong. All it means is that they face different prices. They are both optimizing. That is a difficult idea to wrap your head around. Ponder it.

STEP Is sudden awareness of new American technology a good answer? Why or why not?

This answer makes little sense. American and British people and entrepreneurs moved freely across the Atlantic and were well aware of production methods in each country. The claim that a new technique was suddenly made known to the British is absurd.

STEP Is managerial improvement a good answer for the switch? Why or why not?

This answer is pretty silly. To be credible, it requires an explanation for the sudden change from stupid, lazy, and careless producers of firearms to smart, energetic, and focused ones. There is no evidence of an explosion in managerial aptitude or a burst in managerial education. For this argument to be convincing, we will need a lot more evidence on British management prowess and how it changed over time.

STEP Answer these two questions:

1. What happened to the British labor force?

2. What happened to the Blanchard lathe?

The British labor force underwent a profound structural adjustment. The skilled craftsmen in the Birmingham gun trade died off and were not replaced. No skilled gunstock maker would suggest that his son follow him into the trade. They could see the writing on the wall—the machines were taking over. As the supply of these workers dwindled, the wages of skilled rifle artisans in Birmingham rose.

Perhaps more important is the second shock. The Blanchard lathe was continually improved over time; more modern versions of the lathe wasted a lot less wood. Today, a lathe uses a laser sight to precisely cut the wood. No human could possibly compete with it.

As the lathe wasted less wood, the operating cost of machinery fell. This is a nice example of how the price of an input can represent more than simply the out-of-pocket cost paid for the input. In this example, the price of a lathe is not simply the price paid for the machine itself; it includes the price of the wood used.

So, the shocks to the input cost minimization problem are that the skilled labor wage rose relative to the unskilled and r fell relative to w.

Notice how we first figure out what happened and then we model it. That is, we incorporate the story into one of the variables. In this case, the changing labor force increases the wage of skilled labor and the improving Blanchard lathe decreases r.

STEP Draw graphs that show how the changes mentioned affected the optimal input mix. How do the graphs explain why the British switched to mass-production techniques in 1854?

A high price of skilled labor makes the isocost lines flat (the slope falls in absolute value because the denominator increases). This leads to a more unskilled-labor intensive optimal input mix. As skilled craftsmen disap-

peared and their wages rose, there was greater incentive to use unskilled labor. Notice how the comparison in Figure 11.8 is across time periods.

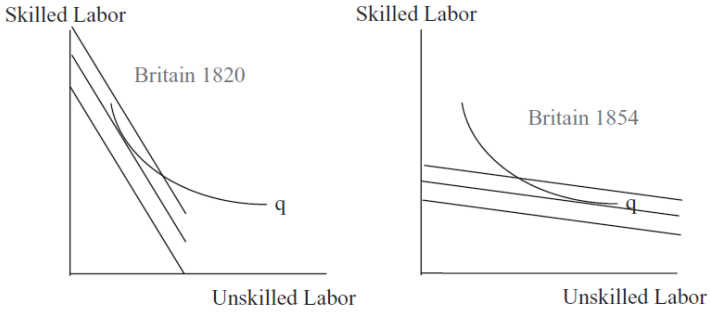


Figure 11.8: The effect of changes in the British labor force.

The price of machinery fell and fell as machines got better and better, making the isocost lines steeper and steeper (r is in the denominator) as shown in Figure 11.9, and leading to the adoption of mass-production techniques in Britain—the Enfield Arsenal was born.

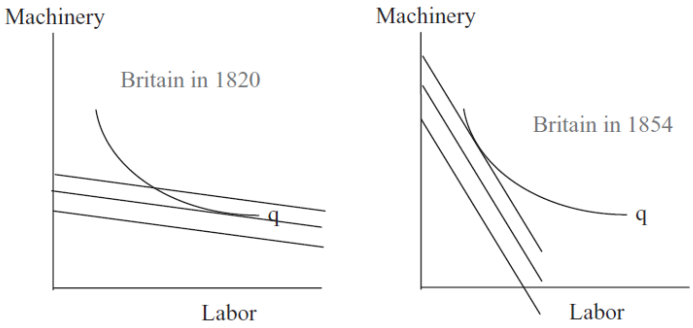


Figure 11.9: The effect of improvement in the Blanchard lathe.

Notice how the Britain in 1854 graphs in Figures 11.8 and 11.9 are the same as the US graphs in Figures 11.6 and 11.7. This shows that when Britain faced the same input prices as the United States, they made the same, optimal decisions.

There are reasons to hope that another type of production function, more diversified than Douglas's, may soon be available, and from these it would be possible to derive cost functions typical for particular industries.

Hans Staehle

11.3 Deriving the Cost Function

We have solved the input cost minimization problem so the next task is comparative statics analysis. We will focus on shocking q (the quantity the firm must produce) and track minimum total cost. The relationship between TC^* and q is called the *cost function*.

The novelty here is that we are not interested in how the optimal values of the endogenous variables, L and K vary as we shock q. Instead, we focus on the objective function, minimum total cost, and how it changes as q changes.

Another important aspect of comparative statics analysis for the input cost minimization problem is that, unlike utility in the Theory of Consumer Behavior, total cost can be cardinally measured. We can compare the total costs of different firms and perform arithmetic on total cost. If the minimum TC for $q = 10$ is \$40 and it rises to \$45 when $q = 11$, we can say TC increased by \$5. Because TC is cardinal, we will be able to interpret and use the Lagrangean multiplier.

As usual, we will explore both ways to do comparative statics:

- Numerical methods using a computer: Excel's Solver and the Comparative Statics Wizard.
- Analytical methods using algebra and calculus: conventional paper and pencil.

Numerical Methods to Derive the Cost Function

STEP Open the Excel workbook *DerivingCostFunction.xls*, read the *Intro* sheet, and proceed to the *OptimalChoice* sheet.

The organization is the same as in the *InputCostMin.xls* workbook. The cost-minimizing way of producing 100 units of output is to use about 183.3 hours of labor with 32.6 machines, which costs $464.38. There is no other combination of L and K that makes 100 units at a lower cost.

What happens if the firm needs to produce more, say, 110 units of output?

STEP Change cell B18 to 110.

The chart updates, showing a new (red) isoquant. The initial combination is not a viable option because it cannot produce 110 units. The firm has to re-optimize.

STEP Run Solver to find the new optimal solution.

The cost-minimizing amounts of labor and capital increase to produce the higher output required and the minimum total cost is now $513.39. We are looking for the minimum total cost. We want to know the cheapest way of producing any given output. This is called the *cost function*.

We can show the comparative statics analysis on the isoquant-isocost graph or on a presentation graph where we plot $TC^* = f(q)$, ceteris paribus. If we connected the points of tangency of isoquants and isocosts, we would get the *least cost expansion path* (LCEP).

Our work thus far has revealed two points on the LCEP and cost function: when $q = 100$, $TC = \$464.38$ and when $q = 110$, $TC = \$513.39$. Let's use the Comparative Statics Wizard to get more data so we can draw the LCEP and cost functions and understand how they are related.

STEP Return cell B18 to 100, then run the Comparative Statics Wizard, applying 10 q shocks in increments of 10.

The *CS1* sheet shows what your results should look like. The *CS1* sheet includes two graphs, the isoquant-isocost graph with the least cost expansion path and the cost function, as shown in Figure 11.10.

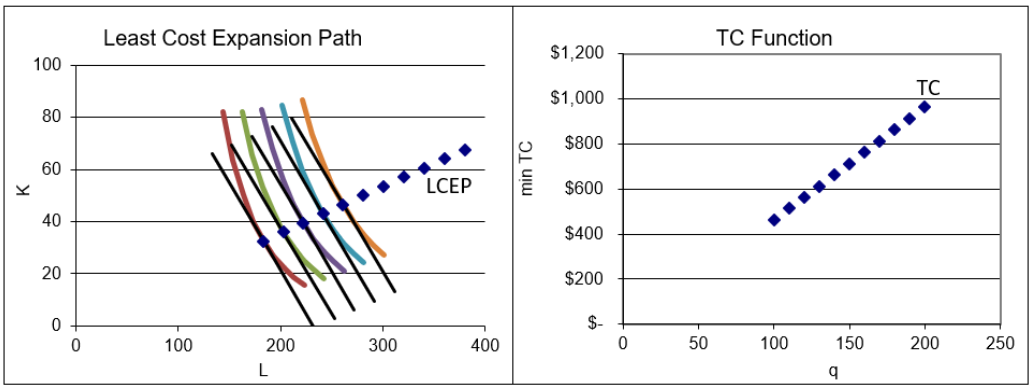


Figure 11.10: Deriving the cost function.
Source: DerivingCostFunction.xls!CS1

Figure 11.10 should remind of you of other graphs we have drawn, such as Engel and demand curves. On the left, using the display of the optimal solution to the input cost minimization problem, we show how different q produce a set of tangency points that comprise the LCEP.

On the right in Figure 11.10, we show only the minimum cost of producing each level of q, and hide everything else. This allows us to highlight the relationship between TC and q.

The two graphs in Figure 11.10 make clear that the source of the cost function is the optimal solution of the cost minimization problem as q varies. Just like demand curves do not come out of thin air, but are derived from utility maximization, cost functions are derived from input cost minimization.

We are interested in the shape of the cost function. It looks like a line, but is it really linear? To find out, we can see if it has a constant slope. If the slope is changing, we know the function is not linear.

STEP In your *CS* sheet, find the slope at different points on the function by computing the change in TC divided by the change in q.

Click the Unhide Column E button (near cell C9 in the *CS1* sheet) if you are stuck or to check your work. It is clear that the slope changes as output changes. This means that the cost function is nonlinear.

Analytical Methods to Derive the Cost Function

We can use the Lagrangean method to find $TC^* = f(q)$. We will leave q as a letter instead of a number so that the reduced-form solution will include q. Then we can plug in any value of q to find minimum cost for that q and easily draw a graph of the cost function.

The solution closely follows the work we did at the beginning of this chapter, but we proceed step-by-step to practice and reinforce the Lagrangean method.

The problem is

$$\begin{aligned} \min_{L,K} TC &= 2L + 3K \\ \text{s.t. } q &= L^{0.75}K^{0.2} \end{aligned}$$

The first step is to rewrite the constraint so that it is equal to zero.

$$q - L^{0.75}K^{0.2} = 0$$

The second step is to form the Lagrangean by adding lambda, λ, times the rewritten constraint to the original objective function. We use an extra-large $\mathcal{L}$ for the Lagrangean function that is not at all related to the L for labor.

$$\min_{L,K,\lambda} \mathcal{L} = 2L + 3K + \lambda(q - L^{0.75}K^{0.2})$$

The third step to finding the optimal solution is to take the derivative of the Lagrangean with respect to each endogenous variable and set each derivative to zero (giving us the first-order conditions).

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial L} &= 2 - 0.75\lambda L^{-0.25}K^{0.2} = 0 \\ \frac{\partial \mathcal{L}}{\partial K} &= 3 - 0.2\lambda L^{0.75}K^{-0.8} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= q - L^{0.75}K^{0.2} = 0 \end{aligned}$$

The fourth, and last, step is to solve this system of equations for L^*, K^*, and λ^*. We move the terms with lambda in the first two equations to the right-hand side and then divide the first equation by the second. The exponents

cancel nicely (see section 11.1) and we get $L = 5.625K$. This is not a reduced-form solution because L is not a function of exogenous variables alone. We substitute this expression for L into the third first-order condition to get optimal K and then optimal L as shown below.

$$
\begin{aligned}
&q - [5.625K]^{0.75} K^{0.2} = 0 \\
&q = 3.6525 K^{0.75} K^{0.2} \\
&\frac{q}{3.6525} = K^{0.95} \\
&\left[\frac{q}{3.6525}\right]^{\frac{1}{0.95}} = (K^{0.95})^{\frac{1}{0.95}} \\
&K^* = 0.25574 q^{\frac{1}{0.95}} \Rightarrow L^* = 1.43854 q^{\frac{1}{0.95}}
\end{aligned}
$$

Finally, we substitute the optimal solutions for L^* and K^* into the original objective function.

$$
\begin{aligned}
&TC = wL + rK \\
&TC^* = 2\left[1.43854 q^{\frac{1}{0.95}}\right] + 3\left[0.25574 q^{\frac{1}{0.95}}\right] \\
&TC^* = 2.877 q^{\frac{1}{0.95}} + 0.767 q^{\frac{1}{0.95}} \\
&TC^* = 3.644 q^{\frac{1}{0.95}}
\end{aligned}
$$

This expression is the total cost function. It gives the cheapest cost of producing any given amount of output. If $q = 100$, $TC = \$464.38$. Not surprisingly, this agrees with our results using numerical methods.

Notice also that the cost function is clearly nonlinear. It is increasing at an increasing rate because the exponent on q is greater than one ($\frac{1}{0.95} \approx 1.05$). The derivative of TC with respect to q, the slope, is not constant because it depends on q. If the exponent was exactly 1, then the slope would be constant and the TC would be a line. The fact that this exponent is only slightly greater than one explains why TC looks almost linear in Figure 11.10.

Interpreting Points Off the Cost Function

When we derived the demand curve from the "maximize utility subject to a budget constraint" optimization problem, we explored what it meant to be off the demand curve (see Figure 4.12). We learned that points to the left or right of the inverse demand curve (with price on the y axis) mean that

the consumer is not optimizing, i.e., the consumer is not choosing a point of tangency between the indifference curve and budget constraint.

We can conduct the same kind of inquiry here, asking this question: What does it mean to be off the cost function?

Unlike the inverse demand curve, where the exogenous variable is on the y axis, the cost function is graphed according to usual mathematical convention, with the exogenous variable, output, on the x axis. Thus, points off the curve are interpreted vertically above or below the cost function.

What does it mean if a point is above the cost curve? Figure 11.11 helps us answer this question. On the left is the familiar isoquant/isocost graph. The cheapest way to produce *q0* units of output is with the L and K combination at the point labeled TC^*. The graph on the right of Figure 11.11 shows that TC^* is the point on the cost function at an output of *q0*.

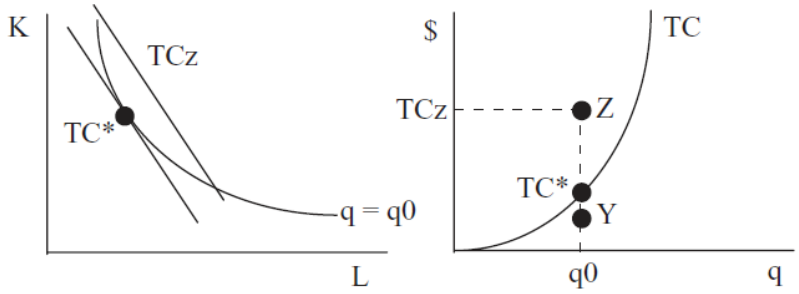


Figure 11.11: Understanding points off the cost function.

Point Z, a point above the cost function, reveals that the firm is producing the level of output *q0* at a total cost *above* the minimum total cost. This means that the firm is choosing an input mix that is not cost minimizing. Point Z on the graph on the left of Figure 11.11 must lie on an isocost above the tangent isocost. We do not know exactly where point Z is on the graph on the left (so we do not know if there is technical or allocative inefficiency), but we do know it has to be somewhere on the isocost labeled *TCz* that has a total cost the same as the cost of producing point Z (on the graph on the right).

Point Y on the right side of Figure 11.11 is below the cost function. How can this point be generated by the graph on the left? It cannot. There is an isocost with a total cost equal to that at point Y, but it is below the iso-

quant and, therefore, unattainable. In other words, point Y does not actually exist. The firm cannot produce *q0* units of output at any cost less than TC^*.

Another way of thinking about TC geometrically, is that there are points above TC, but only empty space below it. Sure, on a printed page, chalkboard, or computer screen, there is white space above and below TC and you can write on it (just like point Y in Figure 11.11), but this is misleading. In fact, below TC there is nothing, total void. If you tried to put a point there, your hand would go through the paper!

This has implications beyond pure theory. The fact that there are no points below the cost function means that we should never fit a line through a cloud of points to estimate a cost function. Instead of a least squares approach to estimating a cost function, estimation techniques in the stochastic frontier literature are based on fitting a curve around the observed points, as in Figure 11.12.

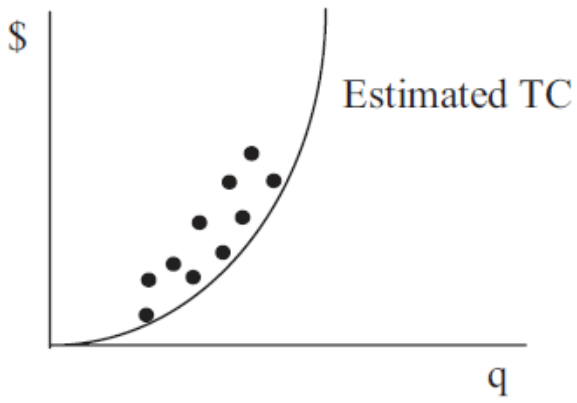


Figure 11.12: Estimating a cost function.

Shifts in the Cost Function

You learned in Introductory Economics that price causes a movement along a demand curve, but other shocks (like increasing income) change demand causing the entire curve to shift. The same thing happens with the cost function. Changing q leads to moving along the TC function, but other exogenous variables cause shifts in the cost function.

STEP Proceed to the *CostFn* sheet.

The sheet displays a cost function charted from the data above it. The data in columns L and M are actually formulas for the reduced-form expressions for L^* and K^*. Column N has the minimum total cost for the benchmark problem and will not change because the cells are merely numbers (so it is labeled "Dead (Initial)"). Column O, however, has the reduced-form expression for TC^* and will update if any of the underlying parameters are changed (hence the "Live" label).

STEP Click on a few cells in columns L, M, N, and O to see the formulas and values.

The general versions of the reduced forms for the Cobb-Douglas production functions are provided and entered in cells. The expressions look daunting (and they are tedious to derive), but the derivation is straightforward: leave every exogenous variable as a letter and find the optimal solution for L,K, λ, and total cost.

Initially, N and O are the same because the exogenous variable values have not been changed yet. Let's do that now.

STEP Change cell B20, the exponent on L, to 0.8.

Your screen looks like Figure 11.13. The increase in labor productivity has shifted down the total cost curve. This makes sense. The increase in c has made it cheaper to produce any given output.

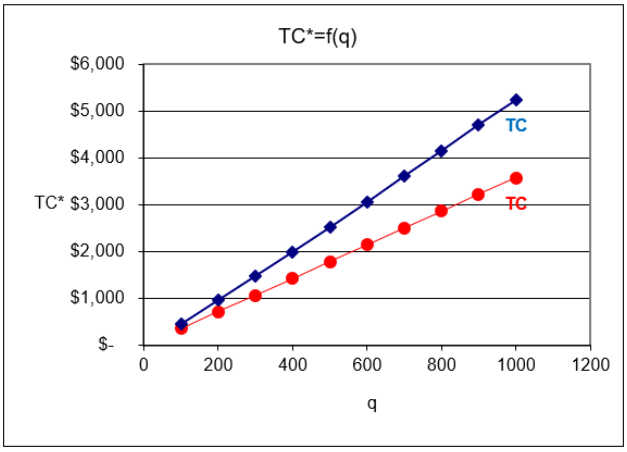


Figure 11.13: Total cost shifts down when labor productivity rises.
Source: DerivingCostFunction.xls!CostFn

You can experiment with other shocks to the cost function. Change input prices, input exponents, or A to see how the cost function shifts. Click the Reset button or ctrl-z (undo) after every trial. Connect what you see on the screen with the shock you applied. Changes in q have no visible effect because you simply move along the cost function.

Interpreting λ*

We end this chapter by showing that the Lagragean multiplier, λ^* has a useful interpretation in the input cost minimization problem. We will see that λ^* gives an easier way to derive a cost function than solving the constrained cost minimization problem with q as a letter and finding $TC^* = f(q)$.

The cost function shortcut uses the fact that λ^* gives the instantaneous rate of change in the optimum value of the objective function as the constraint varies. Thus, λ^* signals how relaxing the constraint would impact the goal.

For utility maximization, we could relax the constraint by increasing income. The budget constraint in the Lagrangean is $m - p_1x_1 - p_2x_2 = 0$ so as m rises, the consumer will be able to reach greater maximum utility. The Lagrangean multiplier tells us how much more utility is gained as income increases. Unfortunately, utility is ordinal so λ^* does not have a useful interpretation in the Theory of Consumer Behavior.

Things are different in the constrained input cost minimization problem. The objective function in this case is minimum total cost and is measured on a cardinal scale. We can directly observe minimum total cost and meaningfully compare how it changes within a firm and across firms. This means we can apply the interpretation of λ^* to input cost minimization.

The constraint in the Lagrangean is $q - f(L, K)$. If we vary the constraint by having the firm produce one more unit of output, we know total cost would rise as we moved to a higher isoquant. The value of λ^* tells us by how much minimum total cost would rise.

For example, at $q = 100$ in *DerivingCostFunction.xls*, λ^* is about \$4.89. You can confirm this by numerical methods (using Excel's Solver and getting the Sensitivity Report) or by analytical methods, solving for λ^* from the three first-order conditions. Either way, you will get (almost exactly) the same answer.

But what does this tell us? The \$4.89 value means that if we increase output by an infinitesimally small amount, minimum total cost will go up by \$4.89-fold. Let's use Excel to work on this.

STEP Click the Reset button in the *CostFn* sheet and take a look at the highlighted cell with a yellow background (P8). Click on it and read the formula.

The value of P8 is \$4.99. That is close to the value of λ^* of \$4.89, but not quite exactly the same. What is going on?

STEP Go to the *CS1* sheet and take a look at the highlighted, yellow-backgrounded cell (E15) (click the Unhide Column E button if needed). Its value is \$4.90.

This is much closer to λ^*'s value of \$4.89. Why? Because the change in q is much smaller in the *CS1* sheet than in the *CostFn* sheet. As the change in q approaches zero, the change in TC^* divided by the change in q will approach λ^*.

STEP Return to the *CostFn* sheet and change cell K8 from 200 to 110. This replicates the *CS1* sheet value for λ^*. Next, set K8 to 101. What do you see?

With K8 set to 101 so that $\Delta q = 1$, $\frac{\Delta TC}{\Delta q} = \$4.89$, the value of λ^*. Well, actually, not exactly \$4.89. If we displayed more decimal places in P8 and computed the value of λ^* to more decimal places, the two would not agree. But they would get closer the smaller we made Δq.

Of course, this is nothing more than a demonstration of the idea of the derivative. If you are puzzled as to how $\frac{\Delta TC}{\Delta q}$ can be that close to λ^* in the *CS1* sheet (a one cent difference seems pretty small), given that the change in q is 10 units (which is hardly infinitesimally small), the answer lies in the total cost function: It simply is not very curvy. Because TC^* follows almost (but not quite) a straight line, computing the slope from $q = 100$ to $q = 110$ is close to the slope of the tangent line at $q = 100$.

The purpose of the work above was to convince you that $\lambda^* = \frac{dTC}{dq}$. The Lagrangean multiplier gives the instantaneous rate of change in minimum total cost with respect to output.

STEP You can confirm the claim that $\lambda^* = \frac{dTC}{dq}$ by changing the parameters in the *CostFn* sheet and keeping your eye on the rose-backgrounded cell H31. It computes the difference between λ^* in H13 and $\frac{dTC}{dq}$ in H30. The difference is always zero because these two things, λ^* and $\frac{dTC}{dq}$ are equivalent.

You might ask, "So what?" In other words, what can we do with the knowledge that $\lambda^* = \frac{dTC}{dq}$? A lot. For one thing, we can easily derive the cost function. After all, the rate of change in total cost as output changes is *marginal cost* (MC). Thus, $\lambda^* = \frac{dTC}{dq} = MC(q)$. This means we can easily get the total cost function by simply integrating λ^* with respect to q.

Furthermore, as we will see when we solve the output profit maximization problem, we usually want marginal revenue and marginal cost, so knowing that $\lambda^* = \frac{dTC}{dq}$ can be a real shortcut. If we have λ^*, then we do not have to derive $TC^* = f(q)$ and then take the derivative to get MC.

The Cost Function has Parents

This section included some complicated ideas, but we end by prioritizing things. There is no doubt that the most important idea is that the cost function has a source and does not appear from nowhere. This is captured by Figure 11.10—the cost function is derived by doing comparative statics analysis on the input cost minimization problem.

Although we are often interested in the response of an endogenous variable to a shock, comparative statics in the input cost minimization problem is focused on how the objective function, minimum total cost, is affected by shocking q. Minimum total cost as a function of q is the cost function.

By explaining what it means to be above or below the cost function in terms of the isoquant–isocost graph, we emphasized the idea that the cost function shows the cheapest way to produce any given output. A good way to remember this is to ponder the striking fact that there is no space below the cost function, meaning that it is impossible to produce the given output any cheaper than the cheapest way possible.

Changes in other parameters besides output cause the entire cost function to shift because minimum total cost depends on all of the exogenous variables. If q changes, we move along the cost function; other shocks shift TC.

Finally, we explained a mathematically sophisticated idea: λ^* provides information on the rate of change of the optimum value of the objective function as the constraint is relaxed. This interpretation of the Lagrangean multiplier holds for every constrained optimization problem.

We did not apply this interpretation in the Theory of Consumer Behavior because utility (the objective function) cannot be cardinally measured. In the old days, when utility was believed to be cardinally measured in utils, λ^* was the marginal utility of money. λ^* would tell you the rate of change in maximum utility if you gave the consumer an infinitesimal increase in income.

Since total cost is directly observable and countable, λ^* can be correctly interpreted as marginal cost, $\frac{dTC}{dq}$. This gives a shortcut to the cost function and MC.

Exercises

1. With the production function, $q = L^{0.75}K^{0.5}$, and exogenous variables $w = 2$, $r = 3$, use Excel to create a graph of the cost function for the same q values as the one in the *CS1* sheet. Copy and paste your graph in a Word document.

2. How is the cost function you just derived different from the one in the *CS1* sheet? Which variable is responsible for generating this difference?

3. From the cost functions in the *CS1* sheet and question 1, what can you deduce about cost functions derived from Cobb-Douglas production functions?

4. If someone solves an input cost minimization problem and finds that $\lambda^* = 50$, what does this mean?

References

The epigraph is from page 333 of Hans Staehle, "The Measurement of Statistical Cost Functions: An Appraisal of Some Recent Contributions," *The American Economic Review*, Vol. 32, No. 2, Part 1 (June, 1942), pp. 321–333, www.jstor.org/stable/1803513. Staehle was optimistic in 1942 that

advances in statistics and data collection would enable economists to estimate cost functions for particular industries. Unfortunately, it is fair to say that Staehle's dream of the discovery of flexible functional forms remains unfulfilled. Empirical work on cost functions usually finds that firms face linear (or nearly linear) total costs (yielding horizontal average and marginal costs) over large ranges of output.

Only 11 percent of firms report that their MC curves are rising. By contrast, about 40 percent claim that their MC curves are falling.

Alan Blinder, Elie Canetti, David Lebow, and Jeremy Rudd

11.4 Cost Curves

In the next chapter, we will work on the firm's second optimization problem: maximize profits by choosing the amount of output to produce. Because profits are revenues minus costs, the cost function plays an important role in the firm's profit maximization problem.

This section is devoted to the terminology of cost curves and an exploration of their geometric properties. Derived from the cost function, a variety of cost curves are used to solve and display the firm's profit-maximization problem. This section defines and derives them.

A basic idea that is easy to forget is that there are many shapes of cost functions. Our work on deriving the cost function used a Cobb-Douglas production function and that gives rise to a particularly shaped cost function. A different production function would give a different cost function. A key idea is that $q = f(L, K)$ determines the shape of $TC^* = f(q)$.

Names and Acronyms

You know that if we track TC^*, minimum total cost, as a function of q, we derive the cost function. Since we will be using other measures of costs, to avoid confusion, we refer to the cost function as the *total cost* (TC) function. The total cost function has units of dollars (\$) on the y axis. We can divide total costs into two parts, *total variable costs*, TVC, and *total fixed costs*, TFC.

$$TC(q) = TVC(q) + TFC$$

If the firm is in the short run, it has at least one fixed factor of production (usually K) and the total fixed costs are the dollar value spent on the fixed inputs (rK). Notice that the total fixed costs do not vary with output. TFC is a constant and does not change as output changes so there is no "(q)" in the TFC function like there is on TVC and TC.

The total variable costs are the costs of the factors that the firm is free to adjust or vary (hence the name "variable costs"), usually L. As output rises, firms need more inputs to produce the increased output so total variable costs rise.

In the long run, defined as a planning horizon in which there are no fixed factors, there are no fixed costs ($TFC = 0$) and, therefore, $TC(q) = TVC(q)$. In other words, the total cost and total variable cost functions are identical.

In addition to total costs, the firm has average, or per unit, costs associated with each level of output. *Average total cost*, ATC (also known as AC), is the total cost divided by the output level.

$$ATC(q) = \frac{TC(q)}{q}$$

Average variable cost, AVC, is total variable cost divided by output.

$$AVC(q) = \frac{TVC(q)}{q}$$

Average fixed cost, AFC, is total fixed cost divided by output.

$$AFC(q) = \frac{TFC(q)}{q}$$

Notice that $AFC(q)$ is a function of q even though TFC is not because AFC is TFC divided by q. Since the numerator is a constant, $AFC(q)$ is a rectangular hyperbola ($y = c/x$) and is guaranteed to fall as q rises. This can be confirmed by a simple example. Say TFC = \$100. For very small q, such as 0.0001, AFC is extremely large. But AFC falls really fast as q rises from zero (and AFC is undefined at $q = 0$). At $q = 1$, AFC is \$100, at $q = 2$, AFC is \$50, and so forth. The larger the value of q, the closer AFC gets to zero (i.e., it approaches the x axis).

It is easy to show that the average total cost must equal the sum of the average variable and average fixed costs:

$$TC(q) = TVC(q) + TFC$$

$$\frac{TC(q)}{q} = \frac{TVC(q)}{q} + \frac{TFC}{q}$$

$$ATC(q) = AVC(q) + AFC(q)$$

We often omit *AFC(q)* from the graphical display of the firm's cost structure (see Figure 11.14) because we know that $AFC(q) = ATC(q) - AVC(q)$. Thus, average fixed cost can be easily determined by simply measuring the vertical distance between ATC and AVC at a given q.

The facts that $AFC(q) = ATC(q) - AVC(q)$ and AFC goes to zero as q rises means that AVC must approach ATC as q rises. Always draw AVC getting closer to ATC as q increases past minimum AVC. Figure 11.14 obeys this condition.

Unlike the total curves, which share the same y axis units of dollars, the average costs are a rate, dollars per unit of output. You cannot plot total and average cost curves on the same graph because the y axes are different.

Another cost concept that we get from the total cost function is *marginal cost (MC)*. Like average costs, MC is a rate and it comes in \$/unit. Marginal cost is often graphed together with the average curves (as shown in Figure 11.14).

Marginal means *additional* in economics. Marginal cost tells you the additional cost of producing more output. If the change in output is discrete, then we are measuring marginal cost from one point to another on the cost curve and the equation looks like this:

$$MC(q) = \frac{\Delta TC(q)}{\Delta q}$$

If, on the other hand, we treat the change in output as infinitesimally small, then we use the derivative and we have:

$$MC(q) = \frac{dTC(q)}{dq}$$

Because TFC does not vary with q, marginal cost also can be found by taking the derivative of *TVC(q)* with respect to q.

Average cost and marginal cost are used to refer to entire functions (see Figure 11.14), but also to specific values. For example, if $ATC = \$10$/unit and $MC = \$3$/unit at $q = 5$, this means that it costs \$10 per unit to make the five units and, thus, the firm had \$50 of total costs to make five units. The MC tells us that the 5^{th} unit costs an additional \$3 so the total cost went from \$47 for 4 units to \$50 for 5 units.

The Geometry of Cost Curves

The average and marginal curves are connected to each other and must be drawn according to strict requirements. Whenever a marginal curve is above an average curve, the average curve must be rising. Conversely, whenever a marginal is below an average, the average must be falling.

For example, consider the average score on an exam. After the first 10 students are graded, there is an average score. The 11th student is now graded. Suppose she gets a score above average. Hers is the marginal score and we know it is above the average so it has to pull the average up. Suppose the next student did poorly. His marginal score is below the average and it pulls the average down. So, we know that whenever a marginal score is below the average, the average must be falling and whenever a marginal score is above the average, the average must be rising. The only time the average stays the same is when the marginal score is exactly equal to the average score.

This relationship between the average and marginal means that the marginal cost curve must intersect the average variable and average total cost curves at their respective minimums, as shown in Figure 11.14. From $q = 0$ to the intersection of MC with ATC, MC is below the ATC and the ATC falls. To the right of the intersection of MC with ATC, MC is above the ATC so the ATC is pulled up. MC and AVC curves share the same relationship.

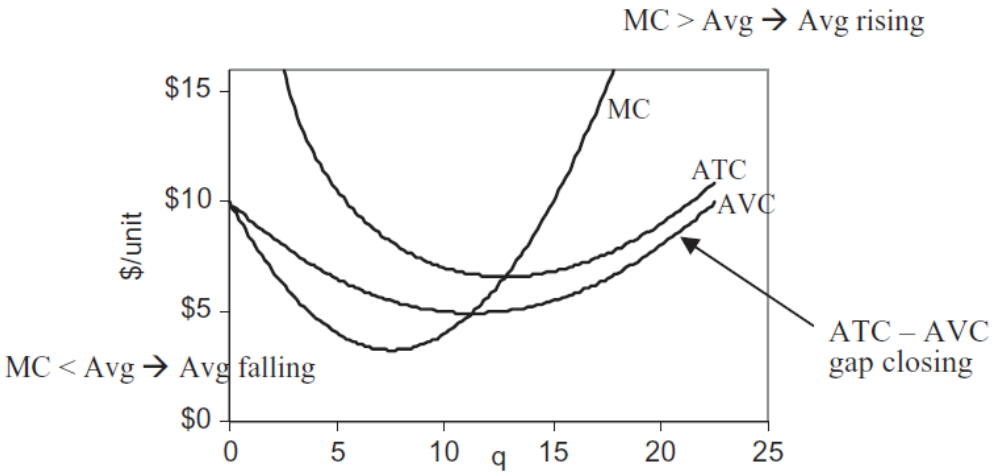


Figure 11.14: Marginal and average relationships.

Figure 11.14 also shows a property that was highlighted earlier: The gap between ATC and AVC must fall as q rises.

You will understand these abstract ideas better by exploring concrete examples. Three cost functional forms will be examined:

1. Cobb-Douglas Cost Curves
2. Canonical Cost Curves
3. Quadratic Cost Curves

Instead of memorizing specific facts or points, look for the pattern and repeated connections. Focus on the relationship between the total and average and marginal curves.

STEP Open the Excel workbook *CostCurves.xls* and read the *Intro* sheet, then go to the *CobbDouglas* sheet to see the first example.

1. Cobb-Douglas Cost Curves

The *CobbDouglas* sheet is the *CostFn* sheet from the *DerivingCostFunction.xls* workbook with the ATC and MC curves plotted below the TC curve. Column I has a formula for the TC curve using L^* and K^*, from which we can compute ATC and MC in columns J and K. Click on an MC cell, for example, cell K4, to see that the cell formula is actually for λ^*. We are using the shortcut that $\lambda^* = MC$.

With L and K both endogenous, there are no fixed factors of production. This means we are in the long run and there are no fixed costs. Thus, $TC = TVC$ and $ATC = AVC$.

It is immediately obvious that the marginal and average curves do not look at all like the conventional family of cost curves as shown in Figure 11.14. In fact, a Cobb-Douglas production function cannot give U-shaped average and marginal cost curves as in Figure 11.14.

Remember that there are many functional forms for cost curves (total, average, and marginal) and the shape depends on the production function. In other words, the production function is expressed in the cost structure of a firm.

STEP Set the exponent on capital, d, to 2 to replicate Figure 11.15.

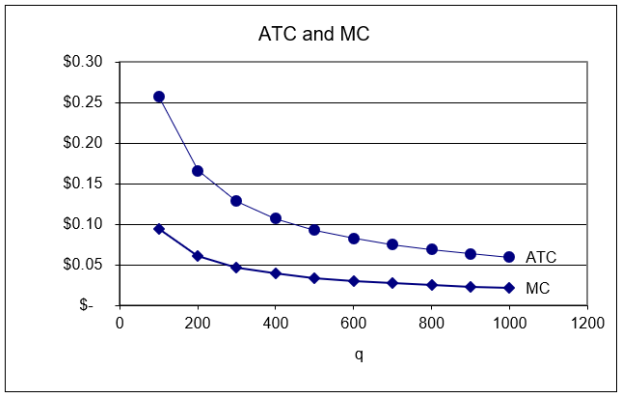


Figure 11.15: Total cost shifts down when labor productivity rises. *Source: CostCurves.xls!CostFn*, after setting $d = 2$.

Because average cost is falling as q rises in Figure 11.15 (and your computer screen), it means that total cost is increasing less than linearly as output rises. The total cost graph on your screen confirms that this is the case. It costs \$33 to make 200 units, but only \$43 to make 400 units. Double output again to 800. How much does it cost? Cell I9 tells you, \$55. This is puzzling. If input prices remain constant, how can we double output and not at least double costs?

The answer lies in the production function. You changed the exponent on capital, d, from 0.2 to 2. Now the sum of the exponents, $c + d$, is greater than 1. For the Cobb-Douglas production function, this means that we are operating under increasing returns to scale. This means that if we double the inputs, we get more than double the output. Or, put another way, we can double the output by using less than double the inputs.

This firm can make 400 units cheaper *per unit* than 200 units. It can make 800 units even cheaper *per unit* because it is taking advantage of the increasing returns to scale.

Increasing returns are a big problem in the eyes of some economists because they lead to a paradox: One firm should make all of the output. There are situations in which increasing returns seem to be justified, such as the case of *natural monopolies*, in which a single firm provides the output for an entire industry because the production function exhibits increasing returns to scale.

The classic examples are utility companies, e.g., electric, water, and natural gas companies. Often, these firms are nationalized or heavily regulated.

We can emphasize the crucial connection between the production function and the cost function via the isoquant map.

STEP Scroll down to row 100 or so in the *CobbDouglas* sheet.

The three isoquants are based on a Cobb-Douglas production function with parameter values from the top of the sheet, except for d, which can be manipulated from the *Set d* radio buttons (above the chart). The three red points are the cost-minimizing input combinations for three different output levels: 100, 120, and 140.

Above the graph, the value of the sum of the exponents, initially 0.95, is displayed. A description of the shape of the total cost function, which depends on the value of $c + d$, and a small picture of that shape is shown. Figure 11.16 has the initial display.

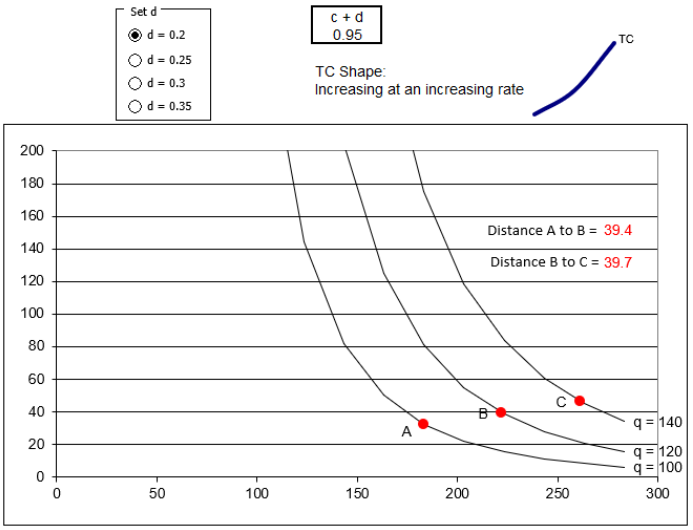


Figure 11.16: Isoquants determine the shape of the cost function.
Source: CostCurves.xls!CobbDouglas.

The spacing between the points is critical. The distance from A to B is a little less than that from B to C. This means that as output is increased from 120 to 140, the firm needs a bigger increase in inputs than when q rose from 100 to 120.

As output continues rising by 20 units, the next isoquant we have to reach is getting farther and farther away, requiring progressively more inputs, and progressively higher costs. This is why TC is increasing at an increasing rate.

STEP Click on the $d = 0.25$ option.

The isoquants shift in because it takes fewer inputs to make the three levels of output depicted. The distance between the isoquants has decreased and TC is linear. Most importantly, the distance between the points is identical.

With $c + d = 1$, the spacing of the isoquants is constant. As q increases by 20, the next isoquant is the same distance away and the firm increases its input use and costs by a constant amount. This is why the TC function is a line, increasing at a constant rate.

STEP Click on the $d = 0.3$ option.

Once again, the chart refreshes and isoquants shift in. Now the distance between the isoquants is decreasing. As q rises, the isoquants get closer together and the total cost function is increasing at a decreasing rate.

STEP Click on the $d = 0.35$ option.

This produces even stronger increasing returns and a TC function that bends faster than $d = 0.3$.

The fundamental point is that the distance between the isoquants reflects the production function. There are three cases:

1. If the distance is increasing as constant increases in quantity are applied, the total cost function will increase at an increasing rate.

2. If the distance remains constant, the cost function will be linear.

3. If the distance get smaller as output rises, the firm has costs that rise at a decreasing rate.

This holds for all production functions and, in the case of Cobb-Douglas, it is easy to see what is going on because the value of $c + d$ immediately reveals the returns to scale and spacing between the isoquants.

But the advantage of Cobb-Douglas in easily displaying the three cases (depending on the value of $C + d$) means it cannot do all three cases at once. A Cobb-Douglas production function can generate a TC functon that is increasing at an increasing or constant or decreasing rate, but not all three.

The shape of the cost function is dependent on the production technology. Repeatedly cycle through the radio buttons, keeping your eye on the isoquants, the distance between the points, and the resulting total cost function. Your task is to understand and cement the relationship between the production and cost functions.

An accordion is a good metaphor for what is going on. When scrunched up, the isoquants are being squeezed together, which gives increasing returns to scale and TC increasing at a decreasing rate. When the accordion is expanded and the isoquants are far apart, we have decreasing returns to scale and TC rising at an increasing rate.

Do not be confused. The reason why increasing (decreasing) returns to scale leads to TC rising at a decreasing (increasing) rate (they are opposite) is that productivity (returns to scale) and costs are opposites. Increased productivity enables slower increases in costs of production. Production increasing at an increasing rate and costs increasing at a decreasing rate are two sides of the same coin.

2. Canonical Cost Curves

STEP Proceed to the *Cubic* sheet.

This sheet displays the canonical cost structure, in other words, the most commonly used cost function. It produces the familiar U-shaped family of average and marginal costs (which Cobb-Douglas cannot).

The canonical cost curves graph can be generated by a cost function with a cubic polynomial functional form.

$$TC(q) = aq^3 + bq^2 + cq + d$$

The d coefficient (not to be confused with the d exponent in the Cobb-Douglas production function) represents the fixed cost. If $d > 0$, then there are fixed costs and we know the firm is in the short run.

Once we have the cost function, the top curve on the top graph in the *Cubic* sheet, we can apply the cost definitions (from the beginning of this section) to get all of the other cost curves. The other total curves are:

$$TVC(q) = aq^3 + bq^2 + cq$$

$$TFC = d$$

STEP Click on each of the three curves in the top graph of the *Cubic* sheet to see the data that are being plotted.

Now turn your attention to the bottom graph. The curves in the bottom graph are all derived from the top graph. Notice that the y axis label is different, the totals in the top have units of \$, while the average and marginal curves have a y scale of \$/unit (of output).

STEP Click on each of the three curves in the bottom graph to see the data that are being plotted.

Custom formatting has been applied to the numbers in the average and marginal cost cells to display "\$/unit" in each cell. It is easy to forget that "\$" is not the units of average and marginal cost curves.

The average total and average variable costs are easy to compute: simply divide the total by q. You can confirm that column E's formula does exactly this. There is no ATC value for $q = 0$ because dividing by zero is undefined.

We can also divide the equation itself by q to get an average. This is done for *AVC*. The formula in cell F2 is "= a_*(A6^2) + b_*A6 + c_" because dividing $TVC(q) = aq^3 + bq^2 + cq$ by q yields $= aq^2 + bq + c$. Notice that *AVC* for $q = 0$ does exist.

Marginal cost is more difficult to understand than average cost. Marginal cost is defined as the additional cost of producing more output. "More" can be an arbitrary, finite amount (such as 1 unit or 10 units) or an infinitesimally small change in the number of units.

If we use an arbitrary, finite amount of increase in q, then we compute MC as $\frac{\Delta TC}{\Delta q}$. We can also compute MC for an infinitesimally small change, using the derivative, $\frac{dTC}{dq}$. These two computations will be exactly the same only if *MC* is a line.

The two approaches are applied in columns G and H. The derivative of TC with respect to q is:

$$TC(q) = aq^3 + bq^2 + cq + d$$
$$\frac{dTC}{dq} = 3aq^2 + 2bq + c$$

Notice how we apply the usual derivative rule, bringing the exponent down and subtracting one from the exponent for each term. The d coefficient, TFC, disappears because it does not have q in it (or, if you prefer, think of d as dq^0). The expression for MC is entered in column G.

Column H has MC for a discrete-size change. You can vary the size of the change in q by adjusting the step size in cell B3.

STEP Make the step size smaller and smaller. Try 0.1, 0.01, and 0.001.

As you make the step size smaller, the values in column H get closer to those in column G. This, once again, demonstrates the concept of the derivative.

Another way to get the cost function is to use the neat result from Lagrangean method. We can simply use $\lambda^* = MC$ and we have the MC curve. No delta-size change or derivative required. If what we really wanted was the total cost function, then we would have to integrate the λ^* function with respect to q. The constant of integration is the fixed cost, which would be zero in the long run.

The family of cost curves in the *Intro* and *Cubic* sheets (and in Figure 11.14) are the canonical cost curves displayed in countless economics textbooks. You might wonder, if not Cobb-Douglas, then what production function could produce such a cost function? That is not an easy question to answer. In fact, the functional form for technology that would give rise to the canonical cost curves is quite complicated and it is not worth the effort to painstakingly derive the usual U-shaped average and marginal cost curves from first principles.

It is sufficient to know that a production function underlies the polynomial TC function and its resulting U-shaped average and marginal cost curves. We also want to keep in mind that if input prices rise, the cost curves shift up and, if technology improves, they shift down.

3. Quadratic Cost Curves

STEP Proceed to the *Quadratic* sheet to see a final example of cost curves.

It is immediately clear that the quadratic functional form is a special case of the cubic cost function, with coefficients a and c equal to zero.

Look at the top chart and connect the shapes of the TC, TVC, and TFC functions to the functional form $TC(q) = bq^2 + d$. Given the coefficient values in the sheet, this gives $TC(q) = q^2 + 1$, $TVC(q) = q^2$, and $TFC = 1$.

The bottom chart does not look familiar, but it obeys the definitions of average and marginal cost explained earlier in this section. ATC is $TC(q)$ divided by q: $ATC(q) = q + \frac{1}{q}$. Similarly, AVC is $TVC(q)/q$, which is q (a ray out of the origin). MC is the derivative of TC with respect to q, which is $2q$.

Although not the usual U-shaped curves, the MC curve (actually, MC is linear) intersects AVC and ATC at their minimums. When MC is below ATC, ATC is falling, but beyond the point at which MC intersects ATC (at the minimum ATC), MC is above ATC and ATC is rising. As q increases, AVC converges to ATC, which implies that AFC goes to zero.

The shapes of the cost curves are not the usual U-shaped average and marginal curves, but this is another of the many possible cost structures that could be derived from a firm's input cost minimization problem.

The Role of Cost Curves in the Theory of the Firm

Cost curves are not particularly exciting, but they are an important geometric tool. When combined with a firm's revenue structure, the family of cost curves is used to find the profit-maximizing level of output and maximum profits.

Cost curves can come in many forms and shapes, but they all share the basic idea that they are derived by minimizing the total cost of producing output, where output is generated by the firm's production function. Different production functions give rise to different cost functions.

The shape of the cost function, rising at an increasing, constant, or decreasing rate, is determined by the production function. With increasing returns to scale, for example, a firm can more than double output when it doubles its input use. That means, on the cost side, that doubling output will less than double total cost. Returns to scale can be spotted by the spacing between the isoquants. With increasing returns to scale, for example, the gaps between the isoquants get smaller as output rises.

No matter the production function, it is always true that for output levels at which marginal cost is below an average cost, the average must be falling and MC above AVC or ATC means AVC or ATC is rising. It is also true that, in the short run (when there are fixed costs), AVC approaches ATC as output rises.

Lastly, consider the message conveyed by Figure 11.17. The arrows show the progression—average and marginal curves come from the total cost function, which comes from the input cost minimization problem (with the production function expressed in the isoquants).

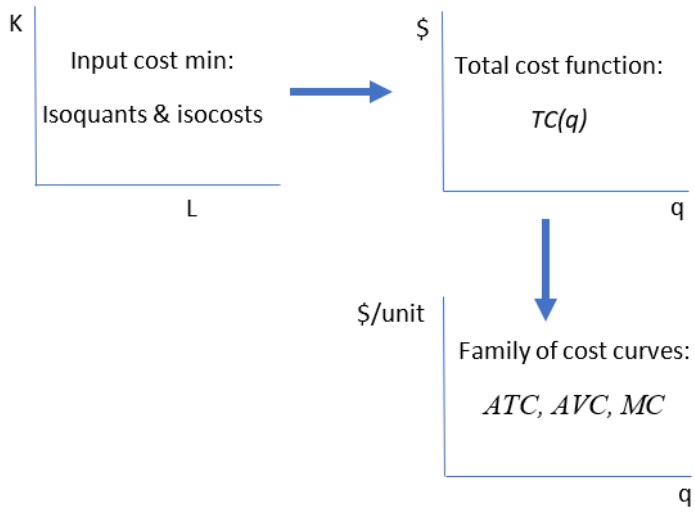


Figure 11.17: Connecting cost graphs.

Economists use graphs to communicate. It may seem like graphs are conjured out of thin air, but this is false. All graphs have a genealogy and a story to tell. When you know where graphs come from, that helps in reading them correctly.

Exercises

1. A Cobb-Douglas production function with increasing returns to scale yields a total cost function that increases at a decreasing rate. Use Word's Drawing Tools to draw the underlying isoquant map for such a production function.

A commonly used specification for production functions in empirical work is the *translog* functional form. There are several versions. When applied to the cost function, you get a result like this:

$$\ln TC = \alpha_0 + \alpha_1 \ln Q + \alpha_2 \ln w + \alpha_3 \ln r + \alpha_4 \ln Q \ln w + \alpha_5 \ln Q \ln r + \alpha_6 \ln w \ln r$$

Notice that the function is a modification of the log version of a Cobb-Douglas function. In addition to the individual log terms there are combinations of the three variables, called interaction terms.

Click the Exercise Questions button at the bottom of the *Q&A* sheet in the *CostCurves.xls* workbook to reveal a sheet with translog cost function parameters. Use this sheet to answer the following questions.

2. Enter a formula in cell B18 for the TC of producing 100 units of output, given the alpha coefficient and input price values in cells B5:B13. Fill your formula down and then create a chart of the total cost function (with appropriate axes labels and a title). Copy and paste your chart in a Word document.

   Hints: $TC = e^{\ln TC}$ and the exponentiation operator in Excel is EXP(). "=EXP(number)" in Excel returns e raised to the power of that number.

3. Compute MC via the change in output from 100 to 110 in cell C19. Report your result.

4. Compute MC via the derivative at $Q = 100$ in cell D18. Report your result.

   Hint: $\frac{d}{dx}(e^{f(x)}) = e^{f(x)} \frac{d}{dx}(f(x))$

5. Compare your results for MC in questions 3 and 4—are your answers the same or different? Explain.

References

The epigraph is from page 218 of Alan Blinder, Elie Canetti, David Lebow, and Jeremy Rudd, *Asking About Prices: A New Approach to Understanding Price Stickiness* (Russell Sage Foundation, 1998). This book reports the results of interviews with more than 200 business executives. The authors explain that asking about a firm's marginal cost "turned out to be quite tricky because the term 'marginal cost' is not in the lexicon of most business people; the concept itself may not even be a natural one" (p. 216). The question was, therefore, phrased in terms of "variable costs of producing additional units."

The results confirmed what many who have attempted to estimate cost curves know: The canonical, U-shaped family of cost curves makes for nice theory, but it is not common in the real world. In fact, many business leaders have no idea what marginal cost is or how to measure it. Do not lose sight, however, of the purpose of the Theory of the Firm. It is not designed to realistically describe a living firm. The Theory of the Firm is a severe abstraction with a primary goal of deriving a supply curve. The next chapter does exactly that.

Chapter 12

Output Profit Maximization

Initial Solution

Deriving the Supply Curve

Diffusion and Technical Change

> There are many occasions, therefore, when several explorers are surprised, and somewhat pained, on meeting each other at the Pole. Of such an occasion the history of the "marginal revenue curve" presents a striking example. This piece of apparatus plays a great part in my work, and my book arose out of the attempt to apply it to various problems, but I was not myself one of the many explorers who arrived in rapid succession at this particular Pole.
>
> Joan Robinson

12.1 Initial Solution

With a total cost function, $TC(q)$, and its associated average and marginal cost curves, we are ready to solve the the firm's output profit maximization problem. The firm chooses the amount of output that maximizes profit, defined as total revenue minus total cost. This is the second of three optimization problems that make up the Theory of the Firm.

All firms face this profit maximization problem, but this chapter works with a perfectly competitive (PC) firm in the short run (SR). There are, of course, many other market structures and types of firms, but perfect competition is the first step from which more sophisticated scenarios arise.

The firm's market structure tells us the environment in which it operates. Its market structure determines the firm's revenue function. A PC firm is the simplest case because it takes price as given. Thus, revenues are simply price times quantity and the revenue function is linear.

Remember that we are not trying to describe the actual operation of a business. In fact, a truly perfectly competitive firm does not exist in the real world. The concept is an abstraction that enables derivation of the supply curve. This is our goal.

Remember also that the short run is defined by the fact that at least one input (usually K) is fixed. In the long run, the firm is free to choose how much to use of every factor. K is fixed not because it is immovable (like a pizza oven or a building), but because the firm has contracted to rent a certain amount. It cannot increase or decrease the amount of K in the short run.

Profit maximization and its graphs may be familiar from introductory economics. This experience will help you, but do not be complacent. Keep your eye on how the economic way of thinking is being applied in this case and make connections with other optimization problems we have explored.

Perfectly Competitive Market Structure

A perfectly competitive firm sells a product provided by countless other firms selling that homogeneous (which means identical) product to perfectly informed consumers. Because the product is homogeneous, there are no quality differences or other reasons for consumers to care about who they buy from. Because consumers are perfectly informed, they know the price of every seller.

Thus, the PC firm's market structure is one of intense price competition. Every firm sells the product at the exact same price because if anyone tried to sell at even a tiny bit higher than the market price, no one would buy from them.

The shorthand term for this environment is *price taking*. The PC firm must take the price and cannot choose its price—price is exogenous to the firm.

In addition to price taking, the market structure of the PC firm is characterized by an assumption about the movement of other firms into and out of the industry: *free entry and exit*. Firms can enter or leave the market, selling the same good as everyone else, at any time.

These two ideas, price taking and free entry, distinguish the PC firm from its polar opposite, monopoly. A monopolist chooses price and has a barrier to entry. Between these two extremes are many other market structures in which real-world firms actually exist.

The PC firm's market structure means that an individual PC firm does not worry about what other firms are doing. Each firm simply chooses its own output to maximize profit and does not watch the other firms to gain a strategic advantage. In this sense, there is no rivalry in perfect competition.

Setting Up the Problem

As usual, we organize the optimization problem into three parts:

1. Goal: maximize profits (π, Greek letter pi), which equal total revenues (TR) minus total costs (TC).
2. Endogenous variable: output (q).
3. Exogenous variables: price of the product (P), input prices (the wage rate (w) and the rental rate of capital (r)), and technology (parameters in the production function).

Unlike the consumer's utility maximization and the firm's input cost minimization problems, this profit maximization problem is unconstrained. The firm does not have a restriction, like a budget constraint or isoquant, that limits its choice of output to a particular range. It can choose any non-negative level of output.

This greatly simplifies the optimization problem. For the analytical method, it means we do not need the Lagrangean method. All we need to do is take a single derivative and set it equal to zero.

Finding the Initial Solution

Suppose the cost function is:

$$TC(q) = aq^3 + bq^2 + cq + d$$

Then we can form the PC firm's profit function and optimization problem like this:

$$\max_q \pi = TR - TC$$

$$\max_q \pi = Pq - (aq^3 + bq^2 + cq + d)$$

As usual, we have two ways to solve this optimization problem: numerically and analytically.

STEP Open the Excel workbook *OutputProfitMaxPCSR.xls* and look over the *Intro* sheet.

The *Intro* sheet is not meant to be immediately understood. It offers highlights of material that will be explained and prints as one landscaped page. It provides a compact summary of the optimal solution of the output profit maximization problem for a perfectly competitive firm in the short run.

STEP Proceed to the *OptimalChoice* sheet to find the initial solution.

The sheet is organized into the components of an optimization problem, with goal, endogenous, and exogenous variable cells.

Initially, the firm is producing nine units of output and making $11.74 of profit. Is this the highest profit it can possibly make?

No. The sheet reveals the information needed to give this answer. By comparing marginal revenue (MR) and marginal cost (MC), we immediately know that the firm would make a mistake (we would say it is inefficient) if it produced just nine units.

The MC of the ninth unit is \$3.52 as shown in cell B22, but what about MR? Perhaps you remember from introductory economics that $P = MR$ for perfectly competitive firms? We can see that the additional revenue produced by the last unit, \$7 (the price), is greater than the additional cost, \$3.52 (cell B22). Thus, the firm should produce more. How much exactly should the firm produce?

STEP Run Solver to find out.

Look carefully at B22. At the optimal solution, $q^* \approx 13.09$, $MC = \$7$ per unit. $P = MC$, a special case of $MR = MC$ for a PC firm, is the equimarginal condition in this problem, analogous to $MRS = \frac{p_1}{p_2}$ and $TRS = \frac{w}{r}$. When the equimarginal condition is met, the firm is guaranteed to be maximizing profits.

To find the optimal solution via the analytical method, we take the derivative of the profit function with respect to q, set it equal to zero, and solve for q^*. Our cubic cost function introduces the complication that the solution has two roots so we have to use the quadratic formula.

STEP Click the Show the Analytical Solution button to see how to solve this problem with calculus.

Cell AC17's formula has the root that maximizes profits (the other root minimizes profits—more on this in the next section). As usual, Solver and calculus agree (not exactly, but they give effectively the same answer).

Representing the Optimal Solution with Graphs

Since this is an unconstrained optimization problem (unlike utility maximization and input cost minimization), the graphical display of the optimal solution is different.

The firm's output profit maximization problem is usually represented by a graph that depicts the family of cost curves along with marginal and av-

erage revenue. Figure 12.1 and the *Intro* sheet shows this canonical graph for a perfectly competitive firm (signaled by the fact that firm demand is horizontal, so marginal revenue equals demand).

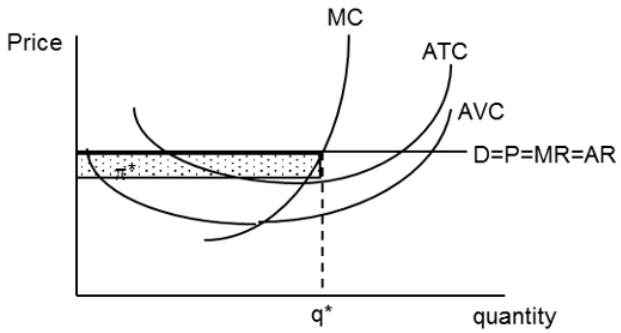


Figure 12.1: The canonical output profit maximization graph.
Source: OutputProfitMaxPCSR.xls!Intro.

Figure 12.1 is the usual display of the optimal solution, but it is actually part of a much larger graphical display.

STEP Proceed to the *Graphs* sheet to see how Figure 12.1 fits into the bigger picture, also shown in Figure 12.2. Zoom out to see all four graphs.

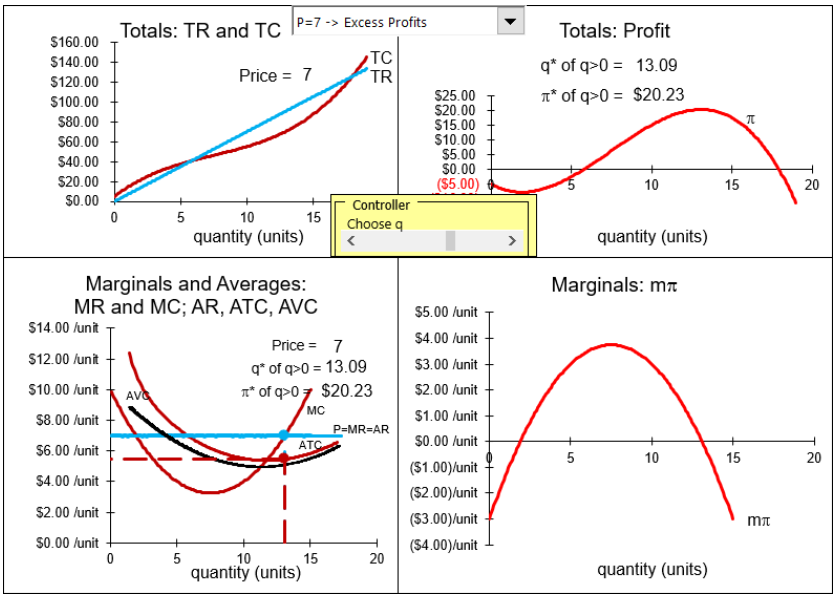


Figure 12.2: Four graphs of output profit maximization.
Source: OutputProfitMaxPCSR.xls!Graphs.

Each of the four graphs in Figure 12.2 and on your screen can be used to show the firm's optimization problem and its solution. We will walk through each one.

1. The top left graph plots total revenue and total cost. TR is linear because the firm's market structure is perfect competition, hence, it is a price taker. The cubic total function produces the shape of TC. The firm wants to choose q to maximize the difference between revenues and costs.

2. The top right graph shows the profit function, which is $TR - TC$. The firm wants to choose q so that it is at the highest point on the profit hill.

3. The bottom right graph displays marginal profit, which can be expressed as the derivative of the profit function with respect to q. The firm can find the maximum profit by choosing q so that marginal profit is zero. This is the first-order condition from the analytical solution.

4. Finally, the bottom left graph is the usual display. The firm chooses q where MR (which equals P given that the firm is a price taker) equals MC. Profits can be calculated as the area of the rectangle $(AR - ATC)q$.

To be clear, all four graphs in Figure 12.2 show the same optimal q and maximum profits, but the graph that is most often used is the bottom left. It highlights the comparison of MR and MC and the family of cost curves provides information about the firm's cost structure. We can also find profits as the area of the rectangle (with blue top and dashed line bottom).

STEP Move the output with the slider control (in the middle of the four charts) to the left and right of q^* to see how the profit rectangle changes.

Only when q is such that $MR = MC$ do you get the maximum area of the profit rectangle. Moving left from optimal q, you can make the rectangle taller, but you must make it shorter to do this and you end up with less area. You can make the rectangle longer by moving right from optimal q, but ATC rises and the rectangle gets thinner, so once again the area falls.

The intersection of MR and MC immediately reveals the optimal q. Profits at any q are also easily seen as the area of a rectangle, length times width, with units in dollars. Because the y axis is a rate, \$/unit, and the x axis is in units of the product, multiplying the two leaves dollars. In other words,

say the product is milk in gallons. Then price, average total, and average variable cost are all in \$/gallon. Suppose that at a price of \$2/gallon, $MR = MC$ at an output of 7,000 gallons and ATC = \$1.50/gallon at this output. Clearly, profits are (\$2/gallon - \$1.50/gallon) x 7,000 gallons, which equals \$3,500.

We can compute profits from the profit rectangle at any level of output. The height of the rectangle is always average revenue (which equals price) minus average total cost. This vertical distance is average profit. When multiplied by the level of output, we get profits, in dollars, at that level of output.

The bottom left graph has another advantage over the other graphs. It can be used to explain a curious and puzzling feature of a firm's short run profit maximization problem. The story revolves around a firm with negative profits and what it should do in this situation.

The Shutdown Rule

The firm has an option when maximum profits are negative: it can simply shut down, close its doors, hire no workers, and produce nothing. The *Shutdown Rule* says the the firm will maximize profits by producing nothing ($q^* = 0$) when $P < AVC$.

The key to whether the firm shuts down or continues production in the face of negative profits lies in its fixed costs. If the firm can do better by shutting down and paying its fixed costs instead of producing and choosing the level of output where $MR = MC$, then it should produce nothing.

Continuing production in the face of negative profits versus shutting down are actually the last two of four possible profit positions for the firm.

1. Excess Profits: $\pi^* > 0$ and $P > ATC$
2. Normal Profits: $\pi^* = 0$ and $P = ATC$
3. Negative Profits, Continuing Production: $\pi^* < 0$ and $P \geq AVC$
4. Shutdown: $\pi^* < 0$ and $P < AVC$

Case 1, excess profits, occurs whenever maximum profits are positive. The example we have been working on is this case. With $P = 7$, we know that $q^* = 13.09$ and $\pi^* = \$20.23$.

STEP In the *Graphs* sheet, click on the pull down menu (over cell R5) and select the *Zero Profits* option.

Your screen now looks like Figure 12.3.

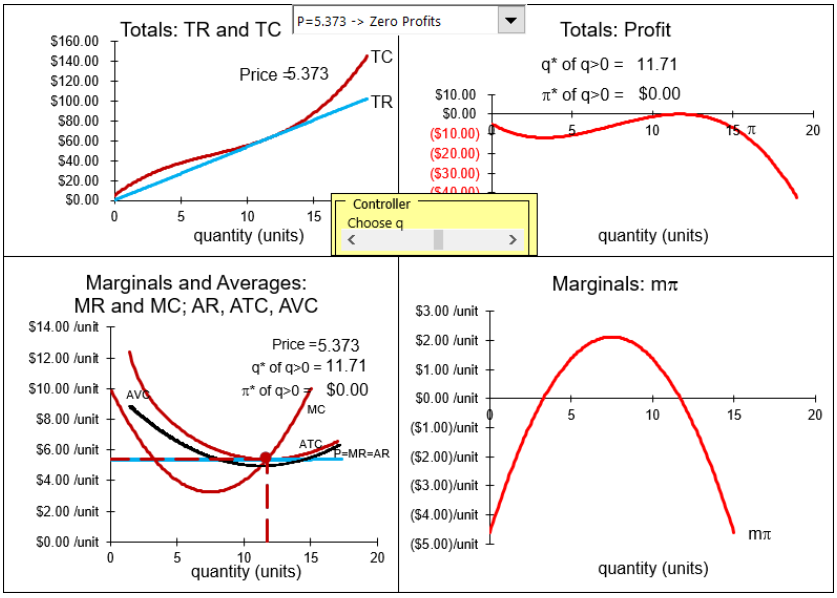


Figure 12.3: Case 2: Normal (zero) profits.
Source: OutputProfitMaxPCSR.xls!Graphs.

Notice that the price ($5.373) in the bottom left chart just touches the minimum of the average total cost curve. The profit rectangle has zero area because it has zero height. The best the firm can do is zero profits—all other choices of q lead to lower (negative) profits.

In the top left graph, you can see that *TR* just touches *TC*. In the top right graph, the top of the profit hill just touches the x axis. These charts confirm what the bottom left chart tells us—with $P = \$5.373$, q^* yields $\pi^* = 0$.

The third and fourth profit cases are the flip side of the first two in the sense that price is so low that profits are now negative. This means firms will leave in the long run, but another question arises: should the firm shut down immediately or continue production?

STEP Click on the pull down menu (over cell R5) and select the *Neg Profits, Cont Prod* option.

With the *Neg Profits, Cont Prod* option selected, $P = 5.10$. The firm produces $q^* = 11.43$ and suffers negative maximum profits of $-\$3.16$. Notice that price is below ATC in the bottom left graph, so that the profit rectangle, $(AR - ATC)q$, will be a negative number. (The area is not negative, but it is interpreted as a negative amount since revenues are below costs.) In the top left graph, the TR line is below the TC curve. In the top right graph, the profit function is below the x axis. There is a maximum, or top of the hill, but it is negative, like a mountain under water.

Keep your eye on the top right graph, reproduced as Figure 12.4. Notice that the top of the profit function is higher than the intercept (where $q = 0$). It is better for the firm to continue production, even though it is earning negative profits of $-\$3.16$ at the optimal output level, because it would make an even lower negative profit of $-\$5$ (the fixed cost) if it shut down.

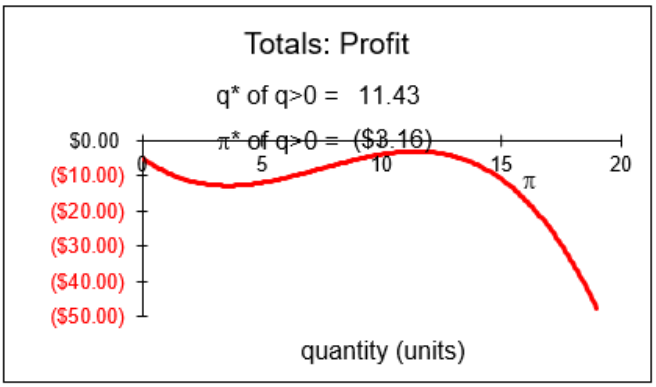


Figure 12.4: Case 3: Negative profits, continuing production.
Source: OutputProfitMaxPCSR.xls!Graphs.

The canonical graph of profit maximization can be used to determine whether the firm should produce or shut down by comparing price to average variable cost. The Shutdown Rule is easy: hire no labor and produce nothing if $P < AVC$.

STEP Look at the bottom left graph on your screen. It confirms that the Shutdown Rule works. Profits are negative because price is below average total cost, but the firm will continue production because $P > AVC$. When the relationship between P and AVC is such that price is greater than average variable cost, it means that the top of the profit function is higher than the y intercept, as in Figure 12.4.

STEP Click on the pull down menu (over cell R5) and select the *Neg Profits, Shutdown* option. Figure 12.5 displays the top right graph.

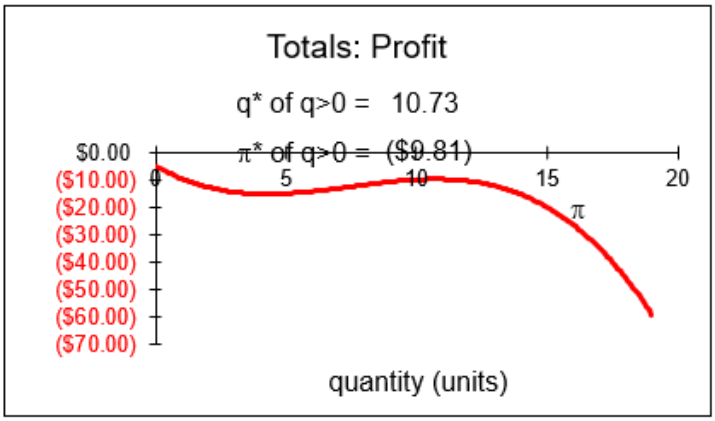


Figure 12.5: Case 4: Negative profits, shut down.
Source: OutputProfitMaxPCSR.xls!Graphs.

In this case, the top of the profit function is below the y intercept. In other words, the maximum profit if the firm produces, $-\$9.81$, is worse than the negative profit incurred if the firm shuts down, $-\$5$. The firm optimizes by choosing $q^* = 0$, that is, shutting down.

STEP Look at the bottom left graph on your screen. Once again, we have confirmation of the Shutdown Rule. With $P = 4.5$, $P < AVC$ and the firm should shut down.

STEP Carefully watch the canonical (bottom left) and profit function (top right) graphs as you change the price (with the pull down menu over cell R5).

As long as $P > AVC$, the top of the profit hill is above the y intercept. If $P = AVC$, the two are exactly equal and the firm is indifferent between producing and shutting down.

$P < AVC$ is the magic cutoff point. When this happens, the top of the hill is below the y intercept (which is the negative profit suffered if the firm produces nothing). Thus, the firm's best choice is to produce nothing.

Here is why the rule works. Multiply the Shutdown Rule by q to get:

$$(P < AVC)q$$
$$Pq < AVCq$$
$$TR < TVC$$

$TR < TVC$ is a restatement of the Shutdown Rule—produce nothing if total revenue cannot cover total variable costs. This makes sense. Why produce if you can't even pay for the variable expenses? You are better off not producing at all.

If total revenue is less than average total cost, then profits are negative. However, the firm can be in a situation where $TR < TC$, but $TR > TVC$. If so, then production makes sense because you will be able to reduce some of the fixed costs you have to pay no matter what you do. Profits are negative, but it is better to produce than not produce because variable costs are covered and fixed costs are at least partially reduced.

STEP For a summary of the four cases and what the Shutdown Rule is doing, click the Show Cases button (over cell AC5).

What's Normal about Zero Profits?

In economics, zero profits are called *normal profits*. This is confusing. Zero sounds bad, not normal. There is a logical explanation, but it requires a clear separation of accounting versus economic profits. They differ because economists include opportunity costs when calculating economic profits.

- Accounting profits = revenues - explicit costs
- Economic profits = revenues - explicit costs - opportunity costs

In economics, without an adjective, "profits" means *economic profits*. So, when profits are zero that means economic profits are zero. Economic profits have had an extra item subtracted, the opportunity costs of using firm resources to make this particular product.

An accountant would subtract explicit (out-of-pocket) costs (wages, rent, etc.) from revenues and if this number is positive, announce that the firm is making money. The economist would then subtract the cost of the profits that could be made by the next best alternative industry that the firm could

be in. If economic profits are zero, it means the opportunity costs are exactly equal to the accounting profit and the firm cannot do better by switching to its next best alternative.

Although this may seem needlessly contorted at first, there is a nice interpretation of economic profits: If positive, the firm will stay in the industry and new firms will enter in the long run; if negative, the firm will exit in the long run; and if zero, there will be neither exit nor entry in the long run. It is in this sense of equilibrium that we say zero profits are normal. With $\pi = 0$, there is stability and no tendency to change in the movement of firms.

The distinction between economic and accounting profits also explains why positive profits are *excess* profits. It is not meant as a pejorative term, but to indicate that the firm is earning greater profits than are needed to keep producing that product in the long run. Excess profits also mean that others are attracted and will enter that industry.

Economists are not concerned with how much money the firm made, but with profits as a signal to entry and exit. Defining economic profits as accounting profits minus opportunity costs gives us a profit measure that tells us whether the firm will stay or leave in the long run.

Shutdown Rule and Corner Solution

The Shutdown Rule is usually covered in introductory economics. Memorization is often all that is achieved. We can do better by properly situating the Shutdown Rule in the landscape of mathematical and economic concepts—it is a corner solution.

Recall that, in the Theory of Consumer Behavior, there are situations in which the MRS does not equal the price ratio, yet the solution is optimal. This is a corner solution.

Food stamps are an example. The fact that food stamps can only be used to buy food creates a horizontal segment on the budget constraint so that a consumer might not be able to make MRS $= \frac{p_1}{p_2}$. At the kink in the constraint, the consumer is optimizing even though the equimarginal condition is not met.

Corner solutions are a general phenomenon. They can be seen whenever a restriction or border blocks further improvement in the objective function. Consider Figure 12.6 which sketches a maximization problem to highlight the difference between an interior and a corner solution. In panel B, the agent cannot choose negative values of the x variable and, therefore, the function is cut off by the y axis.

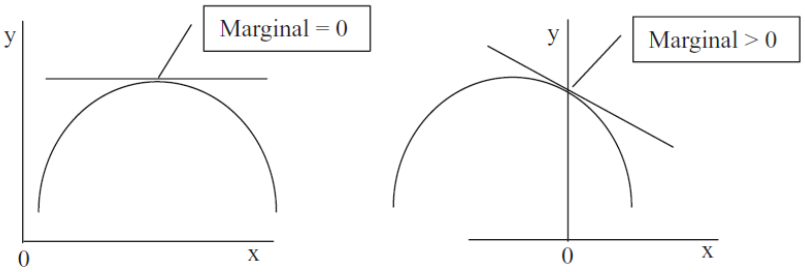


Figure 12.6: Understanding the corner solution.

In panel B, although the marginal condition is not met, we have an optimal solution, defined as doing the best we can without violating any constraints.

Shutting down is another example of a corner solution because, once again, the equimarginal condition is not met at $q = 0$, yet producing nothing is the optimal solution. Shutting down is an unusual example of a corner solution because there *is* a place where the marginal condition is met (there is an output where $MR = MC$), but it is not optimal. The profit function twists in such a way (see Figure 12.5) that profit is decreasing as output rises from zero. This means that profits would go up if we were able to produce negative output. Since we are not allowed to choose $q < 0$, we have a corner solution.

How can we know if we should choose q at $MR = MC$, the interior solution, or shut down, the corner solution? The only way is to compare the profit positions at the two quantities. The good news is that no checking is required for cases 1 and 2. As long as profits are non-negative, there is no way that a profit of minus total fixed cost can be better than the interior solution of q where $MR = MC$. But, whenever, $MR = MC$ yields negative maximum profits, comparing those negative profits to TFC is necessary. Or, you could just use the Shutdown Rule and see if P ¡ AVC, which will give the same, correct answer.

The complexity of the firm's profit maximization problem in the short run, with its shutdown possibility, should increase your sensitivity to lurking problems with analytical and numerical methods. We know neither is perfect so there may be glitches in applying these methods to the firm's profit maximization problem. The *Q&A* sheet provides an example. Be sure to look carefully at questions 2 and 3.

Finding and Displaying the Initial Solution

The output profit maximization problem for a PC firm in the short run is a single-variable (q) unconstrained problem. It can be solved with numerical and analytical methods. The equimarginal rule applied is that $MR = MC$ and since price taking behavior means that $P = MR$ for a PC firm, the equimarginal rule is often shown as $P = MC$.

The firm's profit maximization problem contains a complication in the short run. If maximum profits are negative, it is possible that the firm is better off not producing anything. A shortcut to determine whether or not to produce when $\pi^* < 0$ is the Shutdown Rule, $P < AVC$.

The initial optimal solution is displayed by a canonical graph that superimposes the firm's revenue side (average and marginal revenue) over its cost structure (average and marginal costs). Optimal output is easily found where MR intersects MC (as long as $P > AVC$) and maximum profit is displayed as the area of the appropriate rectangle. The ability to instantly show the optimal solution, maximum profits, and whether or not to shut down explains the popularity of this graph.

You can think of the firm as walking through a series of three steps when solving its profit maximization problem:

1. Choose q where $MR = MC$ in the canonical graph.
2. Compute profits at q^* via $(AR - ATC)q$ (the profit rectangle).
3. If profits are negative, shut down if $P < AVC$.

The PC firm's profit maximization is simpler in the long run. If $\pi < 0$, firms exit the industry; $\pi > 0$ (also known as excess profits) lead to entry. Thus, in long run equilibrium (a state never actually attained), $P = ATC$ and $\pi = 0$ for all firms. This is why zero economic profits are called normal profits.

Exercises

1. Use Excel's Solver to find the optimal output and profit for a firm with cost function $TC = 2q^2 + 10q + 50$ and $P = 40$. Take a screen shot of your optimal solution (including output and profits) and paste it in a Word document.

2. Use analytical methods to solve the problem in the previous question.

3. For what price range will the firm in question 1 shut down? Explain.

4. If fixed costs are higher, will this influence the firm's shutdown decision? Explain.

References

The epigraph is from the foreword (p. vi) of Joan Robinson, *The Economics of Imperfect Competition* (first edition, 1933, followed by many reprints). In a male-dominated profession, Joan Robinson established herself as a well-known, important economist. She helped create the Theory of the Firm, including the canonical graph with average and marginal revenue and cost that is used to this day.

Ironically, however, much of her work was critical of mainstream economics. Her famous Richard T. Ely lecture at the 1971 American Economics Association conference pulled no punches:

> For once the president of the AEA was a dissident. This was the veteran institutionalist and Keynesian John Kenneth Galbraith, a longtime friend of Robinson's and celebrated critic of US capitalism and its apologists in academic economics. Galbraith now offered her the most important platform she had ever occupied. Robinson took full advantage of it, delivering an abrasive, challenging, deliberately provocative indictment of neoclassical economics that was designed to polarize her audience between the old and conservative and the young and progressive. (John Edward King, *A History of Post Keynesian Economics Since 1936* (2002), p. 123.)

As all the functions $\phi_k(D_k)$ are supposed to increase with D_k, the expression for D_k derived from the equation $p = \phi_k(D_k)$ is itself a function of p, increasing with p. [Translation: The supply curve, $q^* = f(P)$ is derived from $P = MC$ and it is upward sloping because MC is upward sloping.]

Augustin Cournot

12.2 Deriving the Supply Curve

The most important comparative statics analysis of the firm's output profit maximization problem is based on tracking q^* (quantity supplied) as price changes, ceteris paribus. This gives us the firm's supply curve.

An important thing to remember is that the supply curve has two parts:

1. MC when $P >$ min AVC
2. Zero otherwise (Shutdown Rule)

As usual, we have numerical and analytical methods at our disposal for the comparative statics analysis that generates the supply curve. Before we begin, we show how Solver can be modified to deal with the shut down possibility and revisit the fact that it is not a silver bullet.

Solver Issues

STEP Open the Excel workbook *DerivingSupply.xls*, read the *Intro* sheet, then go to the *OptimalChoice* sheet to see an implementation of a PC firm's profit maximization problem in the short run.

The sheet looks like the *OptimalChoice* sheet in the *OutputProfitMaxPCSR.xls* workbook (from the previous section), but it has a few additional cells.

The IF statements in cells C4 and C8 of the *OptimalChoice* sheet are a convenient way to incorporate the firm's shutdown option.

STEP Click on C8 to reveal its formula: = IF(max profit >= − d, q, 0). We will use this cell as the correct optimal solution in all cases, including the shutdown case.

It is easy to see that Solver has been run because at $q \approx 10$ in cell B8, $MR = MC$ since $P = 4$ and cell B18 reports $MC = 4$. This q, however, is not the optimal solution because cell B4 shows that $\pi = -15$ (using the common convention that "()" denote negative numbers). This firm would be better off not producing at all and suffering a loss of $TFC = -5$. The Shutdown Rule says the same thing since $P < AVC$ (cell B15 is \$5).

While Solver's answer is wrong (because it found the top of the profit hill, which is lower than the y intercept at $-TFC$), we can add a step to Solver where we check for exactly this situation. This is what cells C8 and C4 do.

The expression $max\_profit \geq -d$ is used to test if Solver's answer (the interior solution) has higher profits than negative total fixed costs (the corner solution). If true, it keeps Solver's solution; if false, the optimal solution is zero (shut down).

Solver will find the best of the positive levels of output in cell B8 and the IF statement in cell C8 checks to make sure that the best solution (of the $q > 0$) is better than shutting down and producing nothing ($q = 0$).

With $P = 4$, the best of all of the positive levels of output, $q = 10$, provides a profit of minus \$15. Cells C4 and C8 show that producing nothing yields a higher profit (and smaller loss) of minus \$5 and is the correct optimal solution.

While this is an improvement over manually checking Solver's answer, there is another potential problem with Solver in this application.

STEP To see the problem, set P (cell B12) to 7 and run Solver.

The optimal q is approximately 13.09 and the firm is enjoying excess profits. Cells B4 = C4 and B8 = C8 because Solver's answer gives profits greater than minus *TFC*. All is well.

STEP Now set cell B8 = 1. Run Solver from this initial value.

Solver's result is disastrous! What happened?

STEP Click the Solver Explained button to see why starting from $q = 1$ leads Solver astray.

The explanation on the sheet makes it clear that the initial or starting value can play a critical role when numerical methods are utilized. This profit maximization problem has a sufficiently complicated surface that a numerical algorithm, such as Solver, cannot easily distinguish between local and global optimal solutions. There is no simple fix. The lesson is that you have to know the optimization problem you are dealing with and be careful interpreting the answers provided by a numerical algorithm.

The explanation of Solver's failure involves the minimum point of the profit function and this provides an opportunity to explain the two roots in the quadratic formula. A picture, in this case, really is worth a thousand words.

STEP Click the Show Analytical Solution button.

Cell Z17 has the other root from the quadratic formula (computed by adding instead of subtracting the square root term). Both roots are places where the profit function is flat (in the top right graph on the sheet). Notice how the dashed lines from the max and min profit points lead to points where marginal profit ($m\pi$) is zero. These are the two roots in the quadratic formula.

The two roots can also be seen in the canonical, bottom left graph as the two points where MR and MC intersect. Of course, we only care about the root that maximizes profits. One way to ensure that $MR = MC$ yields a profit max is to make sure that $MC < MR$ to the left of the intersection. In other words, MC cuts MR from below.

Numerical Methods to Derive the Supply Curve

STEP Set cell B8 back to 10 and $P = 4$ so Solver will converge to the local max at $q = -15$.

STEP Run the Comparative Statics Wizard from $P = 4$ with 0.05 sized shocks 100 times. Track the C4 and C8 cells as endogenous variables. You can safely ignore the warning—you are using the CSWiz to keep track of these cells, but will not include them as changing cells in the Solver dialog box.

Your results will look like those in the *CS1* sheet. Notice that at low prices, the firm is producing nothing. This is the part of the supply curve where the firm shuts down to maximize profits.

The supply curve and inverse supply curves can be graphed with the CSWiz data, as shown in Figure 12.7 and the *CS1* sheet. Of course, the tail runs along the quantity axis all the way to zero. Just as with the demand curve, $q = f(P)$ is the supply curve and flipping the axes, $P = f^{-1}(q)$, gives the inverse supply curve.

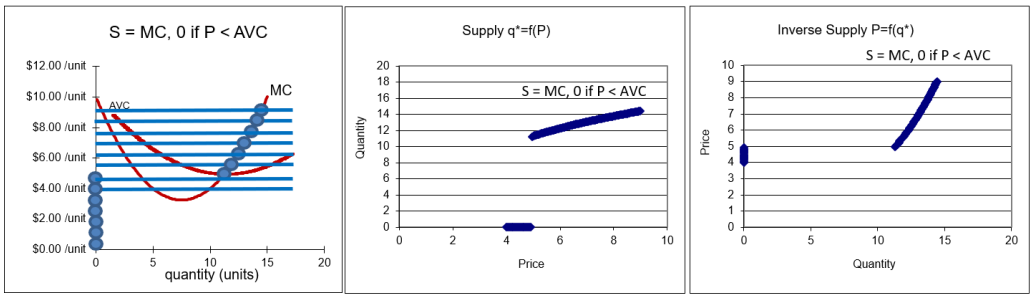


Figure 12.7: Deriving supply and inverse supply curves.
Source: DerivingSupply.xls!CS1.

Figure 12.7 applies our usual graphical exposition. The leftmost chart is the underlying graph from which the other charts are produced. We shock P and track q^*. This gives the supply curve.

Unlike the demand curve, however, notice that the supply curve follows *MC* as long as *P* is not below *AVC*. The discontinuity is at the minimum *AVC*. Row 32 of the *CS1* sheet shows the break occurs for this cost function between \$4.90 and \$4.95. Prices below this minimum *AVC* value result in no quantity supplied since the firm shuts down.

Analytical methods can be used to find the discontinuity. First, we obtain an expression for *AVC*.

$$TC = 0.04q^3 - 0.9q^2 + 10q + 5$$
$$TVC = 0.04q^3 - 0.9q^2 + 10q$$
$$AVC = \frac{TVC}{q} = 0.04q^2 - 0.9q + 10$$

Then we take the derivative of AVC with respect to q and set it equal to zero to find its minimum point.

$$\min_q AVC = 0.04q^2 - 0.9q + 10$$
$$\frac{dAVC}{dq} = 0.08q - 0.9 = 0$$
$$q^* = \frac{0.9}{0.08} = 11.25$$

By plugging this minimum value of output into the AVC function, we know the price at which the discontinuity kicks in.

$$AVC[q = 11.25] = 0.04[11.25]^2 - 0.9[11.25] + 10 = 4.9375$$

In the *CS1* sheet, the discontinuity occurs when price rises from \$4.90 to \$4.95. Our analytical work tells us that the discontinuity is exactly at \$4.9375. Any price below this yields optimal q of zero.

Notice how we used the derivative to find the value of q at which the rate of change for the AVC curve was zero. This is the bottom of the U-shaped AVC curve and prices below this AVC result in shutting down. The lesson is that derivative is a tool that has a variety of uses.

The *CS1* sheet also computes the price elasticity of supply in column E.

STEP Scroll down to see a comparison of slope and elasticities via the Δ and derivative approaches.

In this case, the two approaches are not exactly the same because q^* is non-linear in P. The sheet has all of the details in case you want to refresh your understanding of this concept.

Analytical Methods to Derive the Supply Curve

For the analytical approach, we use a different cost function to give us more practice.

$$TC(q) = q^2 + 20$$

With this quadratic cost function, we can set up and solve the PC firm's profit maximization problem. Because it is a perfectly competitive firm, we know price is given and, thus, $TR = Pq$. Therefore, the optimization problem is:

$$\max_q \pi = Pq - (q^2 + 20)$$

We proceed by taking the derivative with respect to q and setting it to zero, then solving this first-order condition for optimal q.

$$\frac{d\pi}{dq} = P - 2q = 0$$

$$q^* = \frac{1}{2}P$$

This is the supply function. It gives the quantity supplied by a firm at every given price. For example, with $P = 20$, $q^* = 10$.

The inverse supply curve is found by expressing the equation as $P = f(q)$.

$$P = 2q^*$$

The supply function tells us that q^* increases by one-half fold for every increase in P. The size of the change in P does not matter since $\frac{dq}{dP}$ is constant.

The price elasticity of supply is $+1$.

$$\frac{dq}{dP} = \frac{1}{2}$$
$$\frac{dq}{dP}\frac{q}{P} = \frac{1}{2}\frac{P}{\frac{1}{2}P} = 1$$

We can compute the price elasticity of supply from one point to another. We know that at $P = 20$, $q^* = 10$. If $P = 30$, $q^* = 15$. A 50% rise in price led to a 50% increase in quantity supplied so the price elasticity of supply is $+1$. The result is the same as the derivative approach because q^* is linear in P.

A PC firm with a quadratic cost function will not shut down with any price greater than zero. By constructing its family of cost curves and graph of the optimal solution, we can see why. We begin with the cost curves. We know $TVC = 2q$ and $TFC = 20$. Then we can find the average and marginal curves.

$$ATC(q) = \frac{TC}{q} = \frac{q^2 + 20}{q} = q + \frac{20}{q}$$
$$AVC(q) = \frac{TVC}{q} = \frac{q^2}{q} = q$$
$$MC(q) = \frac{dTC}{dq} = \frac{d(q^2 + 20)}{dq} = 2q$$

STEP Proceed to the *Graphs* sheet to see the four graph display of the optimal solution for this problem.

If $P = 20$, then $q^* = 10$ and $\pi^* = \$80$. It is also obvious that there is no positive price at which this firm will shut down because AVC is simply a ray with slope +1 out of the origin. Thus, price can never fall below AVC.

Notice also how there is only one point where $MR = MC$, unlike the two intersections we saw with the cubic cost function. The quadratic cost function cannot produce the S-shape TC needed for the profit function to have a minimum profit at the bottom of a U-shape. The profit function in the top right graph has a single top of the hill (where $m\pi = 0$).

Points Off the Supply Curve

As we did with the demand curve (see Figure 4.12), we can explore the meaning of being off the supply curve. The interpretation is quite similar.

STEP Return to the *CS1* sheet and manipulate the point off the supply and inverse supply curves with the scroll bar in column E.

Figure 12.8 shows what is on your screen, but in Excel you can move the red dot. As you do, the chosen q and profit for that quantity is displayed.

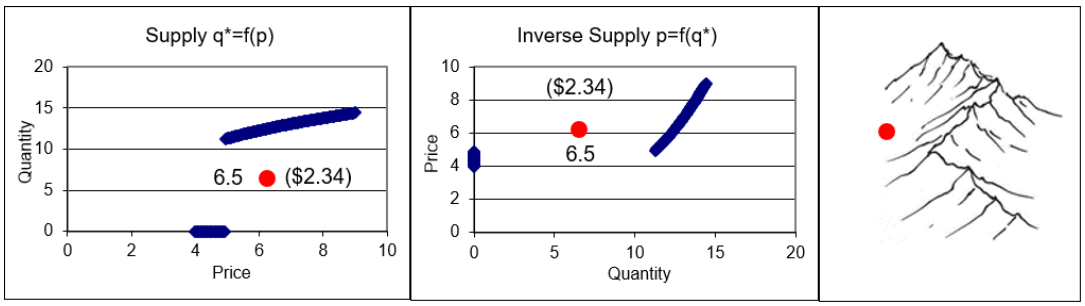


Figure 12.8: Points off the supply curve.
Source: DerivingSupply.xls!CS1.

Profits are maximized when you are on the supply curve. It is clear that the supply curve, like the demand curve, has a hidden third dimension—profit for supply and utility for demand. The right most panel shows the mountain

and how you approach the top at the optimal solution. The ridgeline connecting the mountain tops is the supply curve. Like the demand curve, points off the supply curve are associated with lower values of the objective function.

Notice how the point off the curve moves in a vertical fashion in the supply curve graph and horizontally on the inverse supply curve graph. This happens because price is constant (at $P = 6.25$). With the price on the x axis, points can be above or below the supply curve. Points off the inverse supply curve are to the right or left because P is on the y axis.

Finally, on the inverse supply curve, the inefficiency of being off the curve is obvious because output levels off the inverse supply curve means the firm is not choosing a point where $MR(= P) = MC$.

The Supply Curve has Parents

Like demand and cost curves, supply is derived from an optimization problem. Knowing where key relationships come from separates introductory from more advanced economics and is an important aspect of mastering the economic way of thinking.

The supply curve is a comparative statics analysis of the effects on optimal quantity as price changes, ceteris paribus.

Unlike the demand curve, the supply curve has a discontinuity because the firm will shut down if price falls below AVC. The supply curve depends critically on the firm's cost function. The inverse supply curve is simply MC above AVC and zero otherwise. The firm will choose that level of output where $MR(= P) = MC$ as long as $P > AVC$.

Like the demand curve, points off the supply curve are interpreted as inefficient solutions to the optimization problem. Although possible, no optimizing agent would choose a point off the supply (or demand) curve.

Exercises

1. What happens to the short run supply curve if wages rise? Explain. Use Word's Drawing Tools to create a graph depicting your answer.

2. What happens to the inverse short run supply curve if wages rise?

Explain. Use Word's Drawing Tools to create a graph depicting your answer.

3. What happens to the short run supply curve if the rental rate of capital increases? Explain.

4. What happens to the short run supply curve if the price (P) increases? Explain.

5. Suppose a firm is off its short run supply curve, but at a point where $MR = MC$. Use Word's Drawing Tools to a draw the profit function for this situation and label a point Z that meets the supposed conditions.

References

The epigraph comes from page 92 of the 1897 English translation of Augustin Cournot's *Researches into the Mathematical Principles of the Theory of Wealth*. This book was originally published in French in 1838. It is a remarkable work—truly far ahead of its time.

Cournot (pronounced coor–no) solves profit maximization problems for a variety of market structures, including monopoly, unlimited (today called perfect) competition, and intermediate cases of small numbers of firms. He uses derivatives and integrals with numerous supporting figures, including supply and demand with price on the x axis. Cournot was not bound by Marshall's convention of P on the y axis since Marshall's famous graphs of supply and demand would not appear until 1890.

The mathematical exposition was simply beyond the grasp of many readers in 1838 and the book languished in obscurity until the rise of mathematics in economics. You will hear Cournot's name again in the chapter on Game Theory.

Why was the spread of crops from the Fertile Crescent so rapid? The answer depends partly on that east-west axis of Eurasia.

Jared Diamond

12.3 Diffusion and Technical Change

The Theory of the Firm is a highly abstracted model of a real-world firm, yet there are fundamental ideas that can be applied to observed firm behavior. This section does exactly that, applying the Shutdown Rule to explain differing rates of diffusion of new technology.

The Shutdown Rule, $P < AVC$, says that firms will not produce when price is below average variable cost because profits are maximized (and losses minimized) by shutting down instead of producing at the best of the positive output choices (at $MR = MC$).

Diffusion of new technology is the process by which new methods of production are adopted by firms. The speed of diffusion is critical—the faster firms upgrade and modernize, the richer the society. We will see that some industries have fast and others slow diffusion with the Shutdown Rule playing a key role.

Setting the Table

Consider two thoughts that are both wrong:

1. Always upgrade to have the best equipment or to use "best practice" techniques.

2. Never throw working machinery away or abandon a process that can produce output.

The first statement is wrong because firms would always be replacing almost new machinery, tools, and plant to have the very latest equipment. The second statement is the polar opposite of the first: Now you keep using ancient machinery that was long ago superseded by better technology just because it is still functioning.

There has to be a middle ground between these two extremes and a logical way to determine when to replace equipment.

Consider these two words that are accepted as synonymous in common usage, but are different in the language of the specialized literature of diffusion:

1. *Outmoded*: machinery that is not the best at the time, but is still used.
2. *Obsolete*: machinery that is scrapped (thrown away) yet still functions.

Your phone is *outmoded* if it is not the latest, greatest available version. When you replace your phone with a new one, the old one becomes *obsolete*. At any point in time, a few people have the newest, fanciest model; the rest have versions of outmoded models still in use; and there are many obsolete models that are no longer being used. As time goes by, the newest model becomes outmoded and, eventually, obsolete.

The distinction between outmoded and obsolete sharpens our focus on this question: When does machinery go from being outmoded to obsolete?

Another important idea is *labor productivity*: the ability of labor to make output. This is measured in two ways, output per hour or labor required to produce one unit of output.

The output per hour version is simply the average product of labor, $\frac{q}{L}$. The bigger this ratio, the more productive is labor. You can take the reciprocal and ask, "How much labor is needed to make one unit of output?" This measure, called the *unit labor requirement*, gets smaller as labor productivity improves.

There are two ways of increasing labor productivity:

1. Better labor: increasing education.
2. Better machinery: technical (or technological) change.

Most people only think of the first way. More educated and skilled labor obviously will be more effective in translating labor input into output. But holding labor quality constant, if workers have better technology, such as computers or power tools, then labor productivity rises.

So, if you want to increase ditch digging productivity, you can improve the worker (think ditch digging classes) or you can improve the technology. A

worker with a shovel digs a ditch a lot faster than one without. But the explosion in productivity and output really occurs when you give the worker a backhoe.

But here's the curious thing, after backhoes are invented and brought online, if you look at the entire industry of ditch digging, you will see many different methods being used. Not everyone will instantly adopt the backhoe.

The question we are interested in boils down to explaining the *rate of diffusion*: how rapidly do the latest, best machinery and methods spread?

The mere existence of a new machine (e.g., a backhoe) is not enough to spur economy-wide increases in labor productivity. If the machine is not adopted rapidly, it will have little effect on the economy. We want fast diffusion so new methods spread quickly. This will boost productivity and economic growth.

The rate of diffusion is like adding a drop of red dye in a bucket of water. How rapidly will the water turn red? What factors affect the rate of diffusion? If we stir, the rate of diffusion rockets—how can we "stir" the economy to speed up diffusion?

It turns out that the rate of diffusion of technical change in an economy varies across industries and depends on specific characteristics. We are not searching for an unknown constant, but for the factors that explain wide variation in rates of diffusion—sometimes backhoes are rapidly adopted and other times not.

The rate of diffusion depends on whether machinery is determined to be outmoded versus obsolete. If machines are scrapped and replaced with the latest technology fairly quickly, then the rate of diffusion of technical change will be fast. If old technology is kept online and in production for a long time, then the rate of diffusion of technical change will be slow.

Before we see how the Shutdown Rule plays a critical role in deciding whether machinery is outmoded or obsolete, we review data used by W. E. G. Salter (1960) to support the claim that the rate of diffusion varies across industries. We also introduce a new graph that captures the idea of a distribution of methods or vintages of machinery.

On the Variation of Methods Used Across Industries

Salter presents data on a variety of goods. He focuses on the methods of production used at any point in time. It is quite obvious that there is always a mix of technologies being used. As new plants come online and new machinery is installed, older plants with older machinery remain in operation.

For example, Salter's Table 5, reproduced as Figure 12.9, shows a mix of technologies used in pig-iron production. Notice that the labor productivity of the best-practice plants (the latest technology) rises from 1911 to 1926. The industry average, however, lags behind because the latest technology is not immediately adopted by every manufacturer. The machine charged and cast method (the right most column) is the best technology, but even by 1926, 30.6% of the firms are not using it. These firms remain in operation with older technology. This slow diffusion hampers industry-wide labor productivity.

Table 5. Methods in use in the U.S. blast-furnace industry, selected years, 1911–26

	Gross tons of pig-iron produced per man-hour		Percentage of plants using the following methods		
Year	Best-practice plants	Industry average	Hand-charged and sand-cast %	Mixed types %	Machine charged and cast %
1911	0.313	0.14	50.0	22.7	27.3
1917	0.326	0.15	41.9	34.9	23.2
1919	0.328	0.14	42.0	28.0	30.0
1921	0.428	0.178	22.2	44.3	33.5
1923	0.462	0.213	20.7	39.7	39.6
1925	0.512	0.285	7.2	25.5	67.3
1926	0.573	0.296	6.1	24.5	69.4

Figure 12.9: Slow diffusion in pig-iron production.
Source: DiffusionTechChange.xls!Data.

Figure 12.10 (Salter's Table 6) focuses on the production of five-cent cigars. Salter keeps constant the quality and type of cigar, the five-cent variety, to focus on an apples-to-apples comparison of production methods. Because the measure of productivity is the labor required to make 1,000 five-cent cigars, the *lower* the hours required, the *greater* the labor productivity. The two-operator machine is the best practice, but three other methods are also used. Once again, the point is that a mix of methods are used and all of them combined determines industry-wide productivity.

Table 6. Approximate labour requirements per thousand five-cent cigars for different manufacturing methods, Unites States, 1936

Manufacturing methods in use in 1936	Man-hours per thousand cigars
Hand made	33.38
Machine bunched, hand rolled	27.38
Four-operator machine	15.96
Two-operator machine	11.94

Figure 12.10: Various methods of producing five-cent cigars.
Source: DiffusionTechChange.xls!Data.

Figure 12.11 offers a final example of Salter's point that an economy's labor productivity depends on the technology actually being utilized to make output. The *Range of all plants* column shows substantial variation in output from the best-practice firms to the least productive methods still being used. Notice that lower numbers are higher productivity because, as the title says, we are measuring "labour content per unit of output."

Table 8. Variation in labour content per unit of output in selected industries

			Man-hours per unit of output			Ratio of range to mean	
Industry, time and place	No. of plants	Unit of Output	Mean	Range of all plants	Range of middle 50% of plants	All plants	Middle 50%
Bricks, UK, 1947	17	1000 bricks	1.36	2.12–0.64	1.75–0.93	1.16	0.61
Houses, UK, 1948	160	Standard house	3080	4300–2150	3520–2630	0.66	0.29
Men's shoes, UK, 1949	12	Dozen pairs	9.70	12.34–7.30	11.02–8.53	0.53	0.26
Cement, US, 1935	60	100 barrels	46.7	86.0–25.3	57.9–39.3	1.30	0.40
Beet sugar, US, 1935	59	Ton of beet sliced	1.46	2.81–0.88	1.98–1.20	1.32	0.53
Sole leather, US, 1949	8	1000 lb.	48	–	61–39	–	0.47

Figure 12.11: Variation in labor productivity across six industries.
Source: DiffusionTechChange.xls!Data.

For bricks, with 17 plants in operation, the middle 50% range is from a best 0.93 hours to make 1,000 bricks to 1.75 hours. That is a huge difference and it is just the middle 50%. Take a moment to look at the ranges of the other products in Figure 12.11.

The *Ratio of range to mean* columns measure the rate of diffusion. If somehow every plant adopted the best-practice method, this ratio would be zero. Thus, houses and men's shoes are industries with much faster diffusion than the others.

Pig-iron, five-cent cigars, and products in Figure 12.11 are examples of a widespread phenomenon that was of great interest to Salter. The rate of diffusion of new technology is neither constant nor instantaneously fast. Salter wanted to know what diffusion depends on in the hope of manipulating it. After all, if there is a policy or lever we can pull to speed up diffusion, we would improve productivity and increase output.

A Graph is Born

Salter used an uncommon graph, an ordered histogram, to show how an industry incorporated various technologies in production.

Figure 12.12 (Salter's original Fig. 5) uses rectangles to indicate each method or vintage of machinery. We call this a *Salter graph*.

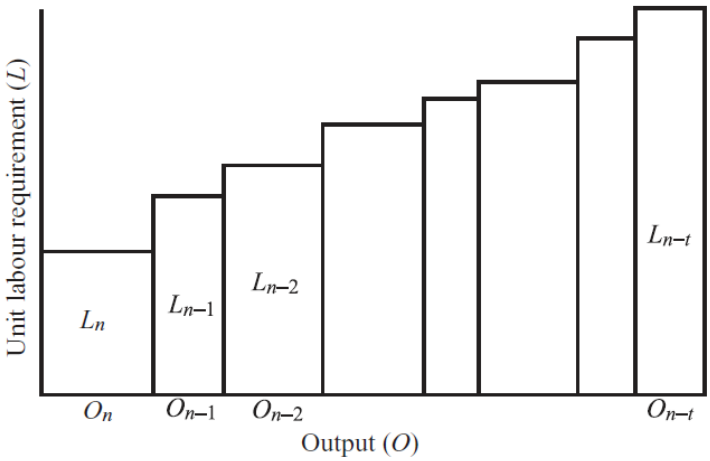


Figure 12.12: Salter graph of the mix of technologies.

The greater the base of each rectangle in Figure 12.12, the greater the share of the industry's output for that particular technology. So, in the middle of the graph, the wider rectangle has a bigger share of the output than the narrower one right next to it. The sum of lengths of the bases have to add up to 100% of the industry output.

The height of each rectangle tells you how much labor is needed to make one unit with that technology. The lower the height (because the y axis shows the labor required to make one unit of output), the greater the labor productivity for that technology.

The Salter graph has to have a stair-step structure because the rectangles are ordered according to when they came online. The oldest technology is to the right and the newest is to the left. The left-most rectangle is the best-practice technology at that time and all of the other rectangles are at different stages of outmodedness.

The Salter graph in Figure 12.12 is actually a single frame of a motion picture. As time goes by, and new techniques are invented and brought online, some of the right most rectangles will "fall over" and be replaced by a new shorter rectangle coming in from the left. Figure 12.13 shows a possibility for the next frame in the movie.

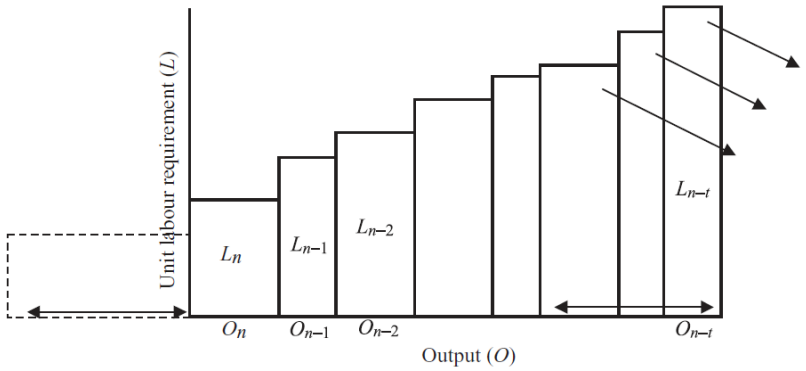


Figure 12.13: Salter graph as time goes by.

The base of the rectangle of the newest technology in Figure 12.13 equals the sum of the widths of the three rectangles representing obsolete technologies, which fall off the graph because they are no longer used.

The wider the base of the newest technology, the better in terms of fast diffusion of technological change and rapid increases in industry-wide productivity. If a new technology swept through an industry like wildfire, the Salter graph would show it as having a very long base, indicating it was producing a large share of industry output.

Another, less favorable possibility is that the newest technology has a small width. This would mean that few firms have adopted the best-practice method and industry-wide productivity will not improve by much. The industry will remain dominated by outmoded methods.

Consider the two Salter graphs in Figure 12.14 (Salter's original Fig. 12). They are enhanced by a strip in the middle, the height of which represents the industry average productivity.

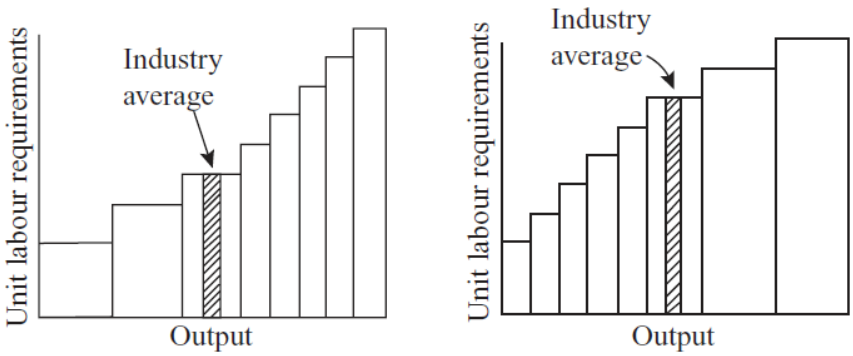


Figure 12.14: A comparison of two industries.

We would much prefer the industry on the left in Figure 12.14 because it has a lower industry average unit labor requirement, which means it has higher productivity. This is a result of much more rapid diffusion of newer, higher productivity technology.

The industry average shaded bar is a *weighted average* of all of the technologies in existence at any point in time. This statistic is the correct way to add up the rectangles with differing widths into a single measure of industry productivity. To understand how to do this, we turn to a concrete example in Excel.

STEP Open the Excel workbook *DiffusionTechChange.xls*, read the *Intro* sheet, then go to the *IndustryAverage* sheet to see how a weighted average is computed and how the Salter graph works.

Cells C9 and C10 show how two technologies contribute to the industry output. Initially, Methods A and B produce 50% of the total output. Because A (the superior, best-practice technology) requires only 1 hour of labor to make a unit of output, whereas B (an outmoded technology) requires 2 hours, the industry average productivity is 1.5 hours per unit of output.

STEP Click on the scroll bar a few times to increase A's share of total output to 90%. Notice how the Salter graph changes as you manipulate the scroll bar.

The Salter graph now shows A's share as a much wider rectangle (indicating much faster diffusion) and the red, industry (weighted) average rectangle is much shorter. Although the simple average does not change, the weighted average falls because more of the output is being generated by the more productive A technology. The weighted average computation (implemented in the formula for cell M10) is:

$$WeightedAverage = \frac{Output_A}{TotalOutput} UnitLReq_A + \frac{Output_B}{TotalOutput} UnitLReq_B$$

STEP Click on the scroll bar to decrease A's share of total output to 10%.

This time, the industry (weighted) average is 1.9 because only 10% of the output is produced with the best-practice technology. This would be an example of slow diffusion.

The contributions of each technology to industry output, weighted by the share of total output, is a good way to show how the rate of diffusion affects industry-wide productivity.

Having seen data that there is substantial variation in the rate of diffusion and that a Salter graph displays this variation, we are ready to explain why we see industries with mixes of technologies. We answer two questions:

1. Why is a machine that works sometimes kept (so it is outmoded) and other times scrapped (so it is obsolete)?

2. What determines the rate of diffusion of technical change?

1. Outmoded versus Obsolete?

We assume that new technologies are being constantly generated in all industries, but some are adopted more quickly. Why is that? Why are some factories and technologies quickly replaced while others remain online? Salter's work pointed to an easily overlooked element: the *cost structure* of the firms in an industry.

STEP Proceed to the *Output* sheet. The opening situation is depicted in Figure 12.15.

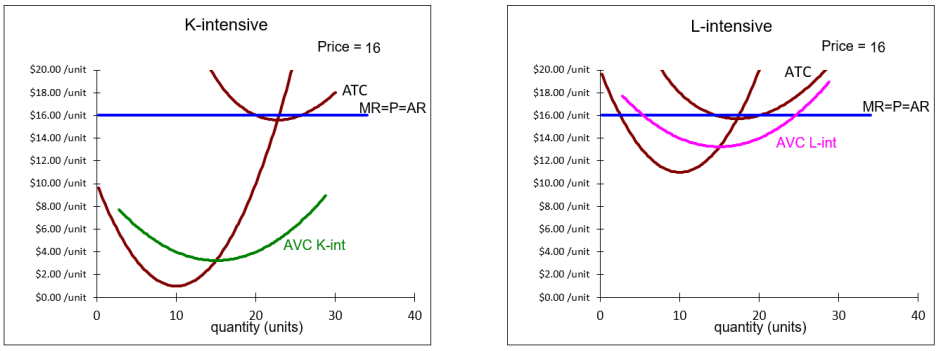


Figure 12.15: Explaining why machinery is outmoded versus obsolete. *Source: DiffusionTechChange.xls!Output.*

The graph shows two firms, one that is labor intensive and the other capital intensive. The capital intensive firm has a larger gap between ATC and AVC because it has higher fixed (capital) costs. The much lower AVC curve will prove to be critical.

Both firms in Figure 12.15 are earning small, but positive economic profits. As time goes by, however, new technologies are introduced and incorporated in newly built factories with shiny, modern equipment. The products from firms with the newest factories with their best-practice methods (the left-most rectangle in a Salter graph) can be made more cheaply so competitive pressure drives the price down.

STEP Click on the scroll bar to lower the price.

Since you know the Shutdown Rule, it is easy to see that the L-intensive firm will shut down first. As soon as you make $P < AVC$, the factory is obsolete and taken offline. The factory on the left will survive as an outmoded technology that is still in operation for much longer. You will have to keep driving the price down for much longer to see it shut its doors.

All firms use the same Shutdown Rule, but differing cost structures is what makes some factories stay in production while others close down.

So, to directly answer the question, Why is a machine that works sometimes kept (so it is outmoded) and other times scrapped (so it is obsolete)? Because the Shutdown Rule, $P < AVC$, determines the difference between outmoded and obsolete technology. Old plants that are kept online, using outmoded machines, operate in an environment in which profits may be negative, but

$P > AVC$. These plants will remain in operation as long as revenues cover variable costs. Once $P < AVC$, we know the machines will be scrapped and become obsolete as the factory is closed down.

2. What Does the Rate of Diffusion Depend On?

Figure 12.15 shows that the firm's cost structure is one of the factors which determine the rate of diffusion of technical change. Industries with capital intensive production and low variable costs will have slow rates of diffusion because plants and technologies will remain online until $P < AVC$.

Steel is a good example of such an industry. Old factories remain in production alongside modern mini-mills. The Salter graph looks like the right panel in Figure 12.14 and the cost structure is given by the left panel in Figure 12.15.

On the other hand, industries who produce in a way that labor is dominant and fixed costs are low will see rapid rates of diffusion of new methods. Legal services are a good example. Cost curves look like the right panel in Figure 12.15 so when new computers and information systems (such as LexisNexis) are developed, they are rapidly adopted and old ways are discarded. Thus, the Salter graph looks like the left panel in Figure 12.14.

Another factor affecting the rate of diffusion is the speed at which price falls. Competition among firms can be intense or muted. If, for example, the government protects an industry from foreign competition with trade barriers, preventing price from falling, the rate of diffusion of new technology and growth of labor productivity are retarded. This has certainly played a role in the rate of diffusion in the steel industry.

So, what determines the rate of diffusion of technical change? There are three factors:

1. New ideas and inventions from research and development (R&D): This is the creativity of the society. Curiosity and willingness to experiment produce a stream of better methods. The faster the flow, the better.

2. The cost structure of the firm: Capital intensive industry with high fixed and low variable costs retards diffusion of new technology. The new ideas are there, but the old ways stay online.

3. The speed at which price falls: If it is slow, we get slow diffusion. We want to encourage competition so price puts pressure on outmoded methods and drives them to be obsolete.

The first factor is the obvious one that everyone thinks of when explaining why technology affects labor productivity and economic growth. Innovation is the implementation of invention—new ideas are the raw material which expand the production function.

But Salter identified another crucial factor: Even if new technology exists, it will be mixed with existing technology and the rate at which it is adopted will depend on the Shutdown Rule. Highly capital intensive industries with low AVC will feel the drag of old technology for a long time because the gap between ATC and AVC will be great. Old methods will stay outmoded as long as $P > AVC$.

The Shutdown Rule compares average variable cost to price. Both matter. Low AVC will keep old methods around, but so will slow decline in P. Although economists usually defend free trade policies on the basis of comparative advantage, this analysis points to another reason for allowing foreign competition in domestic markets. As price is pushed down, firms are forced to modernize, taking old methods offline and investing in the newest technology. Steel tariffs are an example.

You might be confused about the claim that competition makes price fall as time goes by. It seems like inflation, prices rising, is the usual state of affairs. The explanation lies in the difference between real and nominal price.

In nominal terms, also known as current prices, the price of a light bulb is definitely higher today than 10 years ago and much higher than 100 years ago.

But in this application, the correct price to consider is the real price, in terms of actual input use. In real terms, the price of lighting is incredibly lower today. Figure 12.16, created by Nobel Prize winner William Nordhaus, tells an amazing story. In terms of the number of hours of work needed to buy 1,000 lumen hours, the price of light went from incredibly expensive for thousands of years to a free fall since the 1800s. In terms of input use, as technology improves, costs and, therefore, price of the output fall over time.

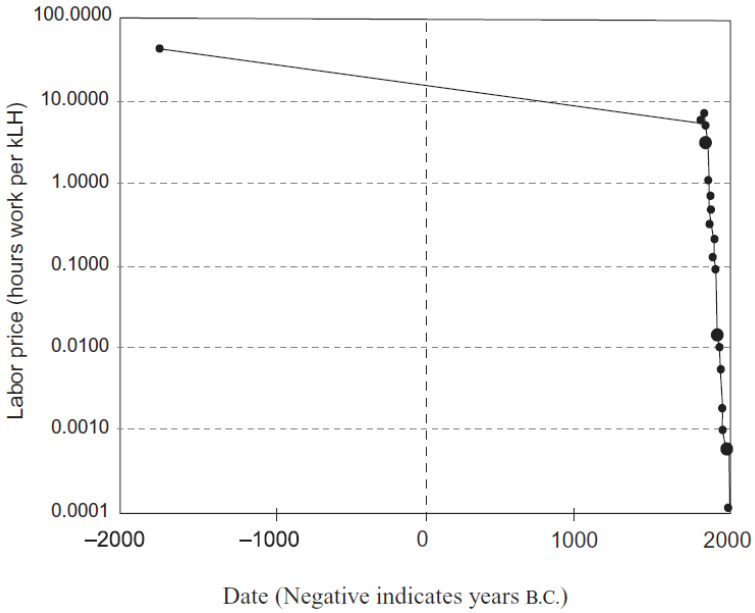


Figure 12.16: Labor price of light: 1750 b.c. to present.
Source: Bresnahan and Gordon (eds.), 1997, p. 54.

Nordhaus argues that "price indexes can capture the small, run-of-the-mill changes in economic activity, but revolutionary jumps in technology are simply ignored by the indexes" (Bresnahan and Gordon, eds., 1997, p. 55). Thus, the real price of lighting, in terms of the labor used, keeps falling and falling as time goes by.

It Is Diffusion, not Discovery, that Really Matters

Wilfred Edward Graham Salter was an Australian economist born in 1929. His promising career was tragically cut short when he died in 1963 after battling heart disease. His dissertation, finished in 1960, was published by Cambridge University Press as *Productivity and Technical Change* and was met with wide acclaim.

Salter was amazed by the ability of markets to incorporate new technology to increase output per person. He realized that scientific knowledge, technology "on the shelf," is not the only or even the most important driver of rapid growth. The new technology has to be implemented, actually used in production, and the faster it is adopted, the faster the economy grows.

Salter's primary contribution was in showing that the rate of diffusion varies tremendously and depends on the cost structures of firms. Industries with high fixed and low variable costs have large $ATC - AVC$ gaps that imply long time spans for outmoded technology.

We want nimble, adaptive firms and startups that challenge established titans. Replacing old with new machinery is necessary for rising productivity. Economies with ossified, rigid institutions are stagnant. There was a silver lining after Germany and Japan's factories were destroyed during World War II. The latest, greatest technology could be used to make all of an industry's output and productivity increased rapidly.

Exercises

1. Sometimes a best practice investment is quickly leapfrogged by newer technology. Google "fiber optic overinvestment" to see an example. Briefly describe what happened and cite at least one web source.

2. Automobile emissions requirements are stricter in Japan than in the United States (where many areas have no vehicle inspection at all). In both countries, newer cars pass inspection (if required) easily, but older cars are more likely to fail inspection and be removed from the operating car fleet. Draw hypothetical Salter graphs, with emissions on the y axis, for the car fleets of Japan and the United States that reflect the stricter emissions standards in Japan.

3. What happens to a late model year Toyota or Honda that has failed an emissions inspection in Japan and, therefore, cannot be used there? Google "Japan used engines" to find out. What effect does this have on the United States Salter graph that you drew above?

4. The National Highway and Traffic Safety Administration maintains a data base of car characteristics by model year. For miles per gallon (MPG) performance, they show the following:

2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004
39.2	37.3	37.2	36.3	36.1	34.8	32.7	33.1	32.1	31.2	30.6	30.3	30.5	29.9

Figure 12.17: MPG by year for US domestic passenger cars.
Source: one.nhtsa.gov/cafe_pic/CAFE_PIC_fleet_LIVE.html.

These data cannot be used to show a Salter graph (with MPG on the y axis) of the US car fleet. Why not? What additional information is needed?

References

The epigraph is from page 183 of Jared Diamond, *Guns, Germs, and Steel: The Fates of Human Societies* (W. W. Norton & Company, originally published in 1997). Diamond argues that geography determines historical development. It is not the people, but fortunate geographical circumstances that guaranteed that western Eurasian societies would become disproportionately powerful. It is geography that enabled the rapid diffusion of technology and knowledge. Diamond, like Salter, is concerned with a point that is easily missed—diffusion is more important than discovery. Visit www.pbs.org/gunsgermssteel for the documentary.

The primary source for the application in this chapter is W. E. G. Salter, *Productivity and Technical Change* (Cambridge University Press; 1st edition, 1960; 2nd edition, 1966; 1st paperback edition, 1969).

For more on technological change and the spread of new ideas, see Timothy F. Bresnahan and Robert J. Gordon, *The Economics of New Goods* (The University of Chicago Press, 1997) and David Warsh, *Knowledge and the Wealth of Nations: A Story of Economic Discovery* (W. W. Norton & Company, 2006).

Richard Preston's *American Steel* (Simon and Schuster, 1991) tells the story of a mini-mill in rural Indiana that uses German cold-casting technology. It is an entertaining tale of entrepreneurship—a billion dollar gamble—and an introduction to the exciting world of business management.

Chapter 13

Input Profit Maximization

Initial Solution

Deriving Demand for Labor

What is the Marginal Revenue Product* of Barry Bonds?

John Palmer in The Sports Economist Blog

13.1 Initial Solution

Recall that the firm's backbone is the production function. Inputs, or factors of production, typically labor (L) and capital (K) are used to make output, or product (q).

In previous chapters, we explored the firm's input cost minimization and output profit maximization problems. This chapter returns to the input side and works on the firm's third optimization problem: input profit maximization.

We continue working with a perfectly competitive (PC) firm, but we extend the assumption of perfect competition to input markets. Thus, not only is the firm one of many sellers of a perfectly homogeneous product with free entry and exit, it is also one of many buyers of labor and capital. Our firm is an output and input price taker.

This means that our PC firm only chooses the amount of input to hire, not how much to pay for it. If it has market power, then the firm not only determines how much to hire, but also gets to choose the input price. In this case, we say the firm has *monopsony power*.

While you have surely heard of monopoly, monopsony may be new to you. They are similar in that one is selling (monopoly) and the other buying (monopsony) and that means price (output or input) is no longer exogenous. A classic example is the only hospital in a small town hiring nurses. Another example is a big box retailer. Walmart is such a big buyer that they have monopsony power. They can negotiate with suppliers and extract cheaper prices from them. Notice that a firm can have both monopoly and monopsony power.

In a Labor Economics course, you study how firms can take advantage of the ability to set input prices to make greater profits. We assume this possibility

away and stay with a PC firm that takes the wage rate (w) and rental rate of capital (r) as given. Our PC firm is such a small buyer that it can hire as much L and K as it wants at the going w and r.

Setting Up the Problem

There are three parts to every optimization problem. Here is the framework for a PC firm.

1. *Goal*: Maximize profits (π), which equal total revenues minus total costs. To distinguish the input from the output side, we use the terms total revenue product (TRP) and total factor cost ($TFacC$). The idea is that labor and capital are used to make product that is sold so price times the number of units produced is the TRP.

2. *Endogenous variables*: labor and capital, in the long run; only L in the short run.

3. *Exogenous variables*: price (of the product, P), input prices (the wage rate and the rental rate of capital), and technology (parameters in the production function).

As usual, we will work with a Cobb-Douglas production function, with $\alpha > 0$, $\beta > 0$, and $\alpha + \beta < 1$.

$$q = AK^{\alpha}L^{\beta}$$

Revenues are the output price multiplied by the output produced, $TR = Pq$. We substitute the production function for q in TR to get total revenue product:

$$TRP = PAK^{\alpha}L^{\beta}$$

The units of TRP are dollars (just like total revenue). The "revenue product" language indicates that we are considering the amount of revenue ($) produced by the inputs.

The costs are simply the amounts spent on labor and capital, $wL + rK$. These are called total factor costs.

The firm chooses L and K to max profits.

$$\max_{L,K} \pi = PAK^{\alpha}L^{\beta} - (wL + rK)$$

Finding the Initial Solution

First the problem is solved using numerical methods, and then the analytical approach is used.

STEP Open the Excel workbook *InputProfitMax.xls* and read the *Intro* sheet,then go to the *TwoVar* sheet to see the problem implemented in Excel.

The sheet is named *TwoVar* because both inputs are choice variables, which means this is a long run profit maximization problem. As usual, the sheet is organized into the color-coded components of an optimization problem, with goal, endogenous, and exogenous cells.

STEP Read the description of the firm, a bakery, and scroll down to the endogenous variables.

On opening, the sheet has 500 hours of labor hired and 100 units of capital rented, yielding a profit of $936. Is this the best this firm can do? Cells B48 and B49 show the marginal revenue product of labor and marginal factor cost. By hiring one more hour of labor, revenues would rise by more than costs, so profits would increase. Clearly, therefore, this bakery is not optimizing.

STEP Run Solver to find the initial solution. Your screen should look like Figure 13.1.

Exogenous variables			
Price (P)	$	2.00	$/loaf of bread
Wage (w)	$	20.00	$/hr
Rental (r)	$	50.00	$/machine
alpha		0.20	
beta		0.75	
technology (A)		30	
Prod Fn (q)		19,086	loaves of bread

Endogenous Variables		
Labor (L)	1,431	hours
Capital (K)	153	machines

Goal			
Profit (π)	$	1,908.55	dollars
Revenue	$	38,171	dollars
Cost	$	36,262	dollars

Figure 13.1: The initial optimal solution.
Source: InputProfitMax.xls!TwoVar.

The firm hires roughly 1,431 hours of labor and rents 153 machines (but click on cells B34 and B35 to see more decimal places). This yields a maximum possible profit of just over $1,900.

Notice that the marginal revenue product and marginal factor cost cells are now exactly equal at $20/hour. This is no coincidence. The equimarginal condition for input profit maximization is that $MRP = MFC$. Since the firm is an input price taker, $MFC = w$ (just like $P = MR$ for a PC firm) so it is also true that $MRP = w$ at the optimal solution.

Finally, notice the breakdown of the firms revenues in rows 44 to 46. Labor's share (wL), capital's share (rK), and profits (whatever is left) add up to 100%. K and L's shares, 75% and 20% equal α and β. Is that a coincidence? No, that's a property of the Cobb-Douglas functional form. The exponent tells you the share of revenues that factor will receive.

We can also solve this problem via the analytical approach. We know the objective function and can substitute in each of the parameter values.

$$\max_{L,K} \pi = PAK^{\alpha}L^{\beta} - (wL + rK)$$

$$\max_{L,K} \pi = 2 * 30 * K^{0.2}L^{0.75} - (2L + 3K)$$

Next, we take derivatives with respect to L and K, set them equal to zero, and use algebra to solve the two equation system of first-order conditions.

$$\frac{\partial \pi}{\partial L} = 0.75 \cdot 2 \cdot 30\, K^{0.2} L^{-0.25} - 20 = 0$$

$$\frac{\partial \pi}{\partial K} = 0.2 \cdot 2 \cdot 30\, K^{-0.8} L^{0.75} - 50 = 0$$

We can move the 20 and 50 to the right hand side and this immediately reveals the equimarginal conditions: $MRP_L = w$ and $MRP_K = r$.

We solve the first equation for L and substitute it into the second equation to solve for optimal K. We use the rule that $(x^a)^b = x^{ab}$ to solve for L.

$$45K^{0.2}L^{-0.25} = 20$$

$$2.25K^{0.2} = L^{0.25}$$

$$\left[2.25K^{0.2} = L^{0.25}\right]^4$$

$$L = 2.25^4 K^{0.8}$$

Substitute the expression for L into the second first-order condition.

$$\begin{aligned}
&0.2 \cdot 2 \cdot 30\, K^{-0.8}\left[2.25^4 K^{0.8}\right]^{.0.75} = 50 \\
&12K^{-0.8}2.25^3 K^{0.6} = 50 \\
&K^{-0.2} = 0.365798 \\
&K^* = 152.6842
\end{aligned}$$

Compute optimal L from the expression for L.

$$L^* = 2.25^4 K^{0.8} = 2.25^4[152.6842]^{0.8} = 1431.414$$

Compute maximum profits.

$$\pi^* = 2{*}30{*}[152.6842]^{0.2}{*}[1431.414]^{0.75}{-}2{*}[1431.414]{-}3{*}[152.6842] = \$1908.55$$

This analytical solution is extremely close to Excel's solution. Practically speaking, as we would expect, the two solutions are the same.

The Short Run

A slightly different version of the firm's input profit maximization problem involves the short run when capital is not variable. By putting a bar over K, we highlight that capital is fixed.

$$\max_L \pi = PA\bar{K}^{\alpha}L^{\beta} - wL - r\bar{K})$$

We do the analytical solution first this time and in general form. There is only one derivative (since there is only one choice variable) and one first-order condition.

$$\begin{aligned}
&\frac{\partial \pi}{\partial L} = \beta PA\bar{K}^{\alpha}L^{\beta-1} - w = 0 \\
&\beta PA\bar{K}^{\alpha}L^{\beta-1} = w \\
&L^{\beta-1} = \frac{w}{\beta PA\bar{K}^{\alpha}} \\
&L^* = \left[\frac{w}{\beta PA\bar{K}^{\alpha}}\right]^{\frac{1}{\beta-1}}
\end{aligned}$$

STEP To see the numerical version of this problem, proceed to the *OneVar* sheet.

Notice that there is only one endogenous variable, L. Capital has been moved to the exogenous list because we are in the short run.

Notice also that there are two graphs. Each one can be used to represent the initial solution.

Below the graphs, you can see that the marginal revenue product of labor does not equal the wage. As you know, this means you need to run Solver because the firm is not optimizing.

STEP Run Solver to find the initial solution. Your screen should look like Figure 13.2.

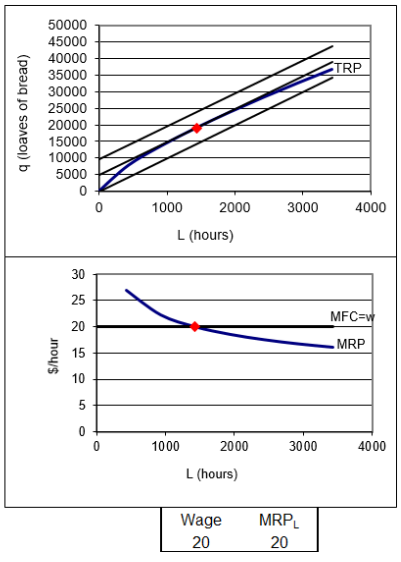
Exogenous variables			
Price (P)	$	2.00	$/loaf of bread
Wage (w)	$	20.00	$/hr
Rental (r)	$	50.00	$/machine
alpha		0.20	
beta		0.75	
technology (A)		30	
Capital (K)		153	machines
Prod Fn (Y)		19,085	loaves of bread

Endogenous Variables			
Labor (L)		1,431	hours

Goal			
Profit (π)	$	1,909	dollars
Revenue	$	38,171	dollars
Cost	$	36,262	dollars

Calculating Distribution of Revenues

		Dollar Payments	Percentage of Total
Labor Share	$	28,628	75%
Capital Share	$	7,634	20%
Profit Share	$	1,909	5%



Wage	MRP_L
20	20

Figure 13.2: The initial optimal solution in the short run.
Source: InputProfitMax.xls!OneVar.

The bottom graph shows that the optimal labor use can be found where the marginal revenue product of labor (the curve) equals the wage (at \$20/hr). This is the canonical graph for the input side profit maximization problem. Like $MR = MC$ on the output side, the intersection of the two marginal relationships instantly reveals the optimal solution.

The top graph is a different way of viewing the exact same problem. It is using the production function as a constraint (the *TRP* curve) and three representative *isoprofit* lines are displayed. Each isoprofit line shows the combination of L and q that gives the same profit. The firm is trying to get on the highest isoprofit (to the northwest) while meeting the constraint. It can roll on the *TRP* curve (like it rolled on the isoquant) until it hits an isoprofit line that is tangent to the *TRP*.

The constrained optimization problem can be written like this:

$$\max_{L,q} \pi = Pq - wL - r\bar{K}$$
$$\text{s.t. } q = A\bar{K}^{\alpha}L^{\beta}$$

The Lagrangean method could be applied to solve this problem. Naturally, the exact same solution is obtained if we use the Lagrangean or the more common approach of directly substituting the constraint (the production function) into the revenue function.

Suppose we wanted to check if the analytical and numerical results are the same. We need to evaluate the expression for optimal L at the parameter values in the *OneVar* sheet.

The expression is complicated enough that entering it in a cell as you would write it is a bad idea. The parentheses are likely to cause confusion. It is better to create *houses* for each part then fill them in. Here's how.

STEP Watch this short video on how to enter a complicated formula in Excel: vimeo.com/415967747.

Entering parentheses as pairs, is a good habit to develop when working in a spreadsheet. It is easy to make an order of operations mistake or get mismatching parentheses if you try to enter the formula like you would on a piece of paper.

STEP Enter the formula in cell M28 (just like in the video) to practice building houses in formulas in Excel.

In so doing, you confirm that the analytical and numerical methods yield substantially the same answer.

Another Short Run Production Function

A Cobb-Douglas production function has many advantages, including that the sum of exponents reveals whether returns to scale are increasing, constant, or decreasing if they are greater, equal, or less than one. However, once the exponents are set, the function can only exhibit those returns to scale.

Likewise, in the short run, with K fixed, our Cobb-Douglas functional form showed the Law of Diminishing Returns because $\beta = 0.75$. A more flexible functional form would enable production to have increasing and diminishing returns as more labor is added.

Like the cubic polynomial we used for the total cost function, a cubic functional form can give us an S-shaped TRP curve.

$$TRP = aL^3 + bL^2 + cL$$

STEP Proceed to the *Graphs* sheet to see this functional form implemented in a set of four graphs that can be used to represent the firm's input profit maximization problem (Figure 13.3).

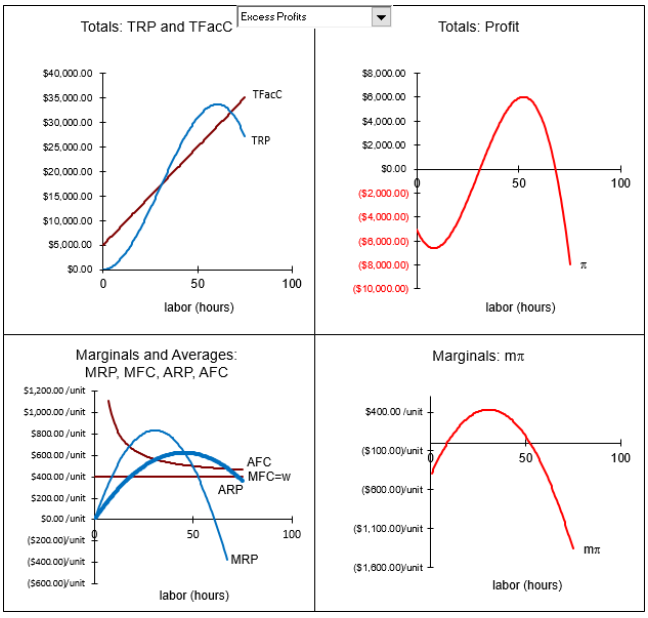


Figure 13.3: Four graphs for input profit maximization.
Source: InputProfitMax.xls!Graphs.

It is striking that these graphs mirror the four graphs we used to describe the firm's output side profit maximization problem. The two top graphs show total revenue and total cost on the top left, along with total profits on the top right. The bottom graphs display a series of marginal and average curves on the bottom left and marginal profit on the bottom right.

If you look carefully, you will notice that things are switched around a bit. Instead of total cost being a curve (as it is on the output side), it is a straight line because total factor cost on the input side in the short run is $wL + r\bar{K}$. On the other hand, total revenue product (so named to distinguish it from total revenue on the output side) is a curve (instead of a straight line).

Unlike the canonical output side profit maximization graph with U-shaped MC, ATC, and AVC curves and a horizontal $P = MR$ line, the bottom left graph has a horizontal MFC line and the MRP and ARP functions are curves and they are upside down.

But there are also key similarities. The equimarginal rule is in play: $MFC = MRP$ reveals the labor use that maximizes profits. Also, a rectangle of $(ARP - AFC)L$ gives an area that is equal to profits. The length of the profit rectangle ranges from zero to the chosen amount of labor hired. The height is the difference between average revenue product, ARP, and average factor cost, AFC. The area of this rectangle is profit because $ARP - AFC$ is profit per hour so multiplying by L, measured in hours, yields profits. Another way to think about this is that multiplying L by ARP yields total revenues (since $L * TRP/L = TRP$) and multiplying L by AFC gives total costs (since $L * TFacC/L = TFacC$). Subtracting the total cost rectangle from the total revenue rectangle leaves the profit rectangle.

Another similarity between output and input profit maximization is that the firm has the same four profit positions.

STEP In the *Graphs* sheet, click on the pull down menu (near cell P4) and cycle through all of the profit positions.

As with the output side, the shock is output price. As it falls, so do maximum profits.

The *Neg Profits, Cont Prod* and *Neg Profits, Shutdown* options show that the firm will shut down when the $w > ARP$. This is analogous to the $P < AVC$ Shutdown Rule. Keep your eye on the total profits in the top right graph

to see that the story is the same—the firm is deciding whether the negative profit at best of the positive levels of L is better than hiring no L at all.

The connection between input and output is simple. The firm shuts down when $w > ARP$ which we can multiply by L to give $wL > TRP$. But wL and TRP are TVC and TR on the output side. Divide both by q and we get $AVC > P$, which is the same as $P < AVC$, the usual output side Shutdown Rule. In addition, the $wL > TRP$ version of the Shutdown Rule supports the claim that revenues must cover variable costs for a firm to produce.

Input Profit Maximization Highlights

At this point, you might be suffering from repetitive stress syndrome—we seem to be going over and over the same ideas. That is an important level to attain in mastering the economic way of thinking. The body of knowledge in economics is grounded in a core methodology of optimization and comparative statics. The framework is used over and over and over again.

Like every optimization problem, the input side profit maximization problem can be organized into a goal, endogenous, and exogenous variables. This problem has a canonical graph (with MFC and MRP as the key elements) and an equimarginal rule $MFC = MRP$.

Because the firm is an input price taker, $MFC = w$. This means that every additional hour of labor adds w to total cost. If the firm was a monopsony, this would not be true and the optimization problem would be more complicated.

Finally, because the input profit maximization problem is the flip side of the output side profit maximization problem, it should not be surprising that we can represent the initial solution with a set of four graphs. The parallelism carries through all the way to the Shutdown Rule, where $w > ARP$ is equivalent to $P < AVC$. We will stress the connections between input and output side again in the next chapter.

Exercises

1. Use the *TwoVar* sheet to compute the long run beta elasticity of L^* from beta = 0.75 to beta = 0.74. Show your work.

2. In the *Q&A* sheet, question 4 asks you to find short run beta elasticity of L^* from beta = 0.75 to beta 0.74. The *InputProfitMaxA.doc* file in the *Answers* folder shows that the answer is about 28. Explain why the short run elasticity (which is admittedly quite large) is much smaller than the long run elasticity that you computed in the previous question.

3. Use Excel to set up and solve (with Solver, of course) the constrained version of the input profit maximization problem in the *OneVar* sheet. Take a screenshot of your solution (including the constraint cell) and paste it in your Word document.

4. In the *Graphs* sheet, select the *Neg Profits, Shutdown* case. Does the top, right graph support the $w > ARP$ Shutdown Rule? Explain.

References

The epigraph, from John Palmer at thesportseconomist.com/what-is-the-marginal-revenue-product-of-barry-bonds, points to two avenues for further reading: sports economics and blogs.

The worlds of economics and sports are increasingly intertwined. There are courses, conferences, and journals dedicated to the economics of sports. For a classic paper on baseball, see Simon Rottenberg's "The Baseball Players' Labor Market," *The Journal of Political Economy*, Vol. 64, No. 3 (June, 1956), pp. 242–258, www.jstor.org/stable/1825886.

There are, of course, many blogs dedicated to economics. The marginalrevolution.com and cafehayek.com are often informative and entertaining. For macroeconomics, see Greg Mankiw at gregmankiw.blogspot.com and Brad DeLong at delong.typepad.com. John Cochrane will give you a free market perspective at johnhcochrane.blogspot.com—plus, *The Grumpy Economist* is a great name for a blog.

> To be sure, we are living in a dessert age. We want things to be sweet; too many of us work to live and live to be happy. Nothing wrong with that; it just does not promote high productivity. You want high productivity? Then you should live to work and get happiness as a by-product.
>
> David Landes

13.2 Deriving Demand for Labor

A profit-maximizing firm with Cobb-Douglas technology and given prices in all markets (P, w, and r) in the short run can be modeled as solving the following optimization problem:

$$\max_L \pi = PA\bar{K}^{\alpha}L^{\beta} - wL - r\bar{K})$$

The previous section found the initial solution for this problem. This section is devoted to comparative statics analysis. How will this firm respond to a change in one of its exogenous variables, ceteris paribus?

Although there are several exogenous variables from which to choose, the responsiveness of optimal L to a change in the wage is of utmost importance. This comparative statics analysis will give us the short run demand for labor.

After deriving the demand for labor in the short run, we will examine the long run demand for labor. A comparison of short and long run wage elasticities of labor reveals that labor demand is more responsive in the long run. We then explore how changes in P affect L^*.

Demand for Labor in the Short Run

We begin with numerical methods for a comparative statics analysis of a change in the wage (also called the wage rate is measured in \$/hr).

STEP Open the Excel workbook *DerivingDemandL.xls* and read the *Intro* sheet, then go to the *OneVar* sheet.

The layout is the same as the *InputProfitMax.xls* workbook in the previous section. It is clear from the graphs and the equivalence of wage and *MRP* below the graphs that the firm is at its optimal solution. The yellow-backgrounded cell, the wage rate, is the shock variable on which we will focus.

STEP Change the wage in the *OneVar* sheet to \$19/hr from the initial value of \$20/hr.

It is difficult to see anything in the top graph, however, the isoprofit line is no longer tangent to the *TRP*. The bottom graph clearly shows that the red diamond (at $L = 1431$ hours) has a marginal revenue product greater than the marginal factor cost (equal to the wage). Cells H40 and I40 show that the wage is less than *MRP*.

STEP Since the firm is no longer optimizing, run Solver to find the new optimal solution.

You will find that, to maximize profits, the firm will hire 1757 hours when the wage falls to \$19/hr, ceteris paribus. At this level of labor use, the marginal revenue product once again equals the marginal factor cost.

Although we have only two data points, it should be clear that the firm will hire that amount of labor where the marginal revenue product equals the wage, in the short run. This means that *the marginal revenue product curve is the firm's (inverse) demand for labor curve.* Quote the firm a wage and it will look to its *MRP* curve to decide how much labor to hire.

We have two points on the demand for labor curve; at $w = \$20$/hr, $L^* = 1431$ hours and at $w = \$19$/hr, $L^* = 1757$ hours. Can we pick more points off of the demand for labor curve?

STEP Set the initial wage back to \$20/hr and use the Comparative Statics Wizard to apply five \$1/hr decreases in the wage. Create charts of the demand for labor and the inverse demand for labor.

Your results should look like those in the *CS1* sheet. The CSWiz output makes common sense. As the wage drops, the firm hires more labor. Look also at the objective function—as wage falls, maximum profits are rising. The key idea here is that firm hiring decisions are driven by profit maximization. The reason why L increases as w falls is that this response is profit maximizing.

Like demand curves in the Theory of Consumer Behavior, the price—the wage in this case—can be placed on the x or y axis. The two displays use the same information and convey the same message.

We can also derive the short run demand for labor via analytical methods. This problem was presented in the previous section. For your convenience, it is repeated below.

We need to leave w as a variable, but for maximum generality we solve for L^* as a function of all parameters.

$$\max_L \pi = PA\bar{K}^\alpha L^\beta - wL - r\bar{K})$$

We take the derivative with respect to L, set it equal to zero, and solve for L^*.

$$\begin{aligned}
\frac{\partial \pi}{\partial L} &= \beta PA\bar{K}^\alpha L^{\beta-1} - w = 0 \\
\beta PA\bar{K}^\alpha L^{\beta-1} &= w \\
L^{\beta-1} &= \frac{w}{\beta PA\bar{K}^\alpha} \\
L^* &= \left[\frac{w}{\beta PA\bar{K}^\alpha}\right]^{\frac{1}{\beta-1}}
\end{aligned}$$

This expression is the demand curve for labor. If we substitute in values for all exogenous variables except w, we can plot L^* as a function of w, ceteris paribus.

Do the numerical methods based on the CSWiz add-in agree with the analytical derivation of the demand for labor?

STEP In the *CS1* sheet, click on cell C16. This is Solver's answer for L^* when the wage is \$20/hr.

Do not be misled by all of the decimal places. That is false precision.

STEP Click on cell E26. It displays L^* when the wage is \$20/hr based on the reduced-form solution.

Do not be misled by the number displayed in cell E26. This is Excel's display for the formula entered into that cell. Excel's memory has a different number.

STEP Widen column E to see more decimal places.

We proceed slowly because things can get confusing here. Consider this hierarchy of truth:

1. Solver is giving a number close to the exact right answer in cell C16.
2. Excel is representing the exact right answer as a decimal in cell E26.
3. The exact right answer is $\frac{w}{\beta PA\bar{K}^{\alpha}}^{\frac{1}{\beta-1}}$ evaluated at $w = \$20/\text{hr}$, along with the other parameter values.

STEP To see that E26 is not the exact answer, make column E very wide, then select cell E26 and click Excel's *Increase Decimal* button repeatedly.

You will see that, eventually, Excel will start reporting zeroes. Excel has finite memory and, therefore, it cannot compute an infinite number of decimal places for the exact answer. The decimal representation of the exact answer stored in Excel's memory is not the exact answer.

To be clear, Excel can display the exact answer if it is an integer or fraction that can be represented with finite memory. For example, $\frac{x}{7}$, evaluated at $x = 14$ is 2 so, no problem for Excel. If 2 is the answer, Excel has it exactly right. Evaluating at $x = 1$ means there is no decimal representation with a finite number of digits. Excel cannot display the exact answer in this case. Enter $= 1/7$ in a cell, widen the column, and click the *Increase Decimal* button repeatedly to see that Excel eventually starts showing zeroes.

Thus, neither E26 nor C16 is the exact answer. They are both so close to the answer, however, that we can say they "substantially agree" and are correct.

We can also use the analytical approach to reinforce the idea that the short-run (inverse) demand for labor is the marginal revenue product of labor.

The first-order condition gives the equimarginal rule.

$$\frac{d\pi}{dL} = \beta PA\bar{K}^{\alpha}L^{\beta-1} = w$$

The term on the left is the *MRP*. Evaluating the $\beta PA\bar{K}^{\alpha}$ portion at their initial values gives 123.0187 (as shown in cell K26 of the *CS1* sheet). Thus, $MRP = 123.0187L^{\beta-1}$ and at $\beta = 0.75$, $MRP = 123.0187L^{0.25}$.

The *CS1* sheet has an inverse demand for labor chart. Is the relationship in this chart the same as the *MRP* function that we just found? Let's find out. By finding the function that fits the data in the inverse demand for labor chart, we can compare this relationship to the *MRP* function.

STEP Right-click on the series in the inverse demand for labor chart and select the *Add Trendline* option. Select the *Power* fit, scroll down and check the *Display equation on chart* option. Click OK. Move the equation (if needed) and increase the font size to see it better. Scroll right to see what your chart should look like.

The answer is clear: The fitted curve that reveals the function for the inverse demand curve for labor is the marginal revenue product of labor curve. The fitted curve's coefficient and exponent are almost exactly that of the *MRP*.

Next, we turn our attention to the wage elasticity of labor demand. We can compute the elasticity at a point or from one point to another. We do the former below and leave the latter as an exercise question.

Elasticity at a point begins by finding the derivative of the reduced-form expression. We substitute in the known value for $\beta P A \bar{K}^{\alpha} = 123.0187$ in the denominator and $\beta = 0.75$ in the exponent.

$$L^* = (\frac{w}{\beta P A \bar{K}^{\alpha}})^{\frac{1}{\beta - 1}} = (\frac{w}{123.0187})^{\frac{1}{0.75-1}} = (\frac{w}{123.0187})^{-4}$$

To take the derivative with respect to w, we isolate w.

$$L^* = (\frac{w}{123.0187})^{-4} = \frac{w^{-4}}{123.0187^{-4}} = (\frac{1}{123.0187^{-4}})w^{-4}$$

Now we can apply our usual derivative rule, moving the exponent to the front and subtracting one from it.

$$\frac{dL^*}{dw} = -4(\frac{1}{123.0187^{-4}})w^{-5}$$

This expression is merely the slope or instantaneous rate of change of optimal labor hired as a function of the wage. To find the elasticity, we must multiply the derivative by the ratio w/L.

$$\frac{dL^*}{dw}\frac{w}{L} = -4(\frac{1}{123.0187^{-4}})w^{-5}\frac{w}{L}$$

But we have an expression for L, so we substitute it in.

$$\frac{dL^*}{dw}\frac{w}{L} = -4(\frac{1}{123.0187^{-4}})w^{-5}\frac{w}{(\frac{1}{123.0187^{-4}})w^{-4}}$$

The 123.0187^{-4} terms cancel. And w^{-5} times w in the numerator is w^{-4} so that cancels with w^{-4} in the denominator. We are left with this.

$$\frac{dL^*}{dw}\frac{w}{L} = -4$$

As has happened before (remember the price and income and cross price elasticity of demand?), the Cobb-Douglas functional form produces a constant wage elasticity of short run labor demand.

This elasticity value says that labor demand is extremely responsive to changes in the wage. We would not expect to find such a large wage elasticity of short-run labor demand in the real world. For a Cobb-Douglas production function, the elasticity is driven by the value of beta. If we had left β in the expression for optimal L instead of using 0.75 (see the first two exercise questions), we would get this expression for the wage elasticity of labor demand:

$$\frac{dL^*}{dw}\frac{w}{L} = \frac{1}{\beta - 1}$$

If we compute the elasticity from one point to another, say from a wage of \$20/hr to \$19/hour (see exercise question 3), we will get a different answer than -4. That makes sense since we know that L^* is non linear in w. As the change in the wage approaches zero, the elasticity computed from one point to another approaches -4.

13.2.1 Demand for Labor in the Long Run

If we relax the assumption that capital is fixed, we change the firm's planning horizon from short to long run. The *TwoVar* sheet implements the firm's long run input profit maximization problem. There are two endogenous variables, labor and capital, and no fixed factors of production.

STEP To derive the firm's long run demand for labor, use the Comparative Statics Wizard from the *TwoVar* sheet. As you did in the short run analysis, apply \$1 decreases in the wage.

Your results should show labor use rising as wage falls, just as in the short run. But what about the elasticity—is it the same in the short and long run?

STEP Use your CSWiz results to compute the wage elasticity of labor demand from a wage of \$20/hr to \$19/hr. Is it close to -4, the point elasticity at w=\$20/hr?

The *CSCompared* sheet is similar, but not the same as your results. It shocks wage by \$1/hr increments in the short and long run.

The difference in the elasticity is dramatic—labor demand is incredibly responsive in the long compared to the short run. The elasticity almost triples, from -3.5 to almost -11. You should find the same result with your CSWiz data for a wage decrease—the long run elasticity is much higher (in absolute value) than in the short run. What is going on?

Figure 13.4 provides an answer to this question. The movement from point A to B is the short run response for a \$1/hr wage increase. As the short run results in the *CSCompared* sheet show, when the wage rises from \$20/hr to \$21/hr, L^* falls from roughly 1,431 hours to 1,178 hours.

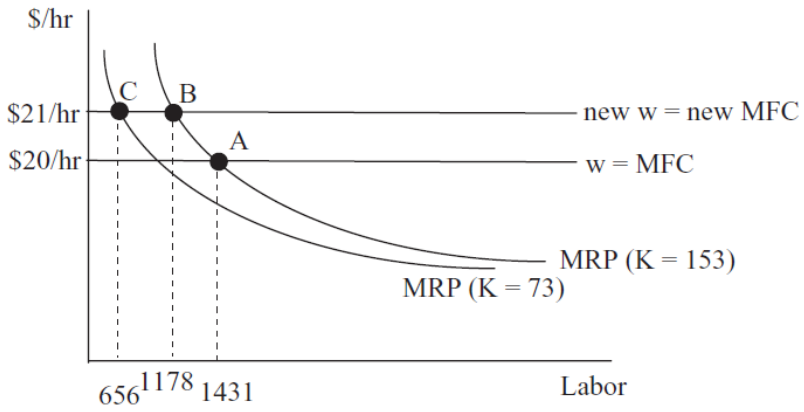


Figure 13.4: Why L^* is more responsive to Δw in the long than short run.

In the short run, capital stays fixed and the firm moves along its marginal revenue product curve (which as we already know is the firm's short run demand for labor) as the wage changes. The $K = 153$ in the parentheses signals that this is the value of K for this MRP schedule.

In the long run, however, the adjustment is different. The data in the *CSCompared* sheet show clearly that the firm will change both labor and capital as the wage rises. Notice that capital falls from 153 machines to 73 machines as the wage rises from \$20/hr to \$21/hr.

This change in capital shifts labor's marginal revenue product curve. As shown in Figure 13.4, the firm's long run response to the change in the wage is from A to C, not simply A to B. It decreases labor use as it moves along the initial *MRP* and then again when *MRP* shifts as K falls. This is the reason why the wage elasticity of labor demand is more responsive in the long run.

Figure 13.5 shows the firm's long run demand for labor and that it is no longer the *MRP* curve. Because capital falls as wage rises, leading to a further decrease in labor hired, the firm is much more responsive to changes in the wage.

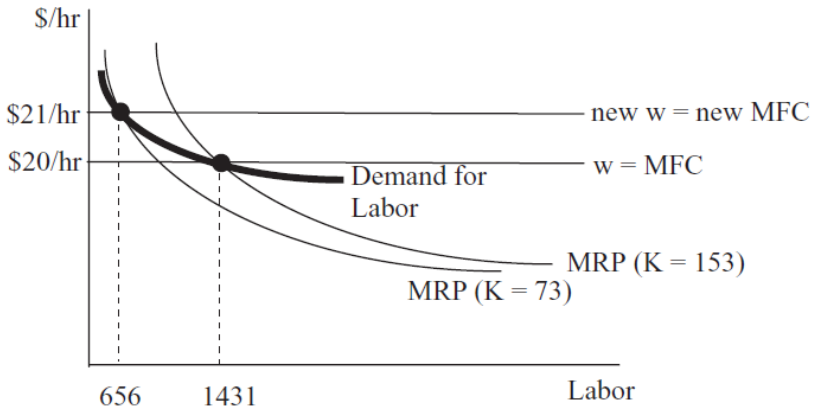


Figure 13.5: The long run demand for labor.

It is clear that the inverse labor demand curve shown in Figure 13.4 is flatter in the long run than the *MRP* curve (which is the short run inverse demand for labor). A wage decrease would stimulate more labor hired in the long than short run because K would rise in the long run.

The Shutdown Rule and the Demand Curve for Labor

Recall that, on the output side, the supply curve is the MC curve when $P > AVC$.If $P < AVC$ where $MR = MC$, then the firm ignores this marginal signal (which is the top of a local profit hill) and shuts down ($q = 0$).

The supply curve has a tail where the quantity supplied is zero when the price falls below average variable cost.

There is a similar tail, with $L = 0$, on the demand curve for labor. The previous section showed that if $w > ARP$, the firm will shut down, hiring no labor and producing no output.

STEP Proceed to the *Graphs* sheet to quickly review this concept. Use the pull down menu to change the firm's output price and place the firm in any of the four profit positions. Select *Neg Profits, Shutdown* to see that the firm will shut down when P is so low that it shifts ARP down so much that $w > ARP$. This is analogous to the $P < AVC$ Shutdown Rule.

The Shutdown Rule means that we have to change our definition of the demand curve for labor to get it exactly right. In the short run, the inverse demand curve is the MRP curve, as long as $w > ARP$; otherwise it is zero, as shown in Figure 13.6.

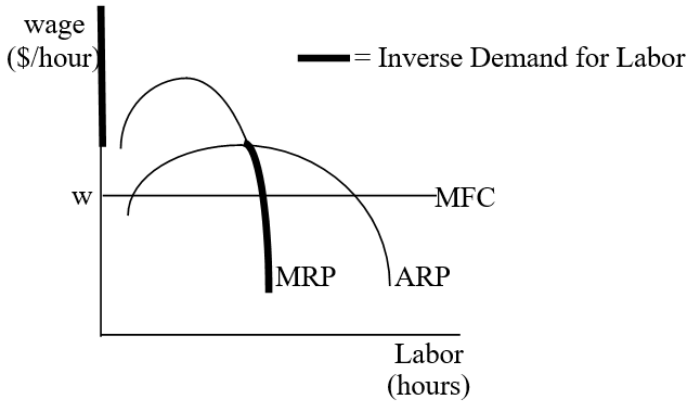


Figure 13.6: The short run inverse demand for labor.

The Shutdown Rule is usually presented from the output side as $P < AVC$. This version of the rule is perfectly compatible with the input side version of the shutdown rule, $w > ARPL$. Either wage increases or output price decreases can trigger a shutdown.

In Figure 13.6, it is easy to see what is happening when wage increases—the horizontal MFC line shifts up and it rises above ARP, the firm shuts down. What is happening on the output side? Remember that as wage rises, cost curves on the output side shift up. At the precise point at which a higher

wage triggers the decision to not hire any labor, the AVC curve will have shifted above P and the firm will decide to not produce any output.

The same story is at work when P falls. On the output side, it easy to see that when the horizontal $P = MR$ line falls below AVC, the firm shuts down. What is happening on the input side? As P falls, the MRP and ARP curves in Figure 13.6 shift down. At the precise moment when P falls below AVC and the firm decides to produce no output, the ARP shift below the horizontal wage line in Figure 13.6 and the firm will decide to hire no labor.

Demand for Labor Depends on *P*

Another comparative statics analysis for input profit maximization revolves around the effect that P has on L^*. This shows how the demand for labor is a derived demand from the desirability of the product. In other words, the stronger the demand for the product, the greater the demand for labor.

Suppose demand for bread rises in our Excel workbook. This increases P, ceteris paribus. What happens to L? We explain the short run response here and leave the long run for exercise questions 4 and 5.

STEP Return to the *OneVar* sheet. Return the wage to \$20/hr. Run Solver.

Instead of simply changing P and running Solver again, we want to see what effect P has on the graphs that show the initial solution.

STEP Change P to \$2.10 and look carefully at the charts.

It is difficult to see that the TRP curve has changed so that it is no longer tangent to the isoprofit line, but the bottom chart clearly shows that the initial solution is no longer optimal. What happened?

From our analytical work, we know that $MRP = \beta P A \bar{K}^{\alpha} L^{\beta-1}$ so it is clear that an increase in P will shift the MRP curve up. That is what you are seeing in the bottom graph on the *OneVar* sheet. Return P to \$2/unit to see that MFC stays constant (w remains unchanged), but MRP is moving.

STEP With P=\$2.10, run Solver. What happens to L^*?

Not surprisingly, the firm wants to hire more labor. The reason is that the MRP curve shifts and a new solution is found where the new $MRP = w$. Labor cost and productivity are unchanged, but the demand for labor is affected by consumer's desire for the product (expressed through the P). We say that that demand for labor is a *derived demand*—the firm's need for labor (and other inputs) comes from the fact that it has customers who want its product.

Figure 13.7 shows what happens as you increase the product price. If the demand for a firm's output is high, the price will be high, and this will induce an increased demand (shift) for labor.

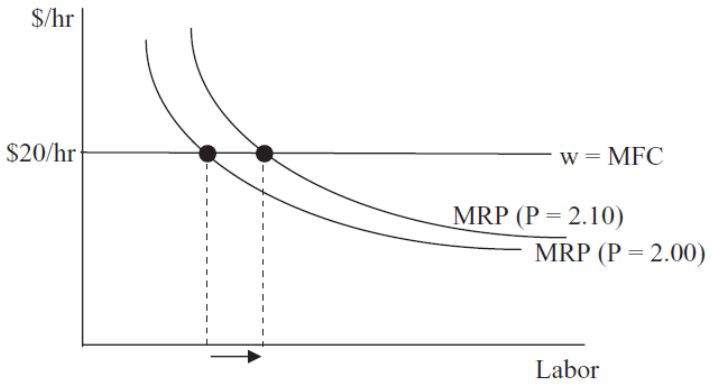


Figure 13.7: Demand for L is a derived demand.

It is easy to see that labor is a derived demand by considering professional sports. Pro athletes in major sports make a lot of money because they are in high demand. Sports teams know that the price of the good they produce (including broadcast and streaming revenue) is high. The output side is most definitely reflected in the input side via the product price.

Marginal Productivity Theory of Distribution

The input side profit maximization problem can be used to examine the distribution of firm revenues. The basic idea is that shares are a function of an input's productivity: The more productive the input, the greater its share.

STEP From the *TwoVar* sheet, run a comparative statics experiment that changes the exponent on labor from 0.75 to 0.755 (5 shocks of 0.001). In the endogenous variables input box, be sure to track not only L and K, but also the shares received in cells C44:C46.

Check your results with the *CS3* sheet. The *CS2* sheet has the outcome of a change in alpha, the exponent on capital. It explains how "large" shocks of, say, 0.1 will cause catastrophic failure as $\alpha + \beta$ approaches +1. This is why the change in beta so small—to stay away from the singularity.

By increasing the exponent on labor in the Cobb-Douglas production function, labor's productivity rises. In other words, labor can make more output, ceteris paribus, as the exponent on labor increases. The firm maximizes profit by using more labor and labor's share of firm revenues rises.

The CSWiz data show that we can immediately determine the percentage share of revenues gained by each input by the input's exponent in the production function. Although a different production function may not have this simple short-cut to determine the percentage share of revenues accruing to each input, it remains true that an input's share will depend on its marginal productivity.

Whereas algebraic convenience and simplicity are often invoked as a rationale for utilizing the Cobb-Douglas functional form, in the case of factor shares, a strong empirical regularity supports the use $AK^{\alpha}L^{\beta}$. About 2/3 of national income has gone to labor and 1/3 to capital. "In fact, the long-term stability of factor shares has become enshrined as one of the "stylized facts" of growth" (Gollin, 2002, pp. 458–459). More recent measurements of factor shares shows that capital is gaining a greater share and this is an active, exciting area of research.

Labor Demand Highlights

The most important comparative statics exercise on the input side is to derive the demand for inputs. This chapter focused on labor demand and showed that the short run demand for labor is the marginal revenue product of labor curve.

In the long run, however, the demand for labor is not the MRP curve because K^* changes as w changes. For this same reason, labor demand is more responsive to changes in the wage in the long run.

Whether in the long or short run, the demand curve for labor is subject to the same Shutdown Rule qualification as the supply curve for output. If the wage is higher than the ARP at the point at which $MRP = MFC$, the firm

will hire no labor. This coincides perfectly with the firm's decision to shut down on the output side, producing no output.

In addition to changes in the wage, this chapter explored the effects of a change in product price. As P increases, L^* rises. In terms of the canonical graph, an increase in P shifts the MRP and leads to a new optimal solution. This leads economists to think of and say that labor demand is a derived demand because the price of the product influences how much labor the firm wants.

This section ended by pointing out that an input's productivity determines its share of firm revenues. As productivity rises, so does the percentage share accruing to that input. Productivity is a key variable in determining input use and distribution of revenues.

Exercises

1. Derive the wage elasticity of short run labor demand for the general case where $L^* = (\frac{w}{\beta P A \bar{K}^{\alpha}})^{\frac{1}{\beta-1}}$. Show your work, using Word's Equation Editor.

2. Does your result from the previous question agree with the -4 value obtained in the text?

3. Compute the wage elasticity of short run labor demand (using the parameter values in the *OneVar* sheet) from w=\$20/hr to \$19/hr. Show your work.

4. Use the Comparative Statics Wizard to analyze the effect of an increase in the product price in the long run. Compute the P elasticity of L^* from $P = 2.00$ to $P2.10$. Copy and paste your results in a Word document.

5. Is L^* more responsive to changes in P in the short run or long run? Explain why.

References

The epigraph is from page 523 of David S. Landes, *The Wealth and Poverty of Nations: Why Some are So Rich and Some So Poor* (paperback edition, 1999; originally published, 1998). Landes was an economic historian interested in economic development. He asked really difficult, fascinating questions: "How

and why did we get where we are? How did the rich countries get so rich? Why are the poor countries so poor? Why did Europe ('the West') take the lead in changing the world?" (p. xxi). His answers are opinionated and clear.

The idea that a profit-maximizing firm will use and reward factors according to productivity has a normative or ethical dimension. John Bates Clark, one of the first well-known American economists, argued in *The Distribution of Wealth* (1899) that the equimarginal principle was not only efficient, but also fair. Paying factors according to productivity showed that capitalism was just. For more modern reading on morality or ethics in economics, from one end of the spectrum to the other, see Robert Nozick, *Anarchy, State, and Utopia* (1974) and John Rawls, *A Theory of Justice* (1971).

You are undoubtedly familiar with the Nobel Prize in Economic Sciences, but the John Bates Clark Medal is given every two years "to that American economist under the age of forty who is adjudged to have made the most significant contribution to economic thought and knowledge." See www.aeaweb.org/about-aea/honors-awards/bates-clark for a complete list of winners—it is peppered with Nobel Prize winners.

In his paper reconciling time series and cross section data, Douglas Gollin, "Getting Income Shares Right," *The Journal of Political Economy*, Vol. 110, No. 2 (April, 2002), pp. 458–474, www.jstor.org/stable/10.1086/338747, says that Cobb and Douglas "were among the earliest authors to point out that, for the United States, the labor share of income appeared to be roughly constant over time, regardless of changes in factor prices" (pp. 460–461). As mentioned in this section, this remarkable constancy of labor shares has crumbled recently, as labor's share has fallen. For a more recent review of labor's share, see conversableeconomist.blogspot.com/2018/02/behind-declining-labor-share-of-income.html.

[That long run responses are more elastic than short run responses] is commonly believed to be empirically true, simply as a matter of assertion. It is interesting and noteworthy that this type of behavior is in fact mathematically implied by a maximization hypothesis.

Eugene Silberberg

Chapter 14

Consistency

We have considered three separate optimization problems in our study of the perfectly competitive (PC) firm. Figures 14.1, 14.2, and 14.3 provide a snapshot of the initial solution and the key comparative statics analysis from each of the three optimization problems.

This chapter ties things together with the fundamental point that these three problems are tightly integrated and are actually different views of the same firm and same optimal solution. Change an exogenous variable and all three optimization problems are affected. The new optimal solutions and comparative statics results are *consistent*—i.e., they tell you the same thing and are never contradictory.

Figure 14.1 shows the input side cost minimization problem. Quantity is exogenous in this problem and the firm looks for the input mix that minimizes the total cost of producing the given q.

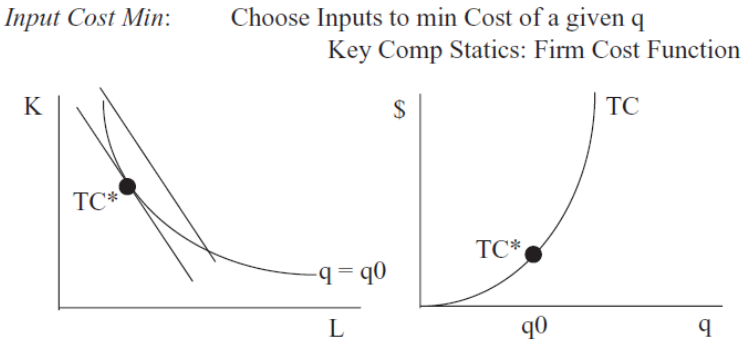


Figure 14.1: Initial cost minimization and cost function.

The right panel in Figure 14.1 shows the cost function that comes from tracking minimum total cost as q varies, ceteris paribus.

Figure 14.2 shows output side profit maximization. The PC firm (since $P = MR$ is a horizontal line) gets average and marginal cost curves from the cost function and finds the quantity that maximizes profit.

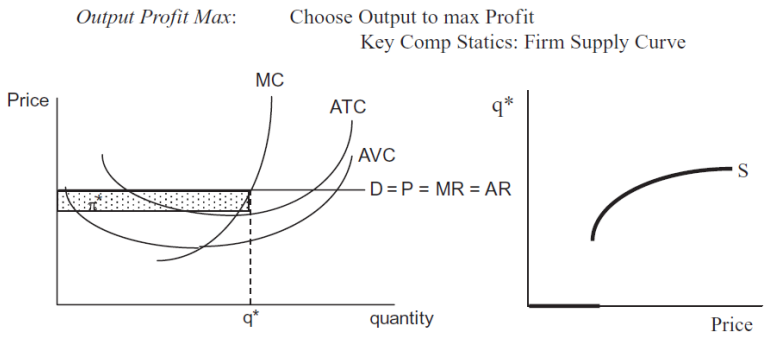


Figure 14.2: Initial profit maximization and the supply curve.

The right panel in Figure 14.2 shows where supply curves come from: shock P, ceteris paribus, and track optimal q.

Figure 14.3 returns to the input side, but this time the firm solves a profit maximization problem, choosing how much labor to hire.

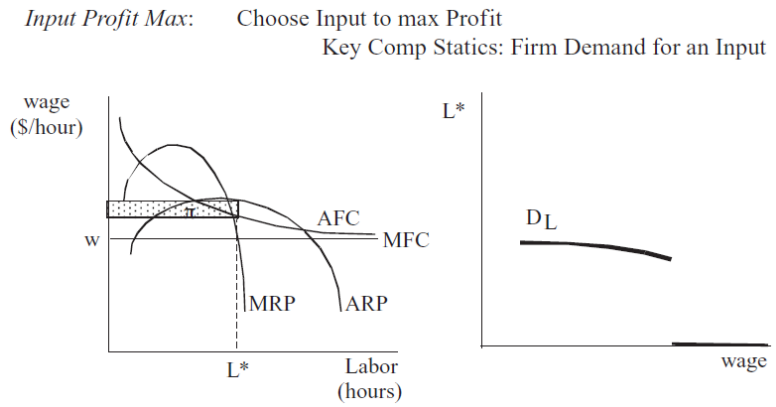


Figure 14.3: Initial profit maximization and the demand for labor.

The right panel in 14.3 shows how changing w, ceteris paribus, produces the demand curve for labor.

These three optimization problems share a common methodology. In each case, we set up and solve the problem, then do comparative statics analysis. There are other shocks that can be explored, but the one shown here is the most important.

But there is one last crucial concept that is the focus of this chapter: these three problems do not exist in isolation, instead, they are woven together to comprise the Theory of the Firm. The relationships among the three exhibit a consistency that can be demonstrated with Excel.

Perfect Competition in the Long Run

STEP Open the Excel workbook *Consistency.xls* and read the *Intro* sheet; then proceed to the *TheoryoftheFirmLongRun* sheet. Use the Zoom In button to fit the graphs on your screen so that all of them can be seen simultaneously.

The first and most important point is that all three optimization problems, in unison, comprise the Theory of the Firm. Perhaps because they see it in introductory economics, many students think of the output profit maximization graph as the firm. The display in *Consistency.xls* gives a strong visual presentation and constant reminder that the firm has three facets.

Gray-backgrounded cells are dead (click on one to see that it has a number, not a formula). They serve as benchmarks for comparisons when we do comparative statics.

The output and input profit maximization graphs do not have the usual U-shaped curves because the production function is Cobb-Douglas. This functional form cannot generate conventional U-shaped MC and AC curves (or upside down U-shaped MRP and ARP). There is no separate AVC curve because we are in the long run, so $AC = AVC$.

STEP Compare the initial solutions for each of the three problems.

There are several ways in which they agree.

1. L^* and K^* are the same in the Input Profit Max (left) and Input Cost Min (middle) graphs.

2. If you use these amounts of L and K, you will produce 636 units of output, as shown in the Output Profit Max (right) graph.

3. π^* is the same in the Input and Output Profit Max graphs. There is no profit in the Input Cost Min graph because there is no output price (P) and, therefore, no revenue in that optimization problem.

4. Total cost from each side is exactly the same. You can find TC from the Input Profit Max by creating a cell that computes wL^*+rK^*. This will equal \$36,262. From the Output Profit Max side, calculate TC by subtracting revenue, Pq, from profit. Again, you get \$36,262.

We can also see consistency in the ways in which the three optimization problems respond to shocks. As you would expect, the comparative statics results are identical.

STEP Wage increase of 1%. Change cell B2 to 20.2. Use the Zoom In button if needed to see more clearly how the graphs have changed.

Figure 14.4 shows the results.

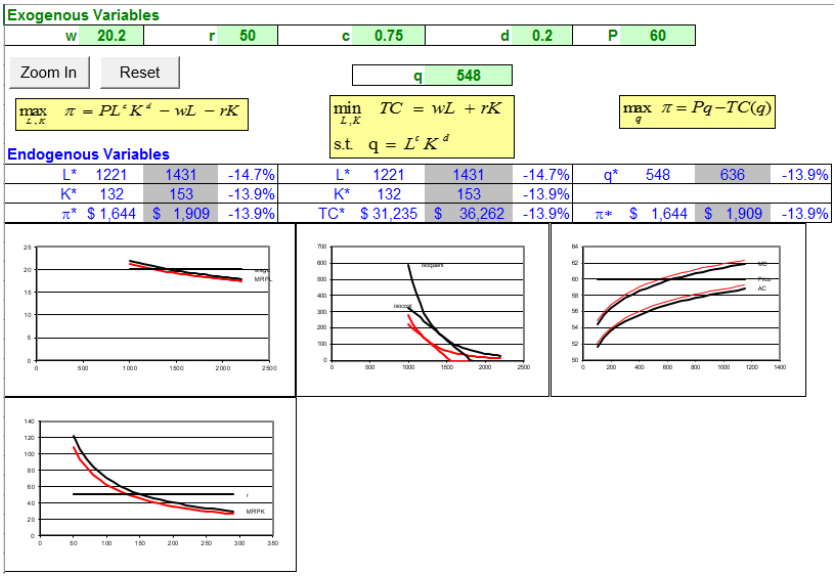


Figure 14.4: Wage shock in the long run.
Source: Consistency.xls!TheoryoftheFirmLongRun

On the Input Profit Max graph, we see that optimal labor use has fallen by 14.7% as wage rose by 1% (so the wage elasticity of labor from wage = \$20/hr

to \$20.20/hr is −14.7). Labor demand collapsed because the horizontal wage line shifted up and because the *MRP* schedule shifted left. The latter effect is due to the fact that optimal K fell.

On the Input Cost Min graph, we see that the firm is minimizing the cost of producing a lower level of output. In other words, we are on a new isoquant. Notice that the changes in L^* and K^* are consistent with the decreases reported from the Input Profit Max results.

The wage increase in the Output Profit Max graph is felt via the shifting up of the cost curves. The firm decreases q^* because *MC* shifted up and therefore the intersection of *MR* and *MC* occurs to the left of the initial solution.

Figure 14.4 and your screen shows how the Theory of the Firm reacts in a consistent manner to a wage shock. Is this true of other shocks? Yes. Here is another example.

STEP Click the Reset button, and then implement a labor productivity increase to 0.751 by changing cell F2.

Figure 14.5 shows the dramatic results of this shock. Input use and output produced have increased by about 18% in response to this tiny change in c.

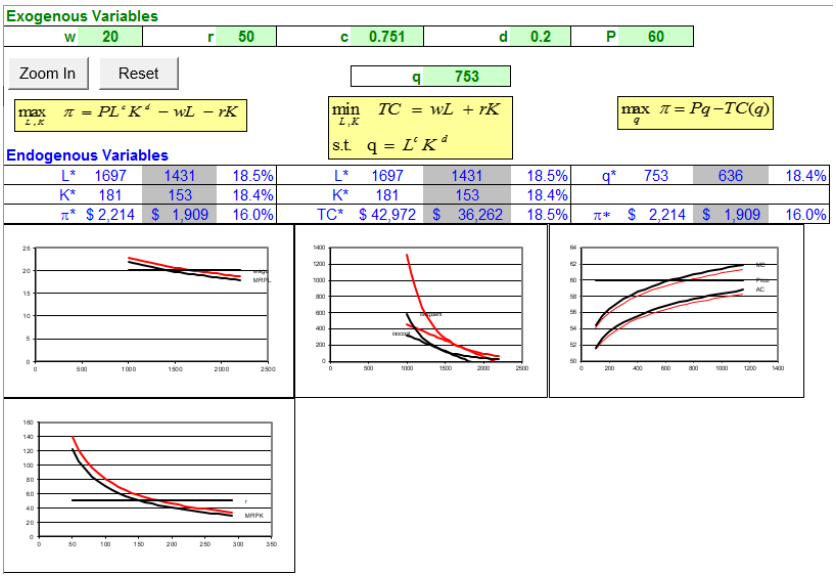


Figure 14.5: Labor productivity shock in the long run.
Source: Consistency.xls!TheoryoftheFirmLongRun

As with the wage shock, comparison of the effects of the change in c on the three optimization problems shows consistency. The two input side problems show that input use is the same and the inputs used will make the desired output on the output side. Profits on the input and output sides are the same. The productivity increase has shifted MRP up and cost curves down.

Other shocks are explored in Q&A and exercise questions. In every case, changing an exogenous variable, ceteris paribus, produces effects felt throughout the three optimization problems and the results are always consistent.

Perfect Competition in the Short Run

STEP Go to the *TheoryoftheFirmShortRun* sheet to explore the comparative statics properties of the three optimization problems in the short run.

This sheet has several differences compared to the previous overall view of the firm in the long run.

- There is an additional exogenous variable, K, because we are in the short run. Its value is set to the long run optimal solution for the initial values of the other parameters.
- There is a missing graph in the input profit max problem. With K fixed, we no longer need to depict its optimal solution.
- There is a straight, horizontal line in the isoquant side graph. With K fixed, the firm will not be able to roll around the isoquant to find the cost-minimizing input mix. It must use the given amount of K.
- There is an extra cost curve in the output profit max graph. Having K fixed means there is a fixed cost so we now have separate average total and average variable costs.

STEP Compare the initial solutions for each of the three problems. As expected, they agree in input use, output produced, and profits generated.

As before, we can change the light-green-backgrounded exogenous variable cells in row 2 and follow the results in the graphs.

STEP Apply a wage increase of 1%. Change cell B2 to 20.2. Use the Zoom In button if needed to see more clearly how the graphs have changed.

Figure 14.6 shows the results of this shock.

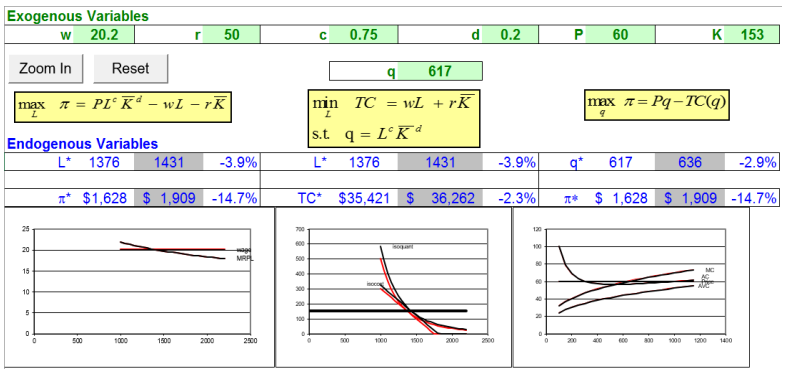


Figure 14.6: Wage shock in the short run.
Source: Consistency.xls!TheoryoftheFirmShortRun

The usual consistency properties are readily apparent. We observe the same change in L^*, q^*, and π^* across the board. Notice that the input profit max problem does not show a shift in MRP because K is fixed.

If we compare the short (Figure 14.4) to the long run (Figure 14.6), we see that the responsiveness of the changes in endogenous variables is greater in the long run. Labor and output fall by more in the long run. Profits, however, fall by less in the long run.

STEP Click the Reset button, then implement a labor productivity increase to 0.751 by changing cell F2.

Figure 14.7 displays the results.

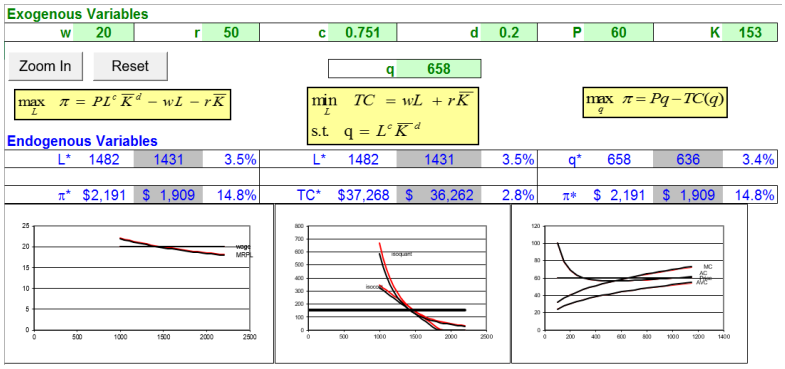


Figure 14.7: Labor productivity shock in the short run.
Source: Consistency.xls!TheoryoftheFirmShortRun

Figure 14.7 shows consistency in the results and, once again, the long run changes are more responsive than in the short run. L and K fall by more and the increase in profits is higher in the long run.

Long versus Short Run

When we compared the short and long run results for shocks in w and c, the long run exhibited greater responsiveness in labor and output. Is there a general principle at work?

Yes. The general law is that long run responses are always at least as or more elastic than in the short run. This is known as the *Le Chatelier Principle*.

Le Chatelier's idea, which he originally applied to the concept of equilibrium in chemical reactions, was introduced to economics by Nobel laureate Paul Samuelson in 1947.

> The Le Chatelier principle explains how a system that is in equilibrium will react to a perturbation. It predicts that the system will respond in a manner that will counteract the perturbation. Samuelson, following the methods of the hard sciences, has transported this principle of chemist Henri-Louis Le Chatelier to economics, to study the response of agents to price changes given some additional constraints. In his extension of this principle, Samuelson uses the metaphor of squeezing a balloon to further explain the concept. If you squeeze a balloon, its volume will decrease more if you keep its temperature constant than it will if you let the squeezing warm it up. This principle is now considered as a standard tool for comparative static analysis in economic theory. (Szenberg, et al., 2005, p. 51, footnote omitted)

In the context of the short and long run responses to shocks by a firm, the Le Chatelier Principle says that long run effects are greater because there are fewer constraints.

When the wage rises, a firm in the short run is stuck with its given quantity of K. In the long run, however, it will be able to adjust both L and K and it is this additional freedom of movement that guarantees at least as great or a greater response in input use and output produced.

For increasing c, the Le Chatelier Principle is reflected in the fact that labor demand is much more responsive in the long run than the short run. In the long run, the firm is able to take greater advantage of the labor productivity shock by renting more machines and hiring even more labor. This is, of course, reflected in the greater profits obtained in the long run in response to the increased c.

A Holistic View of the Firm

Figures 14.1, 14.2, and 14.3 are fundamental graphs for the Theory of the Firm. They represent the three optimization problems that, in unison, comprise the theory. The firm is not merely its output side representation, but includes all three optimization problem, as shown in the *Consistency.xls* workbook.

The input cost min (isoquants and isocosts that can be used to derive the cost function), output profit max (horizontal P with the family of cost curves that yield a supply curve), and input profit max graphs (horizontal w with MRP generating a demand curve for an input) are all intertwined. Not only do they all yield consistent answers for the initial solution, they all provide consistent comparative statics responses.

If we compare short and long run effects of shocks, we see that the firm responds more energetically in the long run. The wage elasticity of labor is greater (in absolute value) in the long run and, via consistency, so is the wage elasticity of output. Similarly, the c elasticities of labor and output are also greater in the long run.

Both of these results are examples of the Le Chatelier Principle: With fewer constraints, responsiveness increases. Since the short run prevents K from varying, the firm is less able to adjust to a shock. It can only vary L and, thus, its adjustment is more restricted and inelastic.

Exercises

1. What happens in the long run when price increases by 1%? Implement the shock and take a picture of the results, then paste it in a Word document. Comment on the changes in optimal labor, capital, output, and profits.

2. Compute the long run output price elasticity of labor demand. Show your work.

3. Apply the same 1% price increase in the short run. Take a picture of the results, then paste it in your Word document. Comment on the changes in optimal labor, capital, output, and profits.

4. Compute the short run output price elasticity of labor demand. Show your work.

5. Compare the price elasticities of labor demand in the long (question 2) and short run (question 4). Is the Le Chatelier Principle at work here? Explain why or why not.

6. With output price 1% higher, increase the wage by 1% in the long and short run. Do these two shocks cancel each other out in either case? Explain.

References

The epigraph is from page 116 of Eugene Silberberg, *The Structure of Economics: A Mathematical Analysis* (1978). This is a classic Math Econ book that was a popular graduate-level text.

Michael Szenberg, Aron Gottesman, and Lall Ramrattan, *Paul Samuelson On Being an Economist* (2005), explore the life and contributions of one of the most important economists of the 20^{th} century.

Instead of using marginal conditions as Cournot had done, Marshall used total ones. Perhaps for that reason, Cournot's marginal-revenue concept was forgotten and had to be rediscovered in the 1930s.

Hans Brems

Chapter 15

Monopoly

Like the perfectly competitive firm, a monopolist has three interrelated optimization problems. Attention is focused on the output profit max problem because that is where the essential difference lies between a perfectly competitive (PC) firm and a monopoly. We know that via consistency, monopoly power manifests itself on the input side also. A monopoly will produce less than a PC firm and, in turn, hire less labor and capital.

Unlike a PC firm, a monopoly chooses output and the price at which to sell the product. This makes the monopoly problem harder to solve. Fortunately, your experience with optimization, comparative statics, and graphical displays give you the background needed to understand and master monopoly.

Definition and Issues

A *monopoly* is defined as a firm that is the sole seller of a product with no close substitutes. The definition is inherently vague because there is no clear demarcation for what constitutes a close substitute.

Consider this example: In the old days, a local cable provider might have an exclusive agreement to provide cable TV in a community. One could argue that the cable provider was a monopoly because it was the sole seller of cable TV. But what are the substitutes for cable TV?

Years ago, cable TV was the only way to access subscription channels such as ESPN and HBO. Commercial broadcasts (with national broadcasters such as ABC, NBC, and CBS and local channels) were a poor substitute for cable TV. In this environment, cable TV would be a good example of a monopoly.

Today, however, cable TV has strong competition from satellite services and streaming services from the web. Even if a firm had an exclusive franchise to deliver cable TV in a community, there are many ways to get essentially the same package of channels. Today, cable TV is not a monopoly.

Of course, cable TV is not a good example of perfect competition either. The cable company does not accept price as a given variable. It is in the middle, somewhere between perfect competition and monopoly. Markets served by a few firms are called *oligopolies*. Add more firms and you eventually get monopolistic competition. The study of how firms behave under a variety of market structures is part of the subdiscipline of economics called Industrial Organization. Figure 15.1 sums things up.

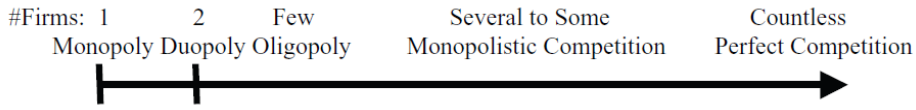


Figure 15.1: A continuum of market structures.

Barrier to Entry

To remain a monopoly, the firm must have a *barrier to entry* to prevent other firms from selling its product. In the cable TV example, the barrier to entry was provided by the exclusive agreement with the community. Such governmental restriction is a common form of a barrier to entry.

Another way to erect a barrier to entry is control over a needed input. ALCOA (the Aluminum Corporation of America) had a monopoly in aluminum in the early 20^{th} century because it owned virtually all bauxite reserves.

If a product requires entry on a large scale, like automobile manufacturing, this is considered a barrier to entry. To compete against established car companies, a firm must not only produce cars, but also many spare parts and figure out how to sell the product.

Like the concept of a close substitute, a barrier to entry is not a simple yes or no issue. Barriers can be weak or strong and they can change over time. Cable TV's barrier was eroded not by changes in legal rules, but by technological change—the advent of satellite TV and the web.

Monopoly's Revenue Function

We know that the firm's market structure impacts its revenue function. The simplest case is a perfectly (or purely) competitive firm. It takes price as given and, therefore, revenues are simply price times quantity. For a perfect competitor, even though market demand is downward sloping, the firm's own individual demand curve is perfectly elastic at the given, market price.

Because the PC firm can sell as much as it wants at the given price, selling one more unit of output makes total revenue (TR) increase by the price of the product. Marginal revenue (MR) is defined as the change in TR when one more unit is sold. Thus, for a PC firm, $MR = P$.

This is not true for a monopoly. A critical implication of monopoly power is that MR diverges from the demand curve. But this is too abstract. We can use Excel to make these concepts clearer.

STEP Open the Excel workbook *Monopoly.xls* and read the *Intro* sheet, then go to the *Revenue* sheet to see how monopoly power affects the firm's revenue function.

The sheet opens with a perfectly competitive revenue structure. Total revenue is a linear function of output and, therefore, $P = MR$ with a horizontal line in the bottom graph. A graph with a linear TR and corresponding horizontal MR means it is a PC firm.

Unlike a PC firm, a monopoly faces the market's downward sloping demand curve. We can model a linear inverse demand curve simply as $P = p_0 - p_1 q$. Because the slope parameter, p_1, in cell T2 is initially zero, TR is linear and MR is horizontal.

STEP To show how monopoly power affects the firm's revenue function, click on the *Price Slope* scroll bar.

Notice that as you increase the slope parameter, MR diverges more from D.

The smaller (in absolute value) the price elasticity of demand, the greater the divergence of MR from D and the stronger the monopoly power.

We will see that the monopolist uses the divergence of MR from D to extract higher profits than would be possible if there were other sellers of the product.

When drawing MR and D in the case of a linear inverse demand curve, keep in mind these two basic rules:

1. MR and D have the same intercept.
2. MR bisects the y axis and D.

We can derive these properties easily. With our inverse D curve, $P = p_0 - p_1 q$, we can do the following:

$$TR = Pq$$
$$TR = (p_0 - p_1 q)q$$
$$TR = p_0 q - p_1 q^2$$
$$MR = \frac{dTR}{dq} = p_0 - 2p_1 q$$

Clearly, both D and MR share the same intercept, p_0. Because the slope of MR is $-2p_1$, it is twice the slope of D, which is simply $-p_1$.

Thus, when you draw a linear inverse demand curve and then prepare to draw the corresponding MR curve, remember the two rules: (1) the intercept is the same and (2) MR has twice the slope so at every y axis value, MR is halfway between the y axis and the D curve.

Figure 15.2, with an inverse demand curve slope of -1, shows the monopoly's revenue function. Unlike the PC firm, TR is a curve and MR diverges from D. MR bisects the y axis and D. The dashed line at \$20/unit, for example, shows the distance from the y axis to MR is 10, the same as MR to D.

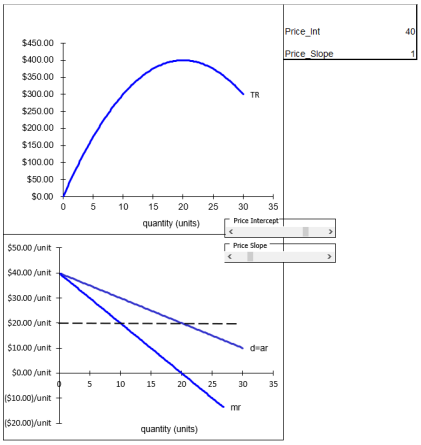


Figure 15.2: TR, D, and MR functions for a monopolist.
Source: Monopoly.xls!Revenues

Notice that where $MR = 0$ at $q = 20$, TR is at its maximum. At this quantity, the price elasticity of demand is exactly -1.

Figure 15.2 shows that MR can be negative. This can happen because there are two opposing forces at work. Increasing quantity increases TR, since $TR = Pq$. However, the only way to sell that extra product is to lower the price (by traveling down the demand curve) so TR falls. When the increase to TR by selling additional output outweighs the effect of the drop in the price, MR is positive. Eventually, however, with a linear demand curve, the monopolist will reach a point at which the increase in revenue for selling one more unit is negative. In the range of output ($q > 20$ in Figure 15.2) where $MR < 0$, the effect of the decreased price outweighs the positive effect of selling more output.

When $MR > 0$, the price elasticity of demand is greater than 1 (in absolute value). When MR is negative, demand is inelastic. The monopolist will never produce on the negative part of MR, which is the same as the inelastic portion of the demand curve.

There is a neat formula that expresses the relationship between MR and P. With an inverse demand curve, $P(Q)$, we know that $TR = P(Q)Q$. From the TR function we can take the derivative with respect to output to find the MR function. We use the Product Rule:

$$MR = \frac{dTR}{dQ} = P + \frac{dP}{dQ}Q$$

If we factor out P from this expression, then MR can be rewritten as:

$$MR = P + \frac{dP}{dQ}Q = P(1 + \frac{dP}{dQ}\frac{Q}{P}) = P(1 + \frac{1}{\epsilon})$$

The Greek letter epsilon (ϵ) is the price elasticity of demand ($\frac{dQ}{dP}\frac{P}{Q}$). The expression shows that $MR = P$ under perfect competition because an individual firm faces a perfectly elastic demand curve. This means epsilon is infinite and its reciprocal is zero.

It also shows that the more inelastic the demand curve (the closer ϵ is to 0), the greater the separation between MR and the demand curve (P). If $\epsilon = 0$, then MR is undefined. With $\epsilon = 0$, inverse demand is a vertical line. The monopoly would charge an infinite price.

Setting Up the Problem

There are three parts to every optimization problem. Here is the framework for a monopolist's output side profit maximization problem.

1. Goal: maximize profits (π), which equal total revenues (TR) minus total costs (TC).

2. Endogenous variables: output (q) and price (P)

3. Exogenous variables: input prices (the wage rate and the rental rate of capital), demand function coefficients, and technology (parameters in the production function).

A monopoly differs from a PC firm only on the revenue side—price is now endogenous. The cost structure is the same. The monopoly has an input cost min problem and it is used to derive a cost function. Increases in input prices shift cost curves up and improvements in technology shift cost curves down. The monopolist has a long and short run, just like a PC firm, and in the short run there is a gap between ATC and AVC that represents fixed costs.

Finding the Initial Solution

We will show the conventional approach to solving the monopoly problem first, then turn to an alternative formulation based on constrained optimization.

The conventional approach is to find optimal q where $MR = MC$, then get optimal P from the demand curve, and then compute optimal π as a rectangle. This is the standard approach and there is a canonical graph that goes along with this approach. Its primary virtue is that it can be easily compared to the perfectly competitive case.

The conventional approach can be demonstrated with a concrete problem. Suppose the cost function is $TC = aq^3 + bq^2 + cq + d$. Suppose the market (inverse) demand curve is $P = p_0 - p_1 q$. Thus, $TR = Pq = (p_0 - p_1)q$.

With this information, we can form the firm's profit function and optimiza-

tion problem, like this:

$$\max_q \pi = TR - TC$$
$$\max_q \pi = (p_0 - p_1)q - (aq^3 + bq^2 + cq + d)$$

We first solve this problem with numerical methods, then analytically.

STEP Proceed to the *OptimalChoice* sheet and look it over.
The profit function has been entered into cell B4. Quantity and price are displayed as endogenous variables, but q is bolded to indicate that it is the primary endogenous variable. In other words, Solver will search for the profit-maximizing output and, having found it, will compute the highest price that can be obtained from the demand curve.

The firm is making \$245 in profits by producing 10 units of output and charging \$34.50 per unit, but this is not the profit-maximizing solution. We know this because the marginal revenue of the 10^{th} unit is \$29/unit, whereas the marginal cost of that last unit is only \$4/unit. Clearly, the firm should produce more because it is making more in additional revenues from the last unit produced than the additional cost of producing that unit.

STEP Run Solver to find the optimal solution.

At the optimal solution, the equimarginal condition, $MR = MC$, is met. With positive profits, this is a clear signal that we have found the answer.

Before you click the Analytical Solution button, try doing the problem on your own. This is a single variable unconstrained maximization because $P = p_0 - p_1 q$ has been substituted into the profit function. Take the derivative with respect to q, set it equal to zero, and solve for optimal q. Substituting in the parameter values to make it a concrete problem makes it easier to do the math:

$$\max_q \pi = (40 - 0.55)q - (0.04q^3 - 0.9q^2 + 10q + 50)$$

You can check your work by clicking the Analytical Solution button. You can also confirm that the two approaches, Solver and calculus, agree.

STEP Proceed to the *OutputSide* sheet to see a familiar set of four graphs.

As usual, the totals are on the top and the average and marginal curves on the bottom. The cost curves are the quite similar to the PC firm's output profit maximization graphs, but the revenue curves are quite different.

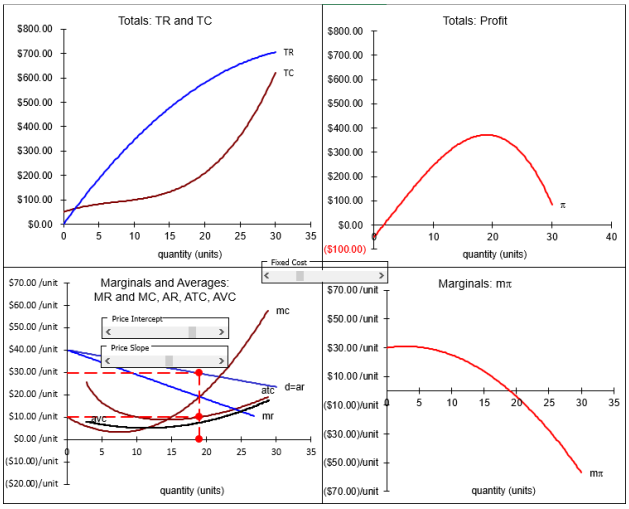


Figure 15.3: Monopoly output profit maximization graphs.
Source: Monopoly.xls!OutputSide

The bottom left-hand corner graph in Figure 15.3 is the canonical graph for a monopolist. It can be used to quickly find q^*, P^*, and π^*. Here's how to read and use the conventional monopoly graph:

1. Finding q^*: Choose q where $MR = MC$. This gives the biggest the difference between TR and TC and puts you on top of the profit hill (in the top right graph).

2. At q^*, travel straight up until you hit the demand curve to get P^*. This is the highest price that the monopolist can get for the chosen level of output.

3. Create the usual profit rectangle as $(AR–ATC)q^*$. It has length q^* and height $AR - ATC$ (where $AR = P$). The area of this rectangle equals the distance of the line segment between TR and TC, which is the height of the profit hill.

Play with the slider controls to improve your understanding of the graphs and relationships.

STEP Click the Fixed Cost slider to manipulate total fixed costs (d in the cubic cost function).

Changes in fixed costs do not affect the monopolist's optimal quantity and price solution. This is just like the perfectly competitive case.

STEP Click the Reset button; explore changes in the price intercept to see how the firm responds. At a low enough price intercept, profits become negative and, just like a PC firm, if $P < AVC$, the firm will shut down.

You can also control the firm's monopoly power by manipulating the inverse demand curve's slope.

STEP Set the Price Slope slider to zero. What happens?

You stripped the monopoly of its price power and it is a PC firm.

No Supply Curve for Monopoly

Monopolists do not have a supply curve. This seems like a strange statement since monopolies produce output and so "supply" whatever good or service of which they are the sole seller. But the key lies in the definition of a supply curve: given price, the supply curve gives the quantity that will be produced.

Because a PC firm is a price taker, it is possible to shock P and see how the optimal output changes. We can derive $q^* = f(P,$ ceteris paribus) and this is called a supply curve.

Unlike a perfectly competitive firm, for which price is exogenous, a monopoly chooses the price. Thus, we cannot ask, "Given this price, what is the optimal quantity supplied?" With price as an endogenous variable, it cannot serve as a shock variable in a comparative statics analysis.

We can (and you just did) shock a monopolist's demand curve parameters such as the intercept and slope, but this is not an exogenous change in the price of the product. The experiment of changing the price cannot be applied to a monopolist and, therefore, the monopolist has no supply curve.

Measuring Monopoly Power

Another common misconception is that monopoly is either zero or one. In fact, it is a continuum and you can have more or less monopoly power. There are several ways to measure it.

STEP Proceed to the *Lerner* sheet.

This sheet demonstrates the point that the more inelastic the demand faced by a monopolist, the greater the monopoly power. In other words, from a profit-maximizing point of view, it is better to have a monopoly over a product that everyone desperately needs (i.e., very inelastic) than to be the sole seller of a product that has a highly elastic market demand curve.

Abba Lerner formalized this idea in a mathematical expression that bears his name, the *Lerner Index*. "If P = price and MC = marginal cost, then the index of the degree of monopoly power is $\frac{P-MC}{P}$." (Lerner, 1934, p. 169). This measure of monopoly power uses the gap between P and MC as a percentage of P.

The Lerner Index takes advantage of the fact that a monopolist will choose that quantity where $MR = MC$, then charge the highest price possible for that quantity. The higher the price that can be charged, the more inelastic is demand and the greater the monopoly power.

The *Lerner* sheet compares two monopolies with the exact same cost structure (assumed for simplicity to have a constant $MC = AC$). They both produce the same profit-maximizing quantity, but Firm 2 faces a more inelastic demand curve than Firm 1 and, therefore, it has a bigger gap between price and marginal cost.

STEP Click on cells B16 and I16 to see the simple formulas for the Lerner Index.

The idea is that the bigger the divergence between price and marginal cost, the greater the monopoly power. Firm 2 has more monopoly power than Firm 1 and more monopoly profits. The Lerner Index for each firm reflects this.

Notice that a perfectly competitive firm that sets $MC = P$ will have a Lerner Index of zero. As the index approaches one, monopoly power rises.

STEP Change Firm 2's demand parameters to 130 for the intercept and 20 for the slope. The y axis is locked down so the entire D and MR functions are not displayed.

The optimal quantity is still 3, but P and profits are higher, as is the Lerner Index.

STEP Make the demand curve more inelastic at $Q = 3$ by setting the demand parameters to 190 and 30.

Optimal P has increased again, along with profits. The Lerner Index reflects the greater monopoly power.

STEP One last time, change the demand parameters to 6010 and 1000. The graph is hard to read because only MR is shown; D is literally off the chart.

Firm 2 continues to produce the same output as Firm 1, but has a much, much higher optimal price and maximum profits. Its Lerner Index is close to one. It cannot rise above one, but the closer it gets, the greater the divergence of P and MC so the greater the monopoly power.

The *Lerner* sheet also shows that the Lerner Index can be expressed as the reciprocal of the price elasticity of demand at the profit-maximizing price. The few algebra steps needed to connect the Lerner Index to the price elasticity start in row 25.

STEP Set Firm 2's demand parameters back to 70 and 10, and then click the Show Elasticity button.

The price elasticity of demand for the two firms is displayed. If you click in the cells, you can see the formula. Notice that the reciprocal of the inverse demand curve's slope is used to compute the price elasticity of demand correctly.

Firm 2's price elasticity of demand at the profit-maximizing price is lower than Firm 1's. The lower the price elasticity and the higher the Lerner Index, the greater the firm's monopoly power.

STEP Proceed to the *Herfindahl* sheet for a quick look at another way to measure monopoly power.

Instead of measuring the markup of price over marginal cost, we can see how big the firms are in an industry. Strictly speaking, a monopoly is one firm so it would have a 100% market share, but in practice, firms have monopoly power even though they are not technically monopolies. Any firm that faces a downward sloping demand curve and has the ability to set its price is said to have monopoly power.

If a market has many firms, each with the same share of total sales, we have a competitive market structure. If, on the other hand, only a few firms exist, the market is monopolized. The question is how to measure the degree of monopolization?

We can sort the firms in an industry from highest to lowest share and then add the shares of the four biggest firms. This gives the *four firm concentration ratio* in cell D5. It turns out this is not a very good way to distinguish between concentrated and unconcentrated industries.

The problem is that the four firm concentration ratio tells you nothing about the sizes of the top four firms or the rest of the market. The four firm concentration ratio is 70%, which seems pretty highly concentrated. The biggest firm's share, 30%, is almost one-third of the entire industry.

STEP Click on the Distribution A button.

The four firm concentration ratio is the same as before (70%), but this industry is clearly much more concentrated. Firm A is even bigger and the others are tiny.

STEP Click on the Distribution B button.

The four firm concentration ratio is the same as before (70%), but this industry is clearly less concentrated. The four top firms are equal so no one firm really dominates.

The primary virtue of the four firm concentration ratio is that it is easy to compute and understand. However, because we have three scenarios with wildly different shares for the top four firms yielding the same four firm con-

centration ratio, we can conclude that this ratio is a poor way to determine whether firms in a market are in a competitive or monopolistic environment. The four firm concentration ratio might be easy to compute and understand, but it is incapable of picking up differences in the distribution of shares.

A better way to judge concentration is via the *Herfindahl Index*. Unlike the Lerner Index, there is confusion about who invented it. Hirschman concludes, "The net result is that my index is named either after Gini who did not invent it at all or after Herfindahl who reinvented it. Well, it's a cruel world" (Hirschman, 1964, p. 761). It is sometimes called the Herfindahl-Hirschman Index (HHI).

Fortunately, its computation is simpler than its paternity. The idea is to square each share and sum, like this:

$$H = \sum_{i=1}^{n} S_i^2$$

The index ranges from $1/n$ to 1 (when using decimal values of shares). The higher the index, the greater the concentration. By squaring the shares, it gives more weight to bigger firms: for example, $0.1^2 = 0.01$, while $0.3^2 = 0.09$.

The *Herfindahl* sheet shows the computation. Notice how each value in column B is squared in column G. The sum of the squares is in cell G15 and it is the value of the Herfindahl Index.

STEP Click on the three buttons one after the other to cycle through them. Notice how the Herfindahl Index changes (but the four firm concentration ratio does not).

For Distribution A, the H value is 0.325. This is quite high. The 0.1375 value with Distribution B means there is more competition in this scenario than the other two.

The Herfindahl Index is not perfect because no single number can completely describe an entire distribution. It is, however, better than the four firm concentration ratio and often used to measure the degree of market competition.

The United States Department of Justice is charged with regulating the conduct and organization of businesses. The mission of the Antitrust Division is to promote economic competition. They use the Herfindahl Index as part of

their Horizontal Merger Guidelines (www.justice.gov/atr/horizontal-merger-guidelines-08192010). Markets with a Herfindahl Index less than 0.15 are "unconcentrated," values between 0.15 and 0.25 are "moderately concentrated," and anything over 0.25 is "highly concentrated."

The Department of Justice deems any proposed merger that increases the Herfindahl Index by more than 0.01 (100 points in the scale they use) in concentrated markets as warranting scrutiny. They can go to court to block mergers to prevent too much concentration. They can also break up companies that have too much monopoly power. This is known as *antitrust law* and is part of the Industrial Organization field of economics.

An Unconventional Approach

The monopolist's profit maximization problem can also be solved by choosing P and q simultaneously subject to the constraint of the demand curve. While this is not the usual way of framing the monopoly's optimization problem, it enables practice with the Lagrangean method of solving constrained optimization problems and reading isoprofit curves.

The analytical solution is based on rewriting the constraint so it is equal to zero ($P - (p_0 - p_1 q) = 0$), forming the Lagrangean, setting derivatives equal to zero, and solving the system of equations for the optimal solution.

$$\max_{p,q} L = Pq - (aq^3 + bq^2 + cq + d) + \lambda(P - (p_0 - p_1 q))$$
$$\frac{dL}{dP} = q + \lambda$$
$$\frac{dL}{dq} = P - 3aq^2 - 2bq - c + \lambda p_1$$
$$\frac{dL}{d\lambda} = P - (p_0 - p_1 q)$$

Set each derivative equal to zero and solve the three first-order conditions for q^*, P^*, and λ^*. From the first equation, $\lambda = -q$, substitute into the second equation:

$$P - 3aq^2 - 2bq - c + [-q]p_1 = 0$$

From the third first-order condition, $P = p_0 - p_1 q$, so

$$(p_0 - p_1 q) - 3aq^2 - 2bq - c - qp_1 = 0$$

Rearrange the terms to prepare for using the quadratic formula.

$$-3aq^2 - 2(b + p_1)q + (p_0 - c) = 0$$

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

$$\frac{2(b + p_1) \pm \sqrt{4(b + p_1)^2 - 4(-3a)(p_0 - c)}}{2(-3a)}$$

STEP Proceed to the *ConOpt* sheet to see formulas based on the Lagrangean solution starting in cell F24.

Naturally, we get the same, correct answer as the unconstrained version.

The *ConOpt* sheet shows that monopoly as a constrained optimization problem can be depicted with a graph. The pink curves are isoprofit curves and the black line is inverse demand. The *MR* curve is not drawn because it is not used. The firm is trying to get to highest isoprofit without violating the demand curve constraint. Clearly, the opening values are not optimal.

STEP Run Solver and get a Sensitivity Report to confirm the value of lambda star is minus optimal quantity. Notice how the Solver dialog box is set up so Solver chooses cells B8 and B9 subject to the constraint.

After running Solver, the graph, reproduced in Figure 15.4, shows the usual tangency result.

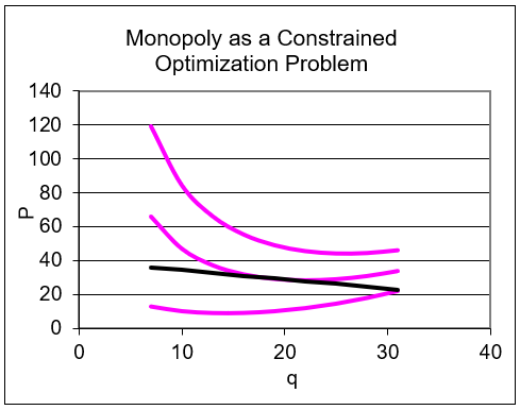


Figure 15.4: The constrained optimization version of the monopoly problem. *Source: Monopoly.xls!ConOpt*

The point of tangency provides the optimal q and P solution, while the value of the isoprofit curve at that point is the level of profits.

Do not be confused. The constrained version is rarely used. The conventional approach is the canonical output profit maximization graph (bottom left in Figure 15,2). This graph shows the optimal q where $MR = MC$ and easily displays P^* from the demand curve and π^* as a rectangle.

Figure 15.4 gives the same optimal solution, but presents the problem in a different way. Understanding that the demand curve serves as a constraint on monopoly is helpful. Monopoly power is not infinite. A monopolist cannot choose a ridiculously high price *and* a high quantity. As price rises, quantity sold must fall.

Monopoly Basics

A monopoly differs from a perfectly competitive firm in that a monopolist can choose the quantity and price, whereas a perfect competitor is a price taker. In addition, a monopolist has a barrier to entry that enables it to maintain positive economic profits even in the long run.

The two are the same, however, in the cost structure (like a perfect competitor, the monopolist derives its cost function from the input cost minimization problem) and the fact that it seeks to maximize profits (where $MR = MC$ as long as $P > AVC$).

We depict the monopolist's optimal solution with a graph that superimposes D and MR over the family of cost curves (MC, ATC, and AVC). Like a PC firm, a monopolist can suffer negative profits in the short run and it will shut down when $P < AVC$.

Monopoly's canonical graph (the bottom left chart in Figure 15.2) belongs in the pantheon of fundamental graphs in economics. Like the indifference curves with a budget constraint or supply and demand, a linear inverse demand with its associated marginal revenue showing optimal q (at the intersection of MR and MC, of course) and optimal P is a truly classic graph.

One way to measure monopoly power is by the Lerner Index. The greater the gap between price and marginal cost, the greater the monopoly power. The greater the price elasticity of demand, the lower the Lerner Index and

the weaker the monopoly power.

The Herfindahl Index is another way to measure the strength of monopolization in a market. It measures industry concentration. Unlike the four firm concentration ratio, it uses market shares of every firm to create a single number that reflects the concentration of an industry. Mergers that boost the Herfindahl Index by more than 0.01 (100 points) in concentrated markets are carefully scrutinized by the Department of Justice because it is presumed that the market will not be competitive.

We concluded this chapter with an unconventional analysis. The monopoly's profit maximization problem can be cast as a constrained optimization problem. In addition to providing practice with the Lagrangean method, this way of looking at monopoly makes quite clear that the monopolist must obey the demand curve.

Exercises

1. De Beers is an internationally famous company that had a monopoly over diamonds. Google "synthetic diamonds" to learn more. Include web citations with supporting evidence in your answers to these two questions.

   (a) What was their barrier to entry when they had a monopoly?

   (b) What happened to their monopoly?

2. Use Word's Drawing Tools to depict a monopoly shutting down in the short run. Explain the graph.

3. In the *ConOpt* sheet, set the demand intercept (cell B13) to 9 and the fixed cost (B18) to 180. Run Solver. Why is Solver generating a miserable result? What is the correct answer?

4. Use Word's Drawing Tools to depict the effect of monopoly from the input side profit maximization perspective. Explain the graph.

   Hint: With perfect competition, L^* is found where $w = MRP$ (where MRP is based on the given, constant price, $PxMP$). With monopoly, however, P and MR diverge.

5. Is the effect of monopoly on the input side consistent with the effect of monopoly on the output side? Explain.

References

The epigraph is from page 149 of Hans Brems, *Pioneering Economic Theory, 1630–1980: A Mathematical Restatement* (1986). This book recasts ideas in the history of economics in mathematical terms. Seeing the thoughts of Smith, Ricardo, Marx, and others presented as mathematical models provides an uncommon perspective.

On the Lerner Index, see Abba P. Lerner, "The Concept of Monopoly and the Measurement of Monopoly Power," *The Review of Economic Studies*, Vol. 1, No. 3 (June, 1934), pp. 157–175, www.jstor.org/stable/2967480.

On the Herfindahl Index, see Albert O. Hirschman, "The Paternity of an Index," *The American Economic Review*, Vol. 54, No. 5 (September, 1964), p. 761, www.jstor.org/stable/1818582.

> Von Neumann hovered for a moment by two rather sloppily dressed graduate students who hunched over a peculiar-looking piece of cardboard. It was a rhombus covered with hexagons. It looked like a bathroom floor. The two young men were taking turns putting down black and white go stones and had very nearly covered the entire board.
>
> Later that evening, at a faculty dinner, he buttonholed Tucker and asked, with studied casualness, "Oh, by the way, what was it they were playing?" "Nash," answered Tucker, allowing the corners of mouth to turn upwards ever so slightly, "Nash."
>
> ---
>
> Sylvia Nasar

Chapter 16

Game Theory

In perfect competition, firms are price takers with no power to affect the market price. Each firm optimizes by choosing q to equalize MC and P.

In monopoly, the sole seller of a product with no close substitutes optimizes by choosing q to equalize MC and MR and then charges the highest price that clears the market (given by the demand curve).

In both market structures, the profits of the individual firm are not affected by what anyone else does. In perfect competition, there are so many other firms that Firm i does not care about what Firm j is doing. In monopoly, there is no other firm to worry about.

What about market structures between the extremes of perfect competition and monopoly? *Oligopoly* is a market dominated by a few firms. Their decisions are interdependent. In other words, what each individual firm chooses does affect the sales and profits of the other firm. To optimize, each firm must anticipate what their rivals will do and then choose its best options. This is clearly a more realistic model than that of perfect competition and monopoly, which rely on idealized, abstract descriptions of firms that have no real-world counterparts.

How do oligopolies behave? We know that, like other firms, they optimize given the economic environment, but because of interdependence, it is much more difficult to analyze.

This chapter opens the door to the analysis of strategic behavior. It presents a few basic ideas from the fields of Game Theory and Industrial Organization.

Interdependence and Nash Equilibrium

It seems obvious when we say that firms are interdependent, but exactly what does this mean? Consider two power companies that generate and sell electricity. This is a good example of a homogeneous product. We assume consumers do not care at all which of the two firms provides electricity to their homes.

To keep it simple, suppose that each power company can choose either a high level of output or a low level of output. Market price is a function of the output decisions of the two firms. Each power company's profits are functions of their own decision to produce and the market price.

Figure 16.1 displays a *payoff matrix*, which shows the possible choices and outcomes. You read the entries in the payoff matrix like coordinate pairs on a graph, the first part is for Firm 1 and the second for Firm 2. The \$300, \$300 pair in the top left of the four entries says that Firm 1 chose high output and Firm 2 chose high output. Each firm ends up with low profits.

		Firm 2			
		High Output		Low Output	
Firm 1	High Output	\$300 profits,	\$300 profits	\$1000 profits,	\$200 profits
	Low Output	\$200 profits,	\$1000 profits	\$800 profits,	\$800 profits

Figure 16.1: The payoff matrix.

If Firm 2 had chosen low output (top right), Firm 1 profits would be much higher, \$1,000, because it made a lot of output and price rose when Firm 2 decided to cut back.

This particular game is a one-shot, simultaneous-move game known as the *Prisoner's Dilemma*. You have probably seen it before. Two criminals are arrested and questioned separately. If both stay silent, they get 1 year in jail. If both confess, they get 3 years. But if one confesses and the other does not, the one who talks gets no jail time and the silent one gets 10 years.

You can match those outcomes to the payoff matrix in Figure 16.1. The outcome that is best for both firms together is \$1600 total, with \$800 for each company. But, like the criminals version of the game, that is going to be an unlikely outcome. Suppose that both agree beforehand that they

are going to collude and both choose low output. Unless they can write a binding agreement that is enforceable (so a cheater can be punished), there is an incentive for each firm to change its decision and choose high output if it thinks that the other firm will stick with low output. As a result, both firms end up with low profits (and both criminals confess).

If you think the other firm is going to cheat, your best move is to also cheat. If you think the other firm is going to honor the agreement, your best move, in the sense of profit maximization, is to cheat and produce a high output (assuming this is a one-time game and you never have to see your opponent again). It looks like cheating, producing high output (or confessing), is the best move no matter what the other firm does. We say that this game has a dominant strategy—produce high output (confess).

This result illustrates the reason why cartels—groups of firms that get together to charge the monopoly price and split the monopoly profits—are unstable. It is difficult for oligopolistic firms to get together and act like a monopoly because there is an incentive for individual firms to cheat on the agreement and produce more to take advantage of high prices.

Because of the interdependence of firms' decision making, competition among firms in an oligopoly may resemble military operations involving tactics, strategies, moves, and countermoves. Economists model these sophisticated decision making processes using game theory, a branch of mathematics and economics that was developed by John von Neumann (pronounced noy-man) and Oskar Morgenstern in the 1930s. One of the most important contributors to game theory is John Nash, a mathematician who shared the Nobel Prize in Economics.

A game-theoretic analysis of oligopoly is based on the assumption that each firm assumes that its rivals are optimizing agents. That is, managers act as though their opponents or rivals will always adopt the most profitable countermove to any move they make. The manager's job is to find the optimal response.

Nash's most important and enduring contribution is the concept named after him, the Nash equilibrium. Once we are in a world where firms are interdependent and one firm's profits depends on what other firms do, we are out of the world of exogenously given price that we used for perfect competition and out of the isolated world of the monopolist. John Nash invented an equilibrium concept that describes a state of rest in this new world of

interdependence.

A *Nash equilibrium* exists when each player, observing what her rivals have chosen, would not choose to alter the move she herself chose. In other words, this is a *no regrets equilibrium*: After observing the outcome, the player does not wish she would have done something else instead.

We will explore in detail a concrete example of a *duopoly* (a market with two firms) with a single Nash equilibrium. Remember, however, that this is simply one example. Some games have one Nash equilibrium, some have many, and some have none. There are many, many games and scenarios in game theory and we will look at just one simple example.

The Cournot Model

Augustin Cournot (pronounced coor-no) was a remarkably creative 19^{th}-century French economist (see the References in section 12.2). Cournot originally set up a model of duopolists who produce the same good and optimize by choosing their own output levels based on assumptions about what the rival will do.

Here is the set up:

- Two firms.
- Each produces the exact same product.
- Constant unit cost.
- Firms choose output levels at the same time.
- Both know the market demand for the product.

The profit of each firm depends on how much it produces and how much its rival produces. If the rival produces a lot, the the market price falls. The interdependence is that one firm's decision about how much to produce affects the price and, thus, the rival's profit.

What strategy should each firm use to choose its output level? The answer depends on its beliefs regarding its rival's behavior.

STEP Open the Excel workbook *GameTheory.xls* and read the *Intro* sheet, then go to the *Parameters* sheet.

Market demand is given by the linear inverse demand curve and, for simplicity, we assume a linear total cost function. This means that $MC = AC$ is a horizontal line.

STEP Proceed to the *PerfectCompetition* sheet.

With many small PC firms, the industry as a whole will produce where demand intersects supply (which is the sum of the individual firm's MCs). The graph shows that a perfectly competitive market will produce 15,000 kwh at a price of 5¢/kwh.

What happens if a single firm takes over the entire market?

STEP Proceed to the *Monopoly* sheet. Use the *Choose Q* slider control to determine the profit-maximizing quantity. Keep your eye on cell B18 as you adjust output. The optimal output is found where $MR = MC$.

The monopolist will produce 7500 kwh and charge a price of 12.5¢/kwh. This solution nets a maximum profit of 56,250 cents.

Not surprisingly, compared to the perfectly competitive results, monopoly results in lower output and higher prices.

Cournot was the first to ask the question, "What happens if the industry is shared by two firms?"

To understand the answer, the concept of residual demand is crucial because it enables us to solve the firm's optimization problem. *Residual demand* is the demand curve facing the firm after the sales from the other firm are subtracted. From there, the reaction function for each firm is derived from a comparative statics analysis. The two reaction functions are then combined to yield the Nash equilibrium, which is the answer to Cournot's question. That is confusing. We turn to Excel to see each step and how it all works.

Residual Demand

To figure out the quantity and price combination with two competing firms, we need to understand how the firms will behave.

STEP Proceed to the *ResidualDemand* sheet.

This sheet shows how Firm 1 decides what to do, given Firm 2's output decision. Think of the chart as belonging to Firm 1. It will use this chart to decide what to do, given different scenarios.

Conjectured Q2, in cell B14, is the key variable. A conjecture is an educated guess. It is based on incomplete information. Firm 1 does not know and cannot control what Firm 2 is going to do. Firm 1 must act, however, so it treats Firm 2's output decision as a conjecture and proceeds based on that projected value.

Conjectured Q2 is an exogenous variable for Firm 1. It does not know what Firm 2 will do and cannot control it. The conjectured output of Firm 2 may be different from Firm 2's actual output. Firm 1 can, however, examine how it would react to different possible values of Firm 2 output.

The *ResidualDemand* sheet opens with *Conjectured Q2* $= 0$. In this scenario, Firm 2 produces nothing and Firm 1 behaves as a monopolist, producing 7,500 kwh and charging a price of 12.5¢/kwh.

STEP Click five times on the scroll bar in cell C14. With each click, *Conjectured Q2* rises by 1,000 units and the red lines in the graph shift left.

The red lines are the critical factor for Firm 1. They represent residual demand and residual marginal revenue. The idea behind residual demand is that Firm 2's output will be sold first, leaving Firm 1 with the rest of the market.

The residual in the name refers to the fact that Firm 2 will supply a given amount of the market and then Firm 1 is free to decide what to do with the demand that is left over.

With each click, Firm 2 was producing more and so the demand left over for Firm 1 was falling. This is why the residual demand shifts left when Firm 2 produces more.

As the *Parameters* sheet shows, the inverse demand curve for the entire market is given by the function $P = 20 - 0.001Q$. If *Conjectured Q2* = 5,000, then the residual inverse demand curve is $P = 20 - 0.001Q - 0.001(5000)$. In other words, we subtract the amount supplied by Firm 2. Thus, the residual inverse demand curve is $P = 15 - 0.001Q$.

Figure 16.2 shows how the residual demand is shifted left by 5,000 kwh when *Conjectured Q2* is 5,000. The key idea is that Firm 2's output is subtracted from the demand curve and what is left over, the residual, is the demand faced by Firm 1.

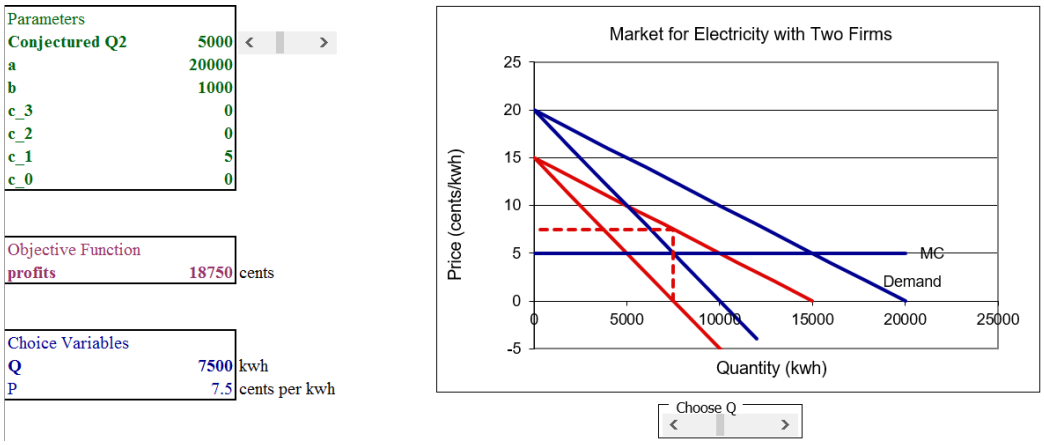


Figure 16.2: Residual demand.
Source: GameTheory.xls!ResidualDemand

Once we have residual demand for Firm 1, we can find the profit-maximizing solution. Firm 1 derives residual MR from its residual demand curve and uses this to maximize profits by setting residual $MR = MC$. In Figure 16.2, Firm 1 is not maximizing profits by producing 7,500 units and charging 7.5¢/kwh. Notice that the price is read from the residual demand curve, not the full market demand curve.

STEP Use the scroll bar (below the chart) to find Firm 1s optimal solution when *Conjectured Q2* is 5,000.

You should have found that optimal Q is 5,000 kwh, optimal $P = 10$¢/kwh and maximum π are 25,000 cents.

The Reaction Function

Now that we know how the duopolist uses residual demand to choose the quantity (and price) that maximizes profits, we can proceed to the next step in answering Cournot's question: "What happens if the industry is shared by two firms?"

We track each duopolist's optimal output as a function of *Conjectured Q2*. This gives the *reaction (or best response) function*. The reaction function is a comparative statics analysis based on shocking *Conjectured Q2*.

STEP Fill in the table in the *Residual Demand* sheet. You are picking points off of Firm 1's reaction function.

You already have two of the rows. In addition to the optimal solution at *Conjectured Q2* = 5,000 which we just found, when *Conjectured Q2* = 0, optimal output is 7,500 and optimal price is 12.5¢/kwh. Fill in the rest of the table.

STEP Check your work by clicking the Check Table button.

The filled in table is giving us Firm 1's reaction function. It is similar to the output of the CSWiz—the leftmost column is the exogenous variable and the other columns are endogenous responses.

Deriving Firm 1's reaction function is an important step in figuring out how two firms will interact. The reaction function gives us Firm 1's optimal response to Firm 2's output decision. We do not know, however, what Firm 2 will actually do. It has a reaction function just like Firm 1. The two firms must interact to determine what will happen in the market.

Finding the Nash Equilibrium

Residual demand enabled us to understand the reaction function. We are now ready for the third and final step so we can answer Cournot's question concerning the results of a duopoly. Remember, perfect competition gives 15,000 kwh of output and monopoly gives only 7,500 (and at a higher price). Presumably, duopoly is between them, but where?

STEP Proceed to the *Duopoly* sheet.

The display is new, but easy to understand. Instead of working with just Firm 1, both are shown. They have the same costs.

The sheet has buttons that make it a snap to see what each firm will do. The analytical solution is used so you do not have to run Solver every time *Conjectured Q2* changes.

STEP Notice that *Conjectured Q2* (in cell B13) is zero. To find the optimal solution, click the Choose q_1* button.

Not surprisingly (given our earlier work with the residual demand graph) since *Conjectured Q2* is zero, Firm 1 chooses to produce 7500 kwh.

But look at cell G13—Firm 1 has optimized, but now we need to ask what Firm 2 would do if Firm 1 made 7,500 kwh? Firm 2 wants to maximize profits just like Firm 1.

STEP Click the Choose q_2* button.

Firm 2's solution makes sense. If Firm 1 makes 7,500 kwh, then Firm 2 maximizes profits by taking the residual demand and producing 3,750 kwh. Their combined output means $P = 8.75$.

This is not, however, an equilibrium solution because Firm 1 is not going to produce 7,500 kwh. Why not?

STEP Look at cell B13. Click on cell B13.

B13's formula, =G20, makes clear how Firm 1's decision is connected to its rival. If Firm 2 says it wants to produce 3,750, then Firm 1 regrets and will change its previous choice. We need to find the optimal output for Firm 1 given Firm 2's new level of output.

STEP Click the Choose q_1* button.

Firm 1 chooses to make 5,625 kwh (based on Firm 2's output of 3,750 kwh), but now we return to Firm 2. Will it produce 3,750 kwh? No. When Firm 1 changed its output, cell G13 updated. Like B13, G13 connects Firm 2's optimal decision to Firm 1's output choice.

It is Firm 2's turn to regret its previous decision. Firm 2 can make higher profits by changing its output when Firm 1 makes 5,625 kwh. How much will Firm 2 want to produce? Let's find out.

STEP Click the Choose q_2* button.

Firm 2 is set, but what about Firm 1? Does it regret making 5,625? Yes, it does because it can make higher profits by changing its decision.

We will not be in equilibrium until both firms are happy with their output choice and do not wish to change it. Since Firm 2 changed its output, Firm 1 will want to change its output.

STEP Click the Choose q_1* button.

You might be thinking that this will never end. That is incorrect. It will end. You can actually see it end.

STEP Repeatedly move back and forth, clicking the Choose q_2* and Choose q_1* buttons, one after the other. What happens?

After repeatedly clicking, you are looking at convergence. Clearly, the two optimal output levels closed in on 5,000—this is the Nash equilibrium solution to this problem and the answer to Cournot's question. The duopoly will produce a combined total of 10,000 kwh with a price of 10¢/kwh. This makes sense since it is in between the perfectly competitive (15,000 kwh) and monopoly outcomes (7,500 kwh).

Manually optimizing for each firm in turn, back and forth, until the equilibrium solutions comes into focus is a great way to understand the concept of a *Nash equilibrium*. It is a position of rest where neither firm regrets its previous decision. In fact, a Nash equilibrium is often referred to as a "no regrets" point. There is, however, a faster way to find the position of rest.

STEP click the Nash Equilibrium button.

This button does all of the hard work for you. It alternately solves one firm's problem given the other firm's output many times. It continues to maximize firm profits until there is less than a 0.001 difference between a firm's optimal output and its optimal output based on the conjectured output of its rival.

STEP To see this, click on cells B20 and G20. They are close to 5,000, but not exactly 5,000.

The Nash Equilibrium button also displays the individual firm's reaction functions (scroll down if needed). In this case, the two reaction functions are identical.

Finally, the Nash Equilibrium button shows the two reaction functions on the same chart and the intersection instantly reveals the Nash equilibrium. Figure 16.3 shows the Nash equilibrium chart with additional elements to help explain it.

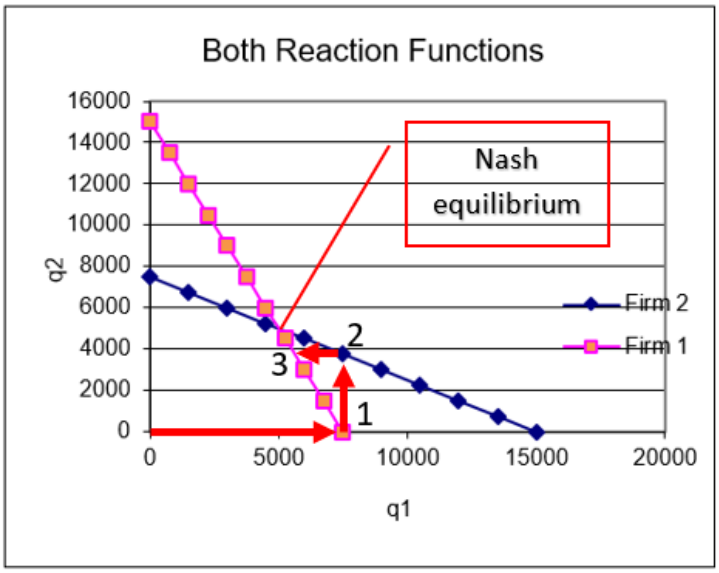


Figure 16.3: Nash equilibrium.
Source: GameTheory.xls!Duopoly

Point 1 in Figure 16.3 represents the first time Firm 1 maximized profits, with *Conjectured Q2* of zero. Point 2 shows Firm 2's optimization based on Firm 1 making 7,500 kwh. You can see, by following the arrows, how this would lead to the intersection as the Nash equilibrium.

You might wonder why the reaction functions are not the same in Figure 16.3 since they are identical when graphed by themselves (as shown below the buttons in the *Duopoly* sheet). The answer lies in the axes—to plot them

both on the same graph, we use the reaction function for Firm 2 and the *inverse* reaction function for Firm 1. Scroll down to see the inverse reaction function starting in row 63.

Remember: A Nash equilibrium exists when each player, observing what her rivals have chosen, would not choose to alter the move she herself chose. Nash equilibrium is a no regrets point for all players.

Figure 16.3 shows that the Nash equilibrium is at the intersection of the two reaction functions. Only there will both firms decline the offer to change their optimal decisions. This is a position of rest.

Evaluating Duopoly's Nash Equilibrium

We know the answer to Cournot's question. Duopoly, at its Nash equilibrium, leaves us in between perfect competition and monopoly. But we can say more about the duopoly outcome. We focus on profits.

STEP In cell D16 in the *Duopoly* sheet, enter a formula that adds the profits of the two firms at the Nash equilibrium. What are industry profits?

You might recall monopoly had maximum profits of 56,250 cents. That is better than the 50,000 cents you just computed with your formula in cell D16 of = B16 + G16.

Can duopolists increase their profits to 56,250 like a monopolist? Yes, they can, but they will not be able to honor their commitments.

STEP Set quantities for both firms (in cells B20 and G20) to 3,750. What happens to profits?

Amazingly, they go up. If the two rivals can agree to simply split the monopoly output of 7,500 kwh, each will make 28,125 cents and match the monopoly outcome.

But this will not last. Why not? Why don't the two firms get together and produce 3,750 units each and make greater joint profits than the Nash equilibrium solution? A single click reveals the answer.

STEP Click the Choose q_1* button or the Choose q_2* button.

If the rival makes 3,750 kwh, the firm maximizes profits at 5,625 kwh. In other words, they have an incentive to cheat—just like in the Prisoner's Dilemma game.

As soon as one takes advantage, the other fires back and they spin back to the Nash equilibrium.

You might suggest writing a contract, but that is illegal and unenforceable in the United States. There are other options and strategies, but they would take us too far from Intermediate Microeconomics. One strong attraction that is easy to see is merger. If the two firms combine into a single entity, they will be a monopoly and enjoy monopoly profits. Presumably, the Department of Justice would object.

Interdependence

Game theory is an exciting, growing area of economics. Its primary appeal lies in the realistic modeling of agents as strategic decision makers playing against each other, moving and countering. This is obviously what a real-world firm does.

The Cournot model is a simple game matching two firms against each other. It illustrates nicely the notion of interdependence and how one firm moves, and then the other responds, and so on. Whereas some games do not have a Nash equilibrium, the Cournot duopolists do settle down to a position of rest.

The *Summary* sheet has the outcomes from perfect competition, duopoly, and monopoly. It is clear that monopoly maximizes firm profits, but perfect competition offers the consumer the lowest price and most output. We will return to this comparison in the third and final part of this book.

We have just scratched the surface of game theory. There are many, many more games. The workbook *RockPaperScissors.xls* lets you play this child's game in Excel. Section 17.7 on Cartels and Deadweight Loss has another application of game theory.

For an entertaining version of the Prisoner's Dilemma in a game show, see this *Golden Balls* episode finale: tiny.cc/splitsteal. And for a really clever twist, watch this one: tiny.cc/ibrahim. Nick's strategy has been outlawed from the show. The Cornell game theory blog has an entry explaining it: tiny.cc/splitstealanalysis.

Exercises

These exercises are based on $c_1 = 5$. If you did the *Q&A* questions and changed this parameter, change it back to its original value.

1. If *Conjectured Q2* is 15,000, why does Firm 1 decide to produce nothing? Use the *ResidualDemand sheet* to support your explanation.

2. Suppose Firm 1 produces 4,500 kwh and Firm 2 produces 6,000 kwh. Does Firm 1 have any regrets? Does Firm 2 have any regrets? Enter these two values in the *Duopoly* sheet and click the Choose q buttons. Which firm changed its mind? Why?

3. Click the Reset All button in the *Duopoly* sheet. Explore the effect of changing Firm 1's cost function so that c_2 (cell B10) is 0.001 (with B11 = 5). How does this affect the Nash equilibrium?

References

The epigraph is from page 75 of Sylvia Nasar, *A Beautiful Mind* (1998). This biography of Nash has won countless awards and was made into an Academy Award-winning motion picture, with Russell Crowe starring as John Nash. Although much of the book is devoted to Nash's personal struggle with schizophrenia, Nasar's book gives a clear and engaging review of game-theoretic concepts before Nash and of the Nash equilibrium.

On the game Nash invented that is mentioned in the epigraph, Nasar writes,

> That spring, Nash astounded everyone by inventing an extremely clever game that quickly took over the common room. Piet Hein, a Dane, had invented the game a few years before Nash, and it would be marketed by Parker Brothers in the mid-1950s as Hex. But Nash's invention of the game appears to have been entirely independent. (p. 76)

A Brilliant Madness, www.pbs.org/wgbh/americanexperience/films/nash/, is an excellent 2002 documentary on Nash's life, struggles, and contributions.

In 1994, Nash, John C. Harsanyi, and Richard Selten shared the Nobel Prize "for their pioneering analysis of equilibria in the theory of non-cooperative games." (www.nobelprize.org/prizes/economic-sciences/1994/summary/)

The first edition of *The Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern was published in 1944.

Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life by Avinash K. Dixit and Barry J. Nalebuff (originally published in 1991) explains and applies game theory to a variety of interesting examples and situations.

Part III

The Market System

The Butterfly Effect
acquired a technical name:
sensitive dependence on
initial con-ditions.

James Gleick

Overview

The first part of this book was the Theory of Consumer Behavior. It modeled a consumer's utility maximization problem and emphasized deriving a Demand Curve as the key result.

The Theory of the Firm comprised the second part. Firm decisions about inputs and outputs were modeled as optimization problems. The key result was deriving a Supply Curve from the perfectly competitive firm's output profit maximization problem.

This third part will put together consumers' demand and firms' supply in an equilibrium model. This will show how individual markets solve society's resource allocation problem. In addition, we will introduce an equilibrium model that incorporates all markets simultaneously.

Before we begin, we review these three key ideas:

1. Optimization versus equilibrium.
2. Partial and general equilibrium.
3. Society's resource allocation problem.

1. Optimization Versus Equilibrium

The stress thus far has been on optimization. Consumers maximize utility, firms minimize costs and maximize profits. We have used numerical and analytical methods, including the Lagrangean, to solve these problems.

The market system, however, is an equilibrium model. There are similarities between optimization problems and equilibrium models. They both rely heavily on comparative statics and we will continue to use numerical and analytical methods, but there are critical differences.

In an optimization problem, an agent explicitly chooses, setting the values of endogenous variables. For example, a consumer picks from available options to maximize utility and a firm manipulates variables to maximize profit. An optimal solution means the best choice is made (from the decision maker's point of view).

Unlike optimization problems, equilibrium models do not have an agent directly controlling or setting values of a variable. Instead, forces within the model drive variables to positions of rest. No agent actually picks the solution in an equilibrium model. Instead, the equilibrium solution means that there is no tendency to change in the endogenous variables (those determined within the model).

The notation we will use is common in economics, but often goes unremarked and unnoticed. A star, or asterisk, means optimal. We have found $x_1{}^* = 25$ and $L^* = 1431$. The star means this value is the best value the agent can choose.

In equilibrium models, the solution is denoted by a subscript "e." We might find that $Q_e = 100$. This means that the system settles down and is at rest at this value.

Unlike optimization problems, an equilibrium solution says nothing about the desirability of the solution. In other words, we cannot conclude that an equilibrium solution is a good one simply because it is the equilibrium solution. We could be at rest at a bad place.

Finally, unlike optimization problems, economists are often interested in the equilibration process, that is, the path followed to the final resting place. If it exists, the type of convergence, direct or oscillatory, can be studied. The equilibration process is beyond the scope of this book, but it helps show the difference between optimization and equilibrium. There is no process in optimization—the agent chooses the best solution and if there is a shock, the agent instantly re-optimizes. Not so with equilibrium. A shock will put forces into play that move the system.

Confusing equilibrium with optimal is common, but bad practice. They are different in the fundamental fact that optimization has an agent choosing and equilibration does not. Never automatically assume that an equilibrium solution is optimal.

2. Partial and General Equilibrium

While all equilibrium models rely on the concept of rest or stability as a key marker of the equilibrium solution, the market system was analyzed in two fundamentally different ways:

1. Partial equilibrium: Focus on a single good or service, in isolation.
2. General equilibrium: Consider all of the goods or services together.

Partial equilibrium was made famous by Alfred Marshall. He not only popularized putting price on the y axis, he made the graphical display of supply and demand curves for individual goods and services popular, especially in the English-speaking world. It is easy to see the equilibrium solution at the intersection of the supply and demand curves and, we will see, the graph can be used to evaluate the equilibrium outcome.

In the rest of Europe, a different tradition arose. Spearheaded by increasingly sophisticated mathematical economists, such as Leon Walras and Vilfredo Pareto, a more holistic approach to the market system was developed. Instead of looking at a single product or industry, all goods and services are simultaneously analyzed.

You are already familiar with partial equilibrium because supply and demand graphs are a staple of high school and introductory economics courses. General equilibrium theory, however, will be new and challenging.

Make no mistake, they are not equal. General equilibrium is superior to partial equilibrium analysis, but it is also more complicated and difficult. In our study of the market system, we will first analyze individual markets using conventional supply and demand graphs, then we turn to general equilibrium analysis.

In both partial and general equilibrium analyses, we first determine the equilibrium solution and then judge it by comparing it to an optimal solution. We avoid the fundamental error of conflating equilibrium with optimal. We may find that an equilibrium solution is, in fact, optimal, but we will also see situations where this is not so and the market fails.

3. Society's Resource Allocation Problem

The partial and general equilibrium models explain how markets function in solving a particularly fundamental optimization problem. It is so important that is it often referred to as *The Economic Problem*.

Figure III.1 depicts the problem. Given scarce resources of labor and capital (representing all inputs), society must decide what to produce, how much of each product to make, and how to distribute the output.

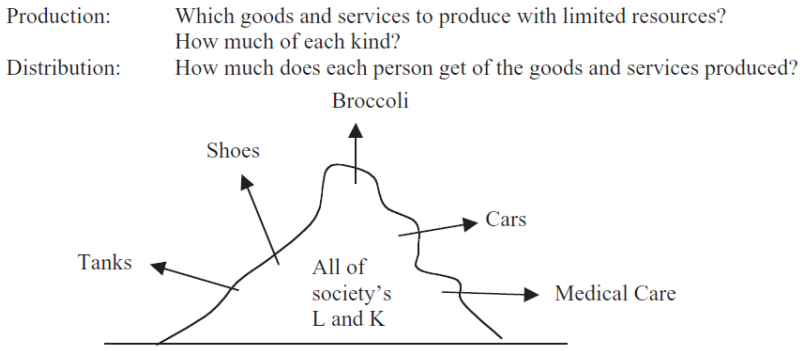


Figure III.1: Society's resource allocation problem.

This problem can be solved by tradition, authority, or the market system. Most people do not realize that the last way is a brand new approach. Of the roughly 200,000 years that humans have been on this planet, traditional and authoritarian arrangements are by far the usual ways to solve society's resource allocation problem. The market system emerged only in the last couple of hundred years.

This may seem incorrect given that money and prices have been around for a long, long time. A moment's reflection should convince you that trading is not a sufficient condition to determine whether a market system is being used to solve society's resource allocation problem. After all, societies in Biblical times had bazaars where people bought and sold goods and the former Soviet Union had stores where people paid rubles for groceries, but neither of these societies had market economies.

Cuba has had not one, but two currencies for decades (tourists must use the Cuban convertible peso or CUC, while Cubans use pesos), but no one would say it has a market economy. No, the presence of money is not a litmus test for a market system.

Societies based on the market system do not use supply and demand to allocate resources for every good and service. It is obvious that military equipment, such as tanks, in Figure III.1, are not produced according to perfectly competitive conditions via supply and demand. There is only one buyer, the government, and a few sellers (manufacturers of military vehicles). Likewise, no modern society uses the market system for medical care.

One could argue that all goods and services are regulated or controlled to some degree and, while there is some truth to this, it is also mostly true that many individual farmers decide what to grow based on market prices and this is a hallmark of the market system.

Unlike other ways of allocating resources, the market system allows each agent to decide how to use their labor and other privately owned resources. In a market system, individual resource owners respond to incentives. Unlike traditional and authoritarian systems, which rely on custom and command to get work done and products made, markets use the lure of gain to attract effort and capital.

The market system takes advantage of individual self-interest, using prices as incentives and signals. Whether self-interest is innate or learned is a deep philosophical question, but there is no doubt that players in a market system are driven to succeed and they calculate (and maximize, as they see it) before deciding what to do.

Although the market system, or simply markets, is the usual terminology today, other names have been used, such as capitalism, private property, free enterprise, price system, and *laissez faire*. Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776) is the first attempt at a comprehensive explanation of how a decentralized system that allows individual resource owners to decide where and how to use society's inputs can give a reasonable solution to society's economic problem. Notice the date—1776—before then, no one had to explain the market system because it did not exist.

This is not a history book, but you should be aware that the market system first emerged in western Europe around the 1700s, give or take a hundred years. It is difficult to pinpoint exactly where and when because there is no single event or marker. From close up, focusing on the 15^{th} to the 20^{th} centuries, it was a long, gradual transformation of society that took a few

hundred years. From far away, on a scale of centuries stretching back thousands of years, it was a sudden, explosive societal change.

One way to convey the stunning explosion in economic output before and after the emergence and spread of the market system is by examining the historical performance of different countries. We know the world was poor for millennia and then things changed fast, but economic historians have painstakingly compiled estimates of output per person to help us understand the evolution. Angus Maddison, for example, devoted his career to measuring long run economic growth around the world. The data are here: www.ggdc.net/maddison/Maddison.htm. There are output and population measures for countries all the way back to 1 AD. Figure III.2 plots real GDP per capita for 12 western European countries.

Figure III.2: Western Europe's historical economic performance.

The hockey stick depicted in Figure III.2 tells a remarkable story. Before the market system, although individual people (kings or other elites) could be rich, almost everyone knew only grinding poverty. Then, suddenly, something happened in western Europe that changed everything. Economies literally took off and the modern world was born. For an excellent, brief review of the rise of the market system, see the second chapter, "The Economic Revolution," in Robert Heilbroner's classic best-seller, *The Worldly Philosophers*.

The intellectual history of research on capitalism and markets is also quite fascinating. A great deal of work revolves around the idea of patterns emerg-

ing without direct, top-down control. Smith invoked the image of an "invisible hand" and Nobel Prize winning economist Friedrich Hayek coined the oxymoron "spontaneous order." In mathematics today, nonlinear dynamics and chaos theory focus on "self-organizing behavior." This idea, a pattern out of nothing, is critical to understanding the market system and the role of supply and demand.

Many have noticed that birds fly in a V, ants can form long chains and never seem to get stuck in traffic, and many animals (bees, locusts, and fish) swarm—they seem to act as if they had a collective mind. How do they do it? They do not rely on a single command center or leader to tell each one what to do. There are no orders given. There is no central direction. Instead, each individual follows simple rules that, taken together, produce a pattern or coherent order.

In computer science, the Game of Life is an artificial world that produces patterns from trivially simple rules. There are many examples on the web, such as this recent one, in honor of John Conway who recently passed away: b3s23life.blogspot.com/2020/01/a-gentleman-and-scholar.html. Search "game of life excel" for spreadsheet versions. LifeWiki has history, explanations, and many examples: www.conwaylife.com/wiki.

The point is that complicated movements of gliders and other objects that would seem to require central control can be generated in a decentralized way. Thus, the Game of Life is just another application—like supply and demand—of the general principle that patterns can be formed not only by top-down direction (like a marching band), but by decentralized systems with no controller at all.

The difficult idea to grasp is that supply and demand analysis is more than two intersecting lines. We are actually studying a pattern-generating system. Supply and demand is the model used by economists to explain how multitudes of interacting agents in markets can solve society's incredibly complicated resource allocation problem.

For the purposes of understanding how the market system works, an individual market will be defined by the commodity bought and sold. Thus, there is a market for broccoli and a market for engineers and a market for tutors. Every good and service allocated by the market system has a supply and demand.

By having a market for each product, we can use each individual market's equilibrium output as the market system's answer to the resource allocation problem. There is no central planner or controller who decides how much gas to produce. Buyers and sellers interact and establish an equilibrium price and quantity that determines how much of society's scarce resources are devoted to gas. If you think the price of gas is too high or we are allocating too much of society's scarce resources to producing gas, there is no one to call. The price and output are determined by the decentralized market system based on the operation of supply and demand.

The idea that a pattern emerges from the interaction of agents is fundamental to the market system. Watch the *The Invisible Hand and the Market System*, freely available at vimeo.com/econexcel/invisiblehand, to get a deeper understanding of these issues.

Organization

The organization of material in this part is straightforward, perhaps deceptively so. Figure III.3 shows the overall view of the book and the two chapters in this part.

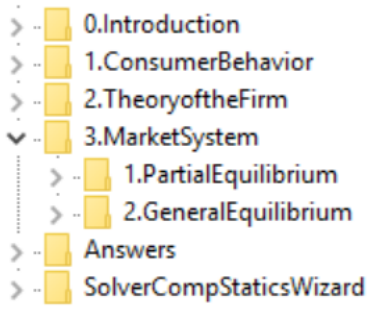


Figure III.3: Content map with focus on the market system.

The partial equilibrium chapter has sections devoted to both theory and applications (including government interventions) of supply and demand analysis. General equilibrium presents only exchange to give you a glimpse of how the model works.

It seems odd to say, but we will ignore a critical, fascinating part of the market system. Even a casual observer would notice that the market system exhibits high rates of innovation and technological change (which is what

produced the striking Figure III.2), but we will limit our analysis to exploring how the market system functions in a static environment in which the only issue is resource allocation (given constant technology).

In the conclusion, after you have mastered partial and general equilibrium analysis, we will return to the question of the dynamic analysis of the market system.

References

The epigraph is from page 23 of James Gleick, *Chaos: The Making of a New Science* (New York: Penguin Books, 1987). This serves as an excellent, friendly introduction to nonlinear dynamics and chaotic systems.

As mentioned in Chapter 3, the online version of *The Wealth of Nations* by Adam Smith, is freely avaiable at www.econlib.org/.

Robert Heilbroner, *The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers* (New York: Touchstone, 1999, 7th edition, originally published 1953), remains one of the best summaries of the history of economics and the market system.

The Game of Life is associated with John Conway, who passed away in April of 2020. See www.nytimes.com/2020/05/16/science/john-conway-math.html for a tribute to this "mathemagician," with many links for further exploration.

Chapter 17

Partial Equilibrium

Supply and Demand

Consumers' and Producers' Surplus

Tax Incidence and Deadweight Loss

Inefficiency of Monopoly

Sugar Quota

Externality

Cartels and Deadweight Loss

Signaling Theory

> Credit for the ubiquitous demand and supply diagrams in principles texts is usually given to Fleeming Jenkin [1870]. ...For the first time, a real visual sense of the market is located. Pride of place goes to the equilibrium price.
>
> Judy Klein

17.1 Supply and Demand

We begin our analysis of the market system by making an obvious, but necessary point: A market demand (or supply) curve is the sum of individual demand (or supply) curves.

STEP Open the Excel workbook *SupplyDemand.xls*, read the *Intro* sheet, then go to the *SummingD* sheet.

The sheet has three consumers, with three different utility functions and different incomes. We assume the consumers face the same prices for goods 1 and 2. We set $p_2 = 10$, but leave p_1 as a variable to derive the individual demand curve for each consumer.

STEP Confirm, by clicking on a few cells in the range B18:D22, that the formulas in these cells represent the individual demand curves for each consumer. Notice that the graphs below the data represent the individual demand ($x_1{}^* = f(p_1)$) and inverse demand ($p_1 = f(x_1{}^*)$) curves.

Given individual demands, market demand can be found by simply summing the optimal quantity demanded at each price.

STEP Confirm, by examining the formula in cell E18, that market demand has been computed by adding the individual demands at $p_1 = 1$. The same, of course, holds true for the other points on the market demand curve.

Because we often display demand schedules as inverse demand curves, with price on the y axis, the red arrow (see your screen and Figure 17.1) shows that market demand is the result of a horizontal summation. At $p_1 = 5$, we read off each of the individual quantities demanded and add them together to obtain the market quantity demanded of 24.3 units.

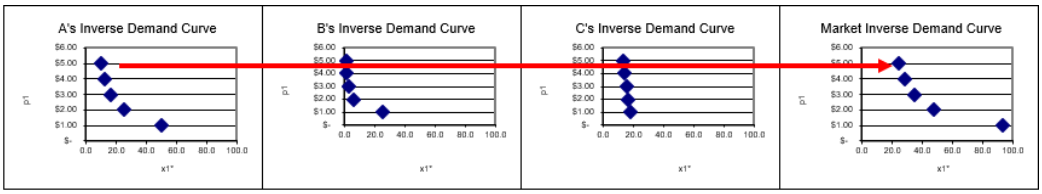


Figure 17.1: Horizontal summation to get market demand.
Source: SupplyDemand.xls!SummingD.

Supply works just like demand. We add individual supply curves (horizontally if we are working with inverse supply curves) to get the market supply curve. Because individual supply curves are P above AVC, we know that the market supply curve is simply the sum of the marginal costs above minimum AVC of all the firms producing the particular good or service sold in this market.

So the way it works is that each of the individual buyers and sellers optimizes to decide how much to buy or sell at any given price. The Theory of Consumer Behavior and the Theory of the Firm are the sources of individual demand and supply.

Once we have the many individual demand and supply curves, we add them up. So market demand and supply are composed of the sum of many individual pieces. Some consumers want a lot of the product at a given price, while others want less (or maybe none at all), but they all get added together to form market demand. The same is true for supply.

Initial Solution

The next step is obvious: market supply and demand are combined to generate an equilibrium solution that determines the quantity produced and consumed. This equilibrium solution is the market's answer to society's resource allocation problem.

The simple story is that price adjusts, responding to surpluses and shortages, until it settles down at its equilibrium level, where quantity demanded equals quantity supplied. This is the intersection of the two curves.

It is confusing, but true that in the supply and demand model, price and quantity are endogenous variables. How can price be endogenous—don't

consumers and PC firms take the price as given? Yes, they do and for individual buyers and sellers, price is exogenous, but, for the system as a whole, price is endogenous.

At the individual agent level, price is given and cannot be controlled by the agent so it is exogenous. But we are now at a different level. We are allowing forces of supply and demand to move the price until it settles down. Thus, at the level of the market, we say price is endogenous because it is determined by forces within the system.

It is worth repeating that equilibrium means no tendency to change. When applied to the model of supply and demand, equilibrium means that price (and therefore quantity demanded and supplied) has no tendency to change. A price that does have a tendency to change (because there is a surplus or shortage) is a disequilibrium price.

We can put these ideas in the same framework that we used to solve optimization problems. There are two ways to find the equilibrium solution and they yield the same answer:

1. Analytical methods using algebra: conventional paper and pencil.

2. Numerical methods using a computer: for example, Excel's Solver.

STEP Proceed to the *EquilibriumSolution* sheet to see how the supply and demand model has been implemented in Excel.

The information has been organized into three main areas: endogenous variables, exogenous variables, and an equilibrium condition. Excel's Solver will be used to find the values of the endogenous variables that meet the equilibrium condition.

As usual, green represents exogenous variables, the coefficients on the demand and supply curves.

Although price and quantity are both endogenous variables, price is bolded to indicate that the model will be solved by finding the equilibrium price and then the equilibrium quantity (demanded and supplied) is determined. This is similar to the approach we took with monopoly where we maximized profits by choosing q, then found P from the demand curve.

Finally, the equilibrium condition is represented by the difference between quantity demanded and supplied.

On opening, the price is too high. At $P = 125$, quantity demanded (Q_d) is 112.5 and Q_s is about 173. Thus, have a surplus ($Q_d < Q_s$) and, therefore, price is pushed down (as firms seek to unload unsold inventory).

STEP Use the scroll bar next to the price cell to set the price below the intersection of supply and demand. The dashed line (representing the current price) responds to changes in the price cell (B12).

Notice how the quantity demanded and supplied cells also change as you manipulate the price, which makes the equilibrium condition cell (B17) change.

With P below the intersection, the market experiences a shortage ($Q_d > Q_s$) and price is pushed up. The force in the market model is the pressure generated by surpluses (excess supply) or shortages (excess demand).

Obviously, the equilibrium price is found where supply and demand intersect. At this price, there is no tendency to change. The forces of supply and demand are balanced. We can find this price by adjusting the price manually and keeping our eye on the chart or by using Excel's Solver.

STEP Open Solver.

The Solver dialog box appears, as shown in Figure 17.2. Notice that the objective is not to Max or Min, but to set an equilibrium condition equal to zero. Notice also that P, price, is being used to drive the market to equilibrium and there are no constraints.

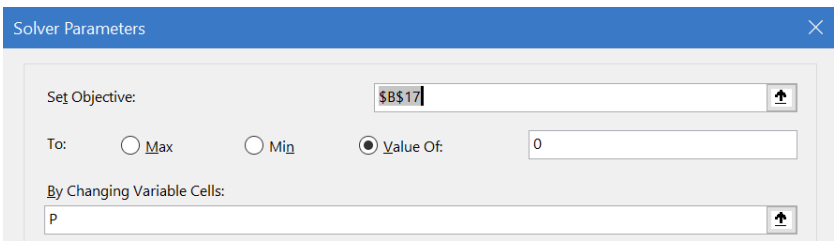


Figure 17.2: Solver dialog box.
Source: SupplyDemand.xls!EquilibriumSolution.

STEP Click Solve to find the equilibrium solution.

The chart makes it easy to see that Solver is correct. At $P = 100$, $Q_d = Q_s = 125$. Without a surplus or shortage, there is no tendency for the price to change and we have found the equilibrium resting point.

The equilibrium quantity, 125 units, is the market's answer to society's resource allocation problem. It says that we should send enough resources from the scarce, finite amount of inputs available to produce 125 units of this product.

We envision a supply and demand diagram for every product and the equilibrium quantity, in each market, is the market's answer to how much we should have of each commodity.

The analytical approach is easier than the math we applied for optimization problems because there is no derivative or Lagrangean. All we need to do is find the intersection of supply and demand.

Given either market supply and demand curves $Q = f(P)$ or inverse supply and demand functions, $P = f(Q)$, we find the equilibrium solution by setting supply and demand equal to each other.

The inverse functions in the Excel workbook are:

$$P = 350 - 2Q_d$$

$$P = 35 + 0.52Q_s$$

Setting the inverse functions equal to each other, we replace the Q_d and Q_s with Q_e because we are finding the value that lies on both of the curves:

$$350 - 2Q_e = 35 + 0.52Q_e$$

$$385 = 2.52Q_e$$

$$Q_e = \frac{315}{2.52} = 125$$

Substituting this solution into either inverse function yields $P_e = 100$.

We can also easily flip the inverse functions, solving for Q in terms of P, to obtain the demand and supply functions:

$$P = 350 - 2Q_d \rightarrow 2Q_d = 350 - P \rightarrow Q_d = 175 - \frac{1}{2}P$$

$$P = 35 + 0.52Q_s \rightarrow 0.52Q_s = P - 35 \rightarrow Q_s = \frac{1}{0.52}P - \frac{35}{0.52}$$

If we set demand equal to supply, using P_e to denote the common value we seek, we find the equilibrium price:

$$175 - \frac{1}{2}P_e = \frac{1}{0.52}P_e - \frac{35}{0.52}$$

$$175 + \frac{35}{0.52} = \frac{1.26}{0.52}P_e$$

$$P_e = \frac{175 + \frac{35}{0.52}}{\frac{1.26}{0.52}} = 100$$

Plugging this equilibrium price into either function gives $Q_e = 125$.

This work shows something obvious, but worth making clear: we can use $P = f(Q)$ functions to find Q_e, then P_e or we can use $Q = f(P)$ functions to find P_e, then Q_e. We get the same result either way since we are merely flipping the axes.

If you think using supply and demand functions ($Q = f(P)$) to get P_e and then Q_e is more faithful to what is going on in the market, you are a Marshallian for that is exactly how he saw markets functioning. And that is why P is on the y axis—so the reader sees it fluctuate up and down until it settles down to its equilibrium value.

We finish our work on the initial solution by pointing out that it is not surprising that numerical methods, using Solver, agree with the analytical approach. Given supply and demand for this product, we know that the market equilibrium solution would call for producing 125 units. The market system would, therefore, allocate the labor and capital needed to make this amount.

Elasticity

We can compute the price elasticity of demand and supply at the equilibrium price (the point elasticity) by applying our usual formula, $\frac{dQ}{dP}\frac{P}{Q}$. This time, we must use the demand and supply curves, $Q = f(P)$.

STEP Click the Show Point Elasticity button to see the calculation.

Although it has text wrapped around it, the number displayed for the price elasticity of demand is based on this part of the formula: $(-1/d1) * (P/Qd)$. With $Q_d = \frac{d_0}{d_1} - \frac{1}{d_1}P$, it is easy to see that $\frac{dQ}{dP} = -\frac{1}{d_1}$ and then we multiply by $\frac{P}{Q}$. Likewise, the price elasticity of supply is the slope of the supply function times the $\frac{P}{Q}$ ratio.

At the equilibrium price and quantity, demand is much more price inelastic than supply. This does not matter right now, but it will in future work.

STEP With $P = 100$, click on the price scroll bar and watch the price elasticities. Keep clicking until you set $P = 125$.

As you increase price, the elasticities change. Even though the slopes are constant, the supply and demand elasticities change because the $\frac{P}{Q}$ ratio is changing. Multiplying the slope by a price-quantity coordinate produces a percentage change measure of responsiveness.

The price elasticity of demand at $P = 125$ is -0.56 means that a 1% increase in the price leads only to a 0.56% decrease in the quantity demanded. This means demand is not very responsive since the percentage change in quantity is less than the percentage change in the price. Notice, however, the demand is more responsive at $P = 125$ than it was at $P_e = 100$.

We will see in future applications of the supply and demand model that the price elasticities play crucial roles. For now, remember that slope and elasticity are not the same and that the price elasticity tells us how responsive quantity demanded or supplied is to a change in the price.

Economists can be sloppy and say things like "demand is elastic" or "inelastic supply." This, of course, is nonsense. All downward-sloping, linear, inverse demand curves that cut both axes have elasticities that range from negative infinity at y-intercept to zero at the x-intercept. Statements like "demand is elastic" typically refer to a specific, usually equilibrium, price.

Long Run Equilibrium

Another concept at play in the model of supply and demand is that of long run equilibrium.

In the long run (when there are no fixed factors of production), a competitive market has another adjustment to make. In addition to responding to pressure from surpluses and shortages, the market will respond to the presence of non-zero profits.

The story is simple. Excess profits (economic profits greater than zero) will lead to the entry of more firms. This will shift the inverse supply curve right, lowering the price until all excess profits are competed away.

If the long run price is too low, firms suffering negative profits will exit, shifting the inverse supply curve left and raising prices. Thus, a long run competitive equilibrium has to look like Figure 17.3.

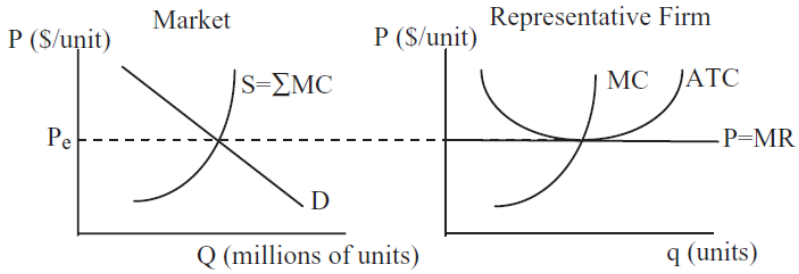


Figure 17.3: Long run equilibrium.

The left panel in Figure 17.3 shows supply and demand in the market as a whole, while the right panel depicts a single firm that is just one of the many firms in this perfectly competitive industry. The two graphs have the same y axis, but the scale of the x axis is different. A single firm can only produce a few units (q), but "millions" (an arbitrary number chosen just as an example) are bought and sold in the market (uppercase Q for emphasis). The idea is that there are many firms, each producing small amounts of the same output. In the aggregate, they make "millions" of units, but one individual firm produces only a tiny amount of the total.

Notice how the market demand curve is downward sloping, but the firm's demand curve is horizontal. This is the classic price taking environment in which a PC firm operates. Notice also that the market supply curve is the sum of the individual firm MC curves because individual firm supply is MC where $P > AVC$. We could chop off the bottom of the market supply curve (below P_e), but that would be confusing.

The long run adjustment process endogenizes the number of firms. This means that forces within the model determine how many firms there will be. This is not true in the short run, where the number of firms is assumed fixed (although they can shutdown if $P < AVC$) and the only adjustment is that market surpluses and shortages are eliminated by price movements.

Notice that the long run equilibrium price meets two equilibrium conditions:

1. Quantity demanded equals quantity supplied so there is no surplus or shortage in the market.

2. Economic profits are zero so there is no incentive for entry or desire to exit.

Long run equilibrium is even more fanciful and unrealistic than our abstract models of the consumer and firm. There has never been and never will be a market in long run equilibrium. Its primary purpose is as an indicator of where a market is heading.

The long run equilibrium model tells us that even though we are at an equilibrium with no surplus or shortage (such as with $P_e = 100$ in the Excel workbook), further adjustments will be made depending on the profit position of the firms. If profits are positive, entry will increase supply and lower price; while negative profits will lead to exit, decreased supply and higher prices.

In the Excel workbook, we do not know if the market is in long run equilibrium when $P_e = 100$ because we do not have a representative firm with its cost curves so we can determine its profits.

A key takeaway is that, like price, the number of firms is endogenous in the long run because there are forces in the model that determine its value. No one sets the number of firms. The interaction of buyers and sellers is generating the number of firms as an equilibrium outcome.

Comparative Statics

Comparative statics analysis with the supply and demand equilibrium model is familiar. Most introductory economics courses emphasize shifts in supply and demand. Here is a quick review, with special emphasis on equilibrium as an answer to society's resource allocation problem.

A change in any variable that affects supply or demand, other than price, causes a *shift* in the inverse supply or demand curve. A change in price causes a *movement along* stationary supply and demand curves. An increase in demand or supply means a rightward shift in inverse demand or supply.

For demand, the shift factors are income, prices of other goods related in consumption (i.e., complements and substitutes), tastes, consumers' expectations about future prices, and the number of buyers. The usual shift factors for supply include input prices, technology, firms' expectations, and the number of sellers.

As usual, comparative statics analysis consists of finding the initial solution, applying the shock, determining the new solution, and comparing the initial to the new solution. In the case of supply and demand, we want to make statements about the changes in equilibrium price and quantity. P_e and Q_e are the endogenous variables in the equilibrium model and we track how they respond to shocks.

For example, new technology lowered costs, What would that do to equilibrium price and quantity? We can use the *EquilibriumModel* sheet to see what happens.

STEP Make sure $P = 100$ so the market is in equilibrium, then click on the *s0* slider to lower the inverse supply curve intercept to 15.

The graph updates as you change the *s0* and a new, red inverse supply curve appears. The original, black line remains as a benchmark, but there is only one demand and supply at any point in time.

At $P = 100$, there is a surplus. We need to find the new equilibrium solution.

STEP Run Solver to find the new P_e and Q_e.

Figure 17.4 shows the result. The equilibrium price falls (from \$100/unit

to roughly \$84/unit) and the equilibrium quantity rises from (from 125 to about 133 units).

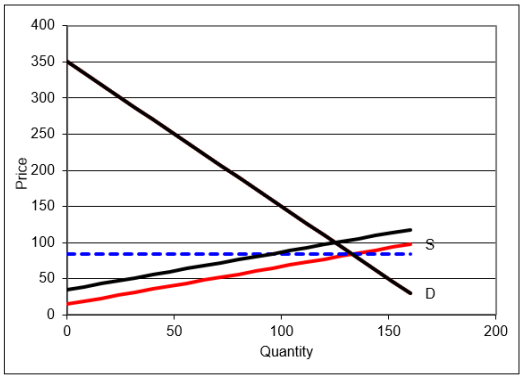


Figure 17.4: Comparative statics with the supply and demand model. *Source: SupplyDemand.xls!EquilibriumSolution.*

The decentralized market system has generated a new answer to society's resource allocation problem. Ceteris paribus, if a product enjoys a productivity increase from a new technology, making it cheaper to produce the product, the system will produce more of it.

This response makes common sense, but it is absolutely critical to understand that the increase in output is not decreed from on high. It is bubbling up from below—output rises because supply shifts and market forces lower prices and raise output.

We do not examine the equilibration process from the initial to the new solution when doing comparative statics analysis. We might directly converge to the new equilibrium, with price falling gradually until $Q_d = Q_s$. Or, price might collapse, falling below the equilibrium price, then rising above it, and so on. This would be oscillatory convergence.

With comparative *statics*, however, the focus is entirely on comparing the new to the initial solution. We may, in fact, be interested in the path to the new equilibrium, but that would take us into comparative *dynamics*—a topic for advanced microeconomics.

Applying Supply and Demand

To escape the usual trap of thinking of supply and demand in purely graphical terms, we apply the model to a real world example. We avoid graphs completely and focus on the mechanics and logic of supply and demand.

The market system uses supply and demand for outputs *and* inputs. This example focuses on labor, but there are many applications of supply and demand for capital—perhaps the stock market is the most prominent.

Consider that most fans of American football would not know the second highest paid position in the NFL. Everyone knows quarterbacks are the highest paid, but what position is second? Are star running backs, wide receivers, or maybe linebackers then next highest paid? No, the answer is left tackles—www.spotrac.com/nfl/positional/.

In *The Blind Side: Evolution of a Game* (2006), Michael Lewis explains that free agency, allowing players to sell their services to the highest bidder, radically altered the pay structure of the NFL. How did this happen? Supply and demand.

First, Lewis (p. 33) explains, there is little supply for the left tackle position.

> The ideal left tackle was big, but a lot of people were big. What set him apart were his more subtle specifications. He was wide in the ass and massive in the thighs: the girth of his lower body lessened the likelihood that Lawrence Taylor, or his successors, would run right over him. He had long arms: pass rushers tried to get in tight to the blocker's body, then spin off it, and long arms helped to keep them at bay. He had giant hands, so that when he grabbed ahold of you, it meant something.
>
> But size alone couldn't cope with the threat to the quarterback's blind side, because that threat was also fast. The ideal left tackle also had great feet. Incredibly nimble and quick feet. Quick enough feet, ideally, that the idea of racing him in a five-yard dash made the team's running backs uneasy. He had the body control of a ballerina and the agility of a basketball player. The combination was just incredibly rare. And so, ultimately, very expensive.

In addition to low supply, there is high demand. The left tackle is charged with protecting the quarterback's blind side, the direction from which defensive ends and blitzing linebackers come shooting in, causing sacks, fumbles, and worst of all, injuries. Because the quarterback is the team's most prized asset, the left tackle position is a highly sought-after bodyguard.

But even more surprising than the fact that blind side tackles are the second highest paid players in the NFL is that this was not always the case. Lewis reports that for many years, linemen were low paid, as shown in Figure 17.5.

Linemen	$398,000
Wide receiver	$504,000
Defensive end	$551,000
Running back	$620,000
Quarterback	$1,250,000

Figure 17.5: NFL salaries in 1990, before free agency.
Source: Lewis, p. 227.

So, why do blind side tackles make so much money today? NFL players did not enjoy free agency until the 1993 season. Up to that time, players were drafted or signed by teams and could move only by being traded.

Then the players' union and team owners signed a contract that enabled free agency for players so they could move wherever they wanted. In return, the players agreed to a salary cap that was a percentage of league-wide team revenue. Free agency meant that a player could sell himself to the highest bidder—in other words, the market would operate to establish player salaries.

At first, everyone was shocked. Teams spent outlandish sums on unknown linemen. Players that most fans never heard of made millions. Then a starting left tackle for the Bills, Will Wolford, announced his deal: $7.65 million over three years to play for the Colts. No one had ever paid so much money for a mere lineman. Not only that, his contract stipulated that Wolford was guaranteed to be the highest paid player on offense for as long as he was on the team.

The NFL threatened to invalidate this outrageous contract. In the end, the deal was allowed, but the commissioner decreed that such terms in a contract could never be used again.

Lewis, pp. 227–228 (emphasis added), explains what had happened:

> The curious thing about this market revaluation is that nothing had changed in the game to make the left tackle position more valuable. Lawrence Taylor had been around since 1981. Bill Walsh's passing game had long since swept across the league. Passing attempts per game reached a new peak and remained there. There had been no meaningful change in strategy, or rules, or the threat posed by the defense to quarterbacks' health in ten years. There was no new data to enable NFL front offices to value left tackles—or any offensive linemen—more precisely. *The only thing that happened is that the market was allowed to function.* And the market assigned a radically higher value to the left tackle than had the old pre-market football culture.

Economics students around the world study supply and demand, but they think it is a graph. It is so much more than an X. It is a model that explains how pressures from buyers and sellers are balanced.

This example shows that markets value commodities by reflecting the underlying demand and supply conditions. Blind side tackles are worth a lot of money in the NFL. Before markets were used, they were grossly underpaid. There were no statistics for linemen like yards rushing or field goal percentage so they could not differentiate themselves. The market system, however, expressing the desires of general managers and reflecting the true importance of the blind side tackle, correctly values the position.

Markets are neither moral nor caring. They are a way to consolidate information from disparate sources. Prices are high when everyone wants something or there is very little of it available. For blind side tackles, with both forces at work, the market system was a bonanza.

Supply and Demand and Resource Allocation

This introduction to the market system via partial equilibrium showed how an individual market settles down to its equilibrium solution. Much of this material is familiar because most introductory economics courses emphasize supply and demand analysis.

There are two fundamental concepts, however, that are critical in gaining a deep understanding of supply and demand.

1. Supply and demand curves do not materialize out of thin air. They are the result of comparative statics analyses on consumer and firm optimization problems. In other words, supply and demand must be interpreted as the reduced-form solutions from utility- and profit-maximizing agents. Figure 17.6 drives this point home by adding representative consumer and firm graphs to supply and demand.

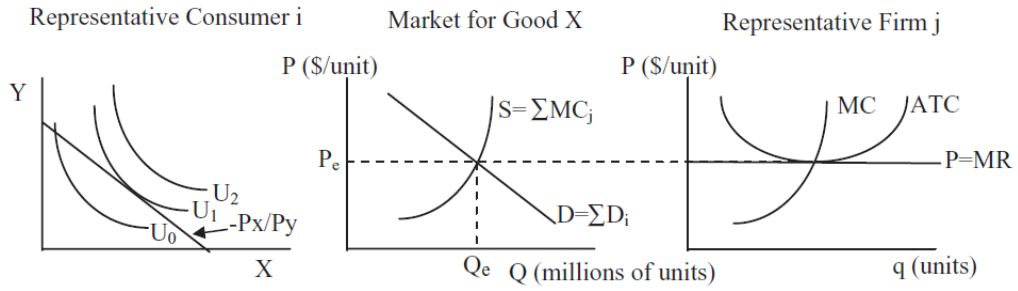


Figure 17.6: An overall view of supply and demand.

The notation in Figure 17.6 is awkward because we are combining consumer and firm theories which have their own individual histories. Thus, X in the left panel is the number of units of the same good that is produced by the firm in the right panel with label "q (units)." Likewise, P in the middle and right panels equals P_x in the left panel. Notational awkwardness notwithstanding, it is true that consumers generate demand for every good and service and the sum of individual demands is market demand. The same holds for supply and firms. Figure 17.6 is a great way to put it all together.

2 Supply and demand is a resource allocation mechanism. It is the equilibrium quantity that is of greatest importance in the supply and demand model because this is the market's answer to society's resource allocation problem. The price is the variable that drives a market to equilibrium, but it is Q_e that represents how much of society's scarce resources are to be allocated to the production of each commodity, according to the market system.

   A picture of this is in the *Intro* sheet. Now that you have finished this section, take another look at it and walk through it carefully.

Introductory economics students are taught supply and demand, but they do not understand that the market demand and supply curves are reduced forms from individual optimization problems. Deriving demand and supply is a bright line separating introductory from intermediate courses.

In addition, introductory courses stress price and equilibration (surpluses and shortages) as students learn the basics of supply and demand. Unfortunately, this means students miss the fundamental point: the equilibrium quantity is the decentralized, market system's answer to how much of society's scarce resources should be devoted to this particular commodity. There are graphs like Figure 17.6 for every good or service allocated by the market.

While the graphics in the *Intro* sheet emphasize the importance of Q_e, Figure 17.7 offers another way to explain what supply and demand is really all about. Filling in the mountain of society's finite resources with a checkerboard pattern conveys that the factors of production are individually owned and controlled. Each square represents the resources controlled by each person. Every person owns a tiny piece of the mountain and decides what do with that labor and capital.

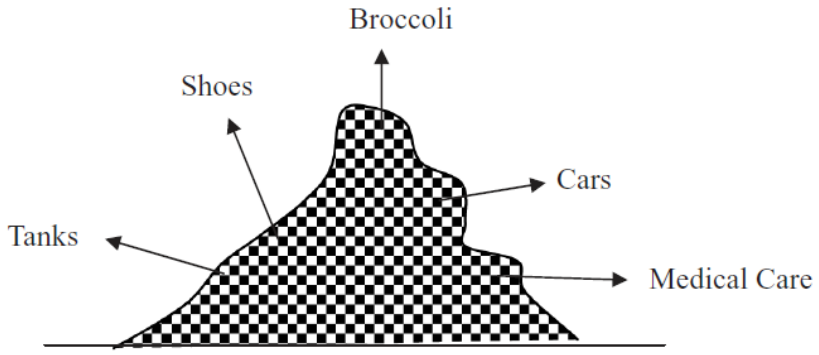


Figure 17.7: Individual ownership of resources.

Every product allocated by the market system has a supply and demand that attracts individual resources owners. Out of this cacophony of interactions, an equilibrium is found and resources flow to the production of an amazing variety of goods and services. This is the truly fascinating aspect of supply and demand. Each agent is self-interested and thinking only of their own gain, but the outcome of the market system establishes a pattern that answers the question of how to use scarce resources.

Of course, the checkerboard pattern in Figure 17.7 makes it seem like everyone controls equal shares, yet there is no question that some people own more resources than others. Inequality in the distribution of resources can be a serious obstacle facing the market system. It will not work well if resources are grossly unequally distributed.

This leads to another common misconception regarding equilibrium and desirability. Can we conclude, by virtue of the fact that the market is in equilibrium, that the market system has correctly solved society's optimization problem? Absolutely not. Equilibrium does not automatically equal optimal. The next section tackles this issue.

Exercises

STEP Click the Reset button in the *EquilibriumSolution* sheet to set the coefficients to their initial values.

1. Use the scroll bar in cell C7 of the *EquilibriumSolution* sheet to set the intercept of the inverse demand curve to 375. Use Excel's Solver to find the equilibrium solution. Take a picture of the answer and paste it in your Word document.

2. Solve the equilibrium model with $d_0 = 375$ via analytical methods. Show your work, using Word's Equation Editor as needed.

3. Because the intercept increased compared with the initial values of the parameters, we know there has been an increase in demand. How has the market responded to this shock? Is the market's response reasonable?

References

The epigraph is from page 111 of Judy Klein, "The Method of Diagrams and the Black Arts of Inductive Economics," published in Ingrid Hahne Rima, *Measurement, Quantification and Economic Analysis: Numeracy in Economics* (1995). As mentioned in section 4.3's References, credit for drawing supply and demand curves is usually given to Jenkin in 1870 and then Marshall in 1890 made the diagrams popular. Klein reviews precursors and how graphs evolved and came to be so important in economics.

The basic questions about resource use that must be answered by society can be traced to Paul Samuelson's *Introductory Economics* textbook (first published in 1948) and Frank Knight's *The Economic Organization* (1933). "Samuelson boiled Knight's five functions down to three: i) What commodities shall be produced and in what quantities?, ii) How shall they be produced?, and iii) For whom are they to be produced? 'These three questions,' Samuelson adds, paraphrasing Knight, 'are fundamental and common to all economies.'" See Ross B. Emmett, "Frank H. Knight and The Economic Organization," Michigan State University Working Paper No. 0405–01, p. 16, papers.ssrn.com/sol3/papers.cfm?abstract_id=922531.

Michael Lewis' book, The Blind Side: The Evolution of a Game was a huge hit in 2006. It was made into a movie in 2009, winning a Best Picture nomination and the Academy Award for Best Actress for Sandra Bullock.

It follows that consumer's surplus is not a concept which can be attributed to Marshall as something rather peculiarly his own. All that belongs exclusively to him is the name.

R. W. Houghton

17.2 Consumers' and Producers' Surplus

Society's resource allocation problem is an especially important optimization problem. It is an easy problem to envision. Pile up of all of society's factors of production and then ask, "How should we use these resources? What should we make? How much of each product should be produced? How should we distribute the output?" These are questions about resource allocation.

An important idea is that of a constraint. Needs and wants by consumers far outstrip available resources. More of one means less of other goods and services.

The previous section showed how supply and demand establishes an equilibrium price and output. The latter is the market system's answer to the resource allocation questions.

Although we are not studying alternative resource allocation methods, it is worth pointing out that if supply and demand is not used, that does not make difficult choices go away. Scarcity means there is not enough to go around. We may decide we do not want to use markets to allocate scarce organs, but we will still need a mechanism to decide whose lives are saved.

This section changes the focus from how supply and demand works to an evaluation of the market system's solution. The approach is clear: We first consider what an optimal allocation would look like, and then check to see whether the market's allocation conforms to the optimal solution.

Finding the Optimal Quantity in a Single Market

To find the optimal solution, we conduct a fanciful analysis. Like the imaginary budget line we used to find income and substitution effects, we work out a thought experiment that actually can never be carried out.

Suppose you had special powers and could allocate resources any way you wanted? Your official title might be *Omniscient, Omnipotent Social Planner*, or OOSP, for short. You are omniscient, or all knowing, so you know every consumer's desires and every firm's costs of production. Because you are omnipotent, or all powerful, you can decide how much to produce of each good and service and how it is produced and distributed.

Because this is partial equilibrium analysis, we focus on just one good or service. The question for you, OOSP, is, "How much should be produced of this particular commodity?"

One way for you to answer this question is to measure the total gain obtained by the consumers and producers of the good. When we compute the gain, we subtract the cost of acquiring the product for consumers and, for firms, the costs of production. The plan is to compute the total net gain for different quantities and pick that quantity at which the total gain is maximized.

The notion of net gain, something above the cost that is captured by consumers and firms, is the fundamental idea behind consumers' and producers' surplus. *Consumers' surplus* is the gain from consumption after accounting for the costs of purchasing the product. *Producer's surplus* is the difference between total revenues and total variable costs. In the long run, it is profit.

We begin with producers' surplus because it is uncontroversial. We will see that consumers' surplus is problematic.

Producers' Surplus

At any given price, if sellers get that price for all of the units sold, they get a surplus from the sale of each unit except the last one. The sum of these surpluses is the producer's surplus. The sum of all of the producer's surpluses in the market is the producers' surplus, *PS*.

The location of the apostrophe matters. Producer's surplus is the surplus obtained by one firm. If the focus is on all of the firms, we use producers' surplus.

STEP Open the Excel workbook *CSPS.xls*, read the *Intro* sheet, then go to the *PS* sheet.

The sheet displays an inverse supply curve given by $P = 35 + 0.52Q_s$. The area of the green triangle is *PS*. To see why, consider the situation when output is 75 units and the price is \$74/unit.

The very last unit sold added \$74 to total cost (given that we know that the supply curve is the marginal cost curve). Thus, the 75^{th} unit sold yielded no surplus. In general, the marginal unit yields no surplus.

But what about the other units? All of the other units are *inframarginal* units. In other words, these are units below the marginal (last) unit and, in general, the inframarginal units generate surplus. The firm is receiving a price in excess of marginal cost for these units, from 1 to 74, and, therefore, it is reaping a surplus each of those units. We can add them up to get producer's surplus.

Consider the 50^{th} unit. The marginal cost of the 50^{th} unit is given by $35 + 0.52 * 50 = \$61$. The firm would have been willing to sell the 50^{th} unit for \$61, but it was paid \$74 for that 50^{th} unit. So, it made \$13 on the 50^{th} unit.

STEP Look at cell Q68. It reports the surplus generated by the 50^{th} unit, \$13, as we computed above. Look at cell Q28. It reports the surplus generated by the 50^{th} unit, \$33.80.

Cell R19 adds the surpluses from all of the inframarginal units. Notice how PS steadily falls from the first to the last unit. The key to PS is that all quantities are sold at the same price, but marginal cost starts low and rises. The firm makes a surplus above *MC* on all output except the last one.

Cell R19 differs from cells B19 and B21 because cell R19 is based on an integer interpretation of output. If output is continuous, then we can compute the *PS* as the area of the triangle created by the horizontal price and the supply curve.

Notice that cell B19 offers another way to understand *PS*. If supply is marginal cost, then the area *under* the marginal cost curve is total variable cost. Because marginal cost is linear, the computation is easy. If *MC* was a curve, we would have to integrate. Total revenue is simply price times quantity. Cell B19 computes $TR - TVC$, the excess over variable cost, which is the producers' surplus.

STEP If $Q_s = 95$, what is *PS*? Use the scroll bar in cell C12 to set quantity equal to 95.

At 95 units of units of output, *MC* is \$84.40. At that price, the 95$^{\text{th}}$ unit has no surplus. But all of the other, inframarginal units generate surplus, adding up to \$2,346.50.

STEP Explore other quantities and confirm that as output rises, so does producers' surplus.

Consumers' Surplus

The idea is the same. At any given price, if a buyer pays that price for all of the units bought, she gets a surplus from the purchase of each unit except the last one. The sum of these surpluses is the consumer's surplus. The sum of all of the consumer's surpluses is the consumers' surplus, *CS*.

STEP Proceed to the *CS* sheet.

Given the inverse demand curve, $P = 350 - 0.2Q_d$, we can easily compute *CS* for a given quantity as the area of the pink triangle.

At $Q_d = 95$, the price so consumers will buy 95 units is \$160/unit. The last unit purchased provides no surplus, but the inframarginal units generate *CS*. The area under the demand curve, but above the price, is a measure of the net satisfaction enjoyed by consumers.

Consumers' surplus comes from the fact that consumers would have paid more for each inframarginal unit than the price they actually paid so they get a surplus for each marginal unit.

STEP Use the quantity scroll bar to confirm that as output rises, so does consumers' surplus.

As mentioned earlier, there is a problem with consumers' surplus. We will finish how OOSP could use *CS* and *PS* before explaining the problem.

Maximizing *CS* and *PS*

Producers' surplus is the amount by which the total revenue exceeds variable costs and measures gain for the firm. Consumers' surplus also measures gain because it is the amount by which the total satisfaction provided by the commodity exceeds the total costs of purchasing the commodity.

Both parties, consumers and producers, gain from trade. This is why a trade is made—both buyer and seller are better off. When you buy something, you part with some money in exchange for the good or service. If the purchase is voluntary, you must value what you are getting more than what you paid for it or else you would not have bought it. Similarly, the seller values the money you pay more than the good or service or else she would refuse to sell at that price. The gains from voluntary trade are captured in the terms consumers' and producers' surplus.

Casting the problem in terms of surplus received by buyers and sellers leads naturally to this question: What is the level of output that maximizes the total surplus? After all, it is clear that as quantity changes the *CS* and *PS* also change.

Thus, OOSP is faced with the following optimization problem:

$$\max_q CS(q) + PS(q)$$

The idea is to maximize the gains from trade for all buyers and sellers. This problem can be solved analytically and numerically. We focus on the latter.

STEP Proceed to the *CSandPS* sheet.

This sheet combines the surpluses enjoyed by producers and consumers into a single chart, shown in Figure 17.8.

Understanding this chart is fundamental. We proceed slowly. The vertical dashed line represents the quantity, which OOSP controls and will choose so that $CS + PS$ is maximized.

There are two prices on the chart, one for the firm and the other for the consumer. The idea is that OOSP uses the quantity to determine the prices needed for firms to be willing to produce the output level and for consumers to want to buy that amount of output.

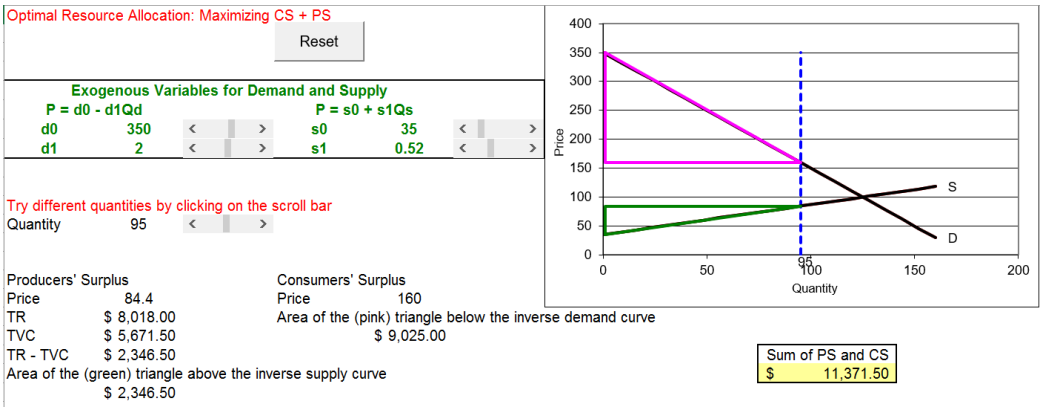


Figure 17.8: *CS* and *PS* at Q=95.
Source: CSPS.xls!CSandPS.

This is *not* an equilibrium model of supply and demand. OOSP cares only about choosing the optimal output. Price for consumers and firms is used only to compute surplus.

In Figure 17.8 (and on your computer screen), producers receive a price of $84.40 for each of the 95 units, yet consumers pay $160.00 per unit. Remember that OOSP, our benevolent dictator, has magical powers so she can charge one price to consumers and give a different price to producers. By adding the values in cells E18 and B21, we get the value in cell J20. It is highlighted in yellow and maximizing it is the goal.

STEP Click on the slider control (over cell C12), to increase output in increments of five units.

As output increases, *CS* and *PS* both rise.

STEP Continue clicking on the slider control so that output rises above 125 units.

Now the sum of *CS* and *PS* is falling. That is confusing because the two triangles are getting bigger. But once the price to consumers falls below the price to the firms, we have to pay the difference. This is explained below in more detail. For now, let's work finding optimal Q.

STEP Launch Solver and use it to find Q^*.

With an empty Solver dialog box, you have to provide the objective (J20) and changing cell (B12). We find that $CS + PS$ is maximized at $Q^* = 125$ units.

In other words, OOSP should order the manufacture of 125 units of this product, allocating the inputs needed from society's scarce resource endowment. This level of output maximizes the sum of *CS* and *PS*.

We have seen this number before. In the previous section, we found that the equilibrium solution was $Q_e = 125$ units. This means that the market's solution is the optimal solution. This is a remarkable result.

No one intended this. No one chose this. No one directed this. Supply and demand established an equilibrium output which answered the question of how much to produce and we now see that it is the same solution we would have chosen if our goal was to maximize consumers' and producer's surplus. This is truly amazing.

Deadweight Loss

If OOSP chooses an output level below 125 and charges a price to consumers based on the inverse demand curve and pays producers a price based on the inverse supply curve, it will generate a smaller value of $CS + PS$.

How much smaller? The amount of surplus not captured is given by the trapezoid between the consumers' and producers' surpluses. This area is called *deadweight loss*, *DWL*. It is a fundamental concept in economics and merits careful attention.

STEP Enter 95 in cell B12, then click the Show DWL button.

Not only do data appear below the button, but the chart has been modified to include a red trapezoid. The area of the trapezoid is displayed in cell D30.

STEP Click on cell D26.

The formula is simply the solution of the intersection of the supply and demand curves. We know this quantity is the solution to the problem of maximizing *CS* and *PS*.

STEP Click on cell D28.

This seemingly complicated formula is not really that hard. It displays the maximum possible total surplus. Two things are being added, *CS* and *PS*. The first part of the formula is *PS*: 0.5*(((s0_ + s1_ *D26) − s0_)*D26). It is half the height of the *PS* triangle times the length (or quantity produced). The second part of the formula uses the same area of the triangle formula to compute the *CS*: 0.5*((d0_ - (d0_ − d1_*D26))*D26).

STEP Click on cell D30.

The formula, = D28 − J20, makes crystal clear that deadweight loss is maximum total surplus minus the sum of *CS* and *PS* at any value of output. In other words, deadweight loss is a measure of the inefficiency of producing the wrong level of output in a particular market. Deadweight loss vaporizes surplus so that it disappears into thin air. Deadweight loss is pure waste.

STEP Click on the slider control (over cell C12) to increase output in increments of five units.

As you increase output, note that the deadweight loss falls as the output approaches the optimal quantity. There is no deadweight loss when the output is at 125 because this is the optimal level of output.

Another way to expressing the efficiency in the allocation of resources of the equilibrium solution is to say it has no deadweight loss. That is, no inefficiency in allocating resources.

As Q approaches Q^* we reach the maximum possible $CS + PS$ and DWL goes to zero. As Q keeps rising, past $Q > Q^*$, we get less total $CS + PS$ and deadweight loss rises. We get deadweight loss on either side of Q^*. The explanation for deadweight loss when $Q > Q^*$ is more complicated. Let's look at some concrete numbers.

STEP Set output above the optimal level, for example, $Q = 150$.

Your screen should look like Figure 17.9. It is true that *CS* and *PS* triangles are large, but with a higher price to firms than consumers, society has to pay for the difference. Once we account for this, the total gain is less than that at $Q = 125$ and we suffer deadweight loss, as shown by the red triangle.

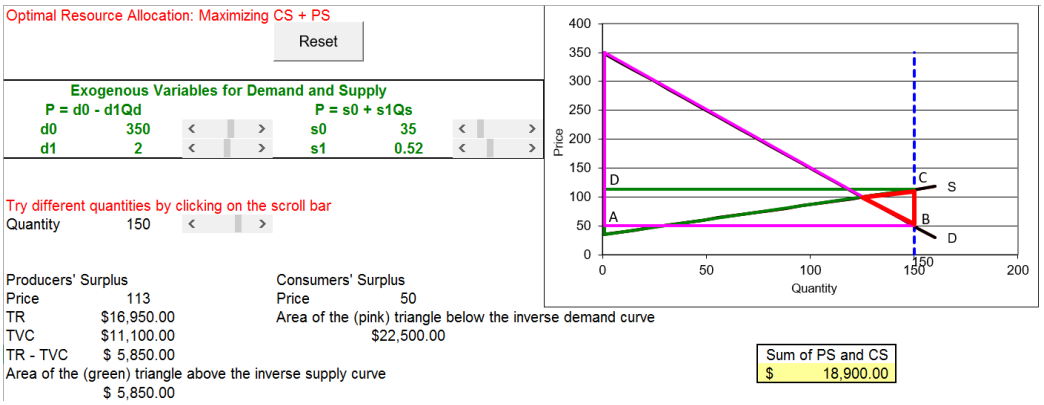


Figure 17.9: *DWL* at Q=150.
Source: CSPS.xls!CSandPS.

Figure 17.9 shows that it is possible to have sellers receive \$113 per unit sold yet have buyers pay only \$50 per unit sold, but someone is going to have to make up that \$63 per unit difference. The total value of the subsidy, \$63/unit times 150 units is \$9,340. This amount (rectangle ABCD in Figure 17.9) must be subtracted from the sum of *CS* and *PS*.

When we add everything up, we get a total surplus of \$18,900 at $Q = 150$, which is lower than the maximum total surplus. Cell J20 uses an IF statement to get the calculation right. The deadweight loss from producing 150 units is \$787.50 (cell D30).

The deadweight loss at $Q = 150$ is given by the area of the red triangle in Figure 17.9. The geometry is easy. We must subtract a rectangle with height 63 and length 150 from the sum of the pink *CS* and green *PS* triangles. This leaves the red triangle as the *DWL* caused by producing too much output.

There is one optimal output and at that value, deadweight loss is zero. Outputs above and below Q^* produce inefficiency in the allocation of resources because we fail to maximize $CS + PS$. This is called deadweight loss.

Price Controls

Price controls are legally mandated limits on prices. A price ceiling sets the highest price at which the good can be legally sold. A price floor does the opposite: The good cannot be sold any lower than the given amount.

To be effective, a price ceiling has to be set below and a price floor has to be set above the equilibrium price.

Most introductory economics students are taught that price ceilings generate shortages and price floors lead to surpluses. For most students, the take-home message is that market forces cannot push the price above the ceiling or below the floor so the market cannot clear and this is why price controls are undesirable.

It turns out that this is not exactly right. Although it is true that ceilings lead to persistent excess demand and floors prevent the market from eliminating excess supply, the real reason behind the unpopularity (among economists) of price controls is the fact that they cause a misallocation of resources.

STEP Proceed to the *PriceCeiling* sheet.

Suppose there is a price ceiling on this good at \$84.40. At this price, there is a shortage of the good because quantity demanded at \$84.40 is 132.8 units (cell B13) while quantity supplied is only 95 (cell B12).

The price cannot be bid up because \$84.40 is the highest price at which the good can be legally sold. Thus, with this price ceiling, the output level is 95. We know this is an inefficient result because we know $Q^* = 125$. This is the real reason why this price ceiling is a poor policy, not because it causes a shortage. The price ceiling fails to maximize total surplus.

To be clear, with this price ceiling, too few resources are allocated to the production of this good or service. There will be only 95 units of it produced, not the optimal 125 units. The fact that there is a shortage is true, but it is the misallocation of resources that is the problem.

While the misallocation of resources is easy to see since the quantity is wrong, deadweight loss is more complicated. It depends on the story about the price control and how agents react.

Suppose, for example, that market players are all honest so there is no illegal selling of the good above the maximum price. In other words, producers do not violate the law. Suppose further that the good is allocated via lottery so there are no lines of buyers or resources spent waiting. This means that consumers' surplus is now a trapezoid instead of a triangle.

STEP Click the Show DWL button.

As shown in Figure 17.10 (and on your screen), a rectangle has been removed from deadweight loss so it is now just the red triangle.

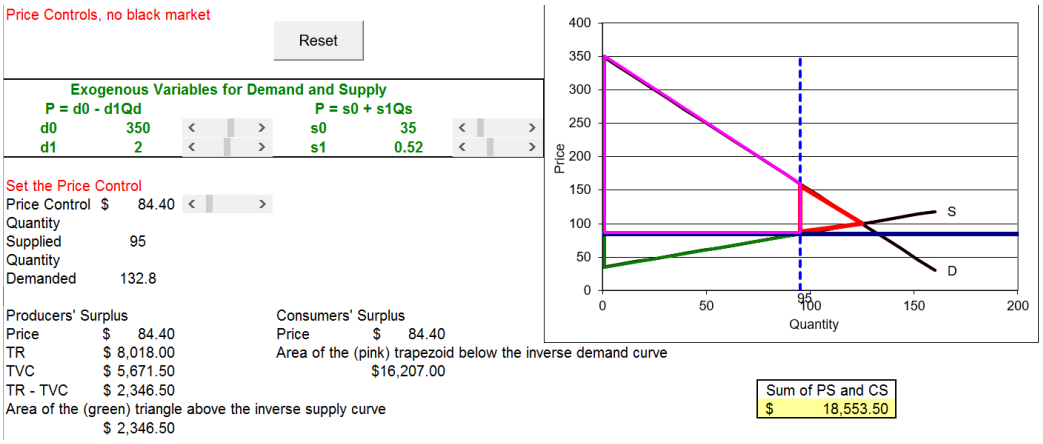


Figure 17.10: *DWL* with no illegal market.
Source: CSPS.xls!PriceCeiling.

In addition to the usual *CS* triangle in Figure 17.10, consumers enjoy the area of the rectangle computed by multiplying a price of \$160 (which is the price consumers are willing to pay for 95 units of the good) minus \$84.40 (the price consumers actually pay) times 95 units.

The good news behind this price ceiling with no cheating story is that the deadweight loss is much smaller than in the *CSandPS* sheet with $Q = 95$ because the lucky consumers who can purchase the good do not have to pay \$160/unit. The bad news is that there is still a deadweight loss of \$1,134. This is a measure of the inefficiency of the price ceiling with no illegal market.

Suppose instead that there are unlawful sales of the product at the illegal market price, \$160/unit (this is the most buyers are willing to pay for 95 units). Suppose, in addition, that somehow there are no wasted resources associated with this illegal market. No police investigations, court cases, or any other resources are spent on stopping criminal sales. Then the producers get the rectangle. With this idealized illegal market, the rectangle is transferred from consumers to producers, but the deadweight loss stays the same. The Q&A sheet asks you to demonstrate this.

If, as is almost surely true, illegal selling results in more resources being spent, then the deadweight loss is larger than the red triangle. Illegal activity often leads to violence (think of illegal drugs, which are a market with a price ceiling of zero) and we would subtract that from $CS + PC$ and thereby increase DWL.

Consider two other stories about the price ceiling. A limited set of buyers are given coupons to buy the product. To buy the good (at the legal price), you must have a coupon. If a rationing coupon scheme is used, the sellers of the coupons get the rectangle. The deadweight loss remains the same.

Suppose, finally, that a price ceiling is set and the good is allocated on a first-come-first-serve basis. In other words, buyers have to wait in line. With this story, the time, effort, and other resources buyers waste standing in line (or paying others to stand in line for them) must be subtracted from the total surplus. The deadweight loss rises. If the entire rectangle is lost, then the deadweight loss is the same as that in the *CSandPS* sheet when 95 units of output are produced.

Price controls are a popular way to modify market results. Unfortunately, from a resource allocation standpoint, price controls suffer from the fact that they fail to maximize total surplus. It is this property and not that they produce shortages that earn price ceilings criticism. We want the allocation mechanism to give optimal Q.

It is confusing that correctly measuring deadweight loss depends on the story, but do not be distracted by the many ways buyers and sellers can respond to price controls. The take-home message is that any deviation from Q^* means that the allocation scheme has failed. Deadweight loss, which gives a measure of the inefficiency in monetary units, depends on the specific implementation of the price control, but the fact that it is not zero is evidence that it has failed.

Caveat Emptor

"Let the buyer beware" is the meaning of the Latin phrase, *caveat emptor*. This idea from contract law is a warning to the buyer that they are responsible for what they are buying. The consumer needs to be careful so they aren't tricked or end up with a poor quality, unsuitable product.

Caveat emptor applies to deadweight loss. On the one hand, deadweight loss is a common way that economists measure inefficiency. It is based on the idea that the maximum total surplus is not attained from a particular output level. But users need to know what they are getting themselves into—deadweight loss has two glaring weaknesses.

The first has to do with our calculation of consumers' surplus. For technical reasons, restrictive assumptions about the utility function must be imposed. For example, a Cobb-Douglas utility function for individual consumers will not work because it has an income effect. A quasilinear utility function will work (no income effect), but it is unlikely that all consumers have quasilinear utility.

Consumers' surplus violates the rule that we should not make interpersonal utility comparisons. We are using the demand curve to add up dollar measures of the extra satisfaction that different people get from consuming a product. That is unsound and breaks a basic tenet of modern utility theory.

The second weakness stems from the use of partial equilibrium analysis. We are calculating deadweight loss based on the impact in a single market of a deviation in output from its optimal level. The focus on one market is too limited. If we apply too many or too few resources to the production of one good, we will cause deviations from optimal output for other goods and services. So, the deadweight loss computation based on one market is a lower bound. To get it exactly right, we would have to analyze effects on other markets and do a general equilibrium analysis.

Regarding deadweight loss, it is *caveat emptor*. Remember that deadweight loss measures inefficiency and it is commonly used in applied work, but it is not exactly right. The best way to think of deadweight loss is as an approximation.

Some economists are appalled at the thought of using deadweight loss. These most strident critics are usually more theoretically oriented. Economists who do empirical work are more likely to argue that deadweight loss is imperfect, but practically speaking, it is a useful way of measuring inefficiency.

Optimal Allocation of Resources

This is an important section. It introduced producers' and consumers' surpluses, which are key elements in the omnipotent, omniscient social planner's objective function.

The idea that there is an optimal level of output for each good and service is fundamental. From this idea we get the procedure for evaluating any allocation scheme or government policy: We compare an observed result to the optimal answer.

It is obvious that quantities below the intersection of supply and demand cannot be optimal because both CS and PS rise as Q increases. The situation with quantity above the intersection of supply and demand is more subtle. To get the calculation right, whenever quantity is above the intersection point, we must subtract from the sum of CS and PS a rectangle that is the difference between prices multiplied by quantity.

The most important and remarkable result from this section is that $Qe = Q^*$. This says that in a properly functioning market, the equilibrium quantity (which is the market system's answer to society's resource allocation problem) yields the socially optimal level of output.

Price controls lead to inefficient allocation of resources. The output generated does not match the optimal output. The deadweight loss associated with a price control depends on the story of how the particular implementation of the price control is enforced and responded to by buyers and sellers.

There is no question that deadweight loss is a linchpin of policy analysis. Countless estimates of deadweight loss and cost–benefit studies have been conducted. It is, however, flawed. Measuring consumers' surplus in value of money terms from a market demand curve in a partial equilibrium setting leaves us on very thin ice. Applications and estimates of deadweight loss should be seen as an approximation to the exact measure of the loss from the misallocation of resources (if such a measure exists).

While deadweight loss is flawed, the notion of a misallocation of resources is not. The idea that there is an optimal solution to society's resource allocation problem is perfectly valid. So is defining an allocation that deviates from optimal as a misallocation of resources. These are bedrock ideas in microeconomic theory.

This should mark the end of this section, but because there is so much confusion about equilibrium and optimal resource allocation, what follows is an attempt to provide some clarity.

Take a moment, before you begin reading the next section, to think about what supply and demand is really all about. What is the point? What are we trying to explain? Read the next section, maybe even repeatedly, to make sure you can answer these fundamental questions.

Equilibrium and Optimal Resource Allocation

The material below is being repeated for emphasis. The Theory of Consumer Behavior and Theory of the Firm are stepping stones to the $Q_e = Q^*$ result. Let's put things in perspective and explain why this is so fundamental.

In the 1700s, it is absolutely true that philosophers and deep thinkers of the day were baffled by the market system. There was active debate about how and why Europe and, especially, England was getting so rich. How could the unplanned, individual decisions of many buyers and sellers produce a pattern, much less a good result? It seemed obvious that a leaderless, fragmented system would produce chaos.

In the previous section, we saw that the equilibrium quantity, Q_e, generated by a properly functioning market is located at the intersection of supply and demand. The market uses a good's price to send signals to buyers and sellers. Prices above equilibrium are pushed down, whereas prices below equilibrium are pushed up. At the equilibrium solution, the price has no tendency to change and output is also at rest. The equilibrium level of output is the market's answer to how much of society's resources will be devoted to producing this particular good.

Our work in this section on consumers' and producers' surplus takes a much different perspective on the resource allocation problem. Instead of examining how the market works, we have created a thought experiment, giving an imaginary social planner incredible powers. Given the goal of maximizing total surplus, OOSP would choose an optimal quantity, Q^*, that should be produced. If we produce less or more than this socially optimal amount, society would forego surpluses that would make producers and consumers better off. Producing the wrong Q yields deadweight loss.

If we compare the market's equilibrium quantity to the socially optimal quantity, we are struck by an amazing result: $Q_e = Q^*$. This critical equivalence means that we do not need a dictator, benevolent or otherwise, to optimally allocate resources. The market, using prices, can settle down to a position of rest where all gains from trade are completely exploited and the sum of producers' and consumers' surplus is maximized.

There is no guarantee, however, that $Q_e = Q^*$—there are conditions under which the invisible hand does not lead the market to optimality. We will see examples where the equality does not hold and the market is said to fail.

As you work on this section and this part of the book, do not lose sight of the main point: The market's ability to generate an equilibrium quantity that is socially optimal is nothing short of amazing and unbelievable. It is equivalent to geese flying in a V. A pattern is generated by the interactions of individuals with no awareness or intent to make the pattern.

Consider this hypothetical: we learn that broccoli cures cancer. Would we need a president, prime minister, or king to tell farmers to grow more broccoli? Of course not. Broccoli would fly of the shelves, its price would rocket, and farmers would *automatically* plant and produce more broccoli. They would not try to figure out what would be best for society, but simply respond to the market signal. That is what the supply and demand model is really all about.

Analogies from biology are many, but this one might be so shocking and different from anything you have seen before that it will convey why supply and demand is so fascinating to economists.

STEP Visit http://tiny.cc/siphonophore to learn about this creature and see it in action.

Exercises

1. From the *CSandPS* sheet, click the Reset button, then set $d_0 = 375$ and use Solver to find the optimal quantity. Take a picture of the cells that contain your answer and paste it in a Word doc.

2. Click the Show DWL button. Suppose there was a price ceiling of \$84.40. What is the story about price ceilings assumed by the chart and *DWL* computations on the sheet?

3. Suppose the government implemented a price support scheme (this is a type of price floor that is used frequently for agricultural products) where they only allowed 95 units to be produced. Cell E16 shows that the market price would be \$185. Compute the deadweight loss and explain it.

References

The epigraph is from the first page of R. W. Houghton, "A Note on the Early History of Consumer's Surplus," *Economica*, New Series, Vol. 25, No. 97 (February, 1958), pp. 49–57, www.jstor.org/stable/2550693 A French engineer, Jules Dupuit (pronounced doo-pwee) presented the idea of *utilite relative* in 1844, but Alfred Marshall independently rediscovered and popularized the notion of consumer's surplus.

Almost immediately after Marshall introduced consumers' surplus, the concept came under attack. It has survived the move from a cardinal to an ordinal perspective on utility and a variety of other criticisms. Economists know that *CS* is built on shaky foundations, but they often use it in practical, policy-oriented, real-world discussions. In a review of the state of *CS*, Abram Bergson concludes, "Despite theoretic criticism, practitioners have continued to apply consumer's surplus analysis through the years. As some have argued, that must already say something about the usefulness (as well as the use) of such analysis, but just what it says has remained more or less in doubt." See "A Note on Consumer's Surplus," *Journal of Economic Literature*, Vol. 13, No. 1 (March, 1975), pp. 38–44, www.jstor.org/stable/2722212

> Harberger triangles, now common fare, were once rare delicacies. ...While the theory of deadweight loss measurement was well-established by the 1950s, economists very rarely estimated deadweight losses prior to the appearance of Harberger's work.
>
> James R. Hines, Jr.

17.3 Tax Incidence and Deadweight Loss

Many goods and services are taxed. Sales taxes (also called value added or *ad valorem* taxes) are a percentage of the monetary amount spent; quantity taxes are levied per unit bought. Quantity taxes are applied, for example, to gasoline, alcohol, and cigarettes.

In chapter 3.4, we examined cigarette taxes. It was shown that, for a particular consumer, lump sum (fixed amount) taxes are better than quantity taxes. In this section, we turn from an analysis of taxes on the individual to their effect on society and the resource allocation problem.

We will use supply and demand in a partial equilibrium setting to evaluate the effects of taxes on goods and services allocated by the market. We work with quantity taxes because our linear supply and demand curves will shift vertically as the tax is applied. Sales taxes are harder to analyze, but the qualitative results we derive for quantity taxes carry over to sales taxes.

There are two basic issues:

1. *Tax incidence*: determining the tax split between buyer and seller.
2. *Deadweight loss*: evaluating the inefficiency generated by the tax.

Our work will show a counterintuitive proposition: It does not matter whether consumers or producers pay the tax. In the end, neither the tax burden nor the deadweight loss depends on who sends tax payments to the government.

Our approach to the second—and more important—issue relies on comparing the output after the tax is imposed to the socially optimal output (based on maximizing consumers' and producers' surplus). Deviations from optimality are said to be inefficient solutions to society's resource allocation problem. We will use deadweight loss to measure the inefficiency. This is known as *welfare analysis*, where welfare means the well-being of a person or group.

It Does Not Matter Who Sends the Tax Payment

Suppose you are renting an apartment for $700 a month. Suppose further that property taxes rise $100. If your landlord raises the rent to $800 a month and you agree, it is easy to see that you are paying for the entire tax increase. The landlord pays the property tax to the government, but you are bearing the burden of the tax.

But what if you refuse to pay the $100 increase and move out. The landlord cannot find anyone to rent the apartment for $800 and, eventually, agrees to rent the apartment for $725 a month to a new tenant. The computation of the tax burden is easy. The new tenant is bearing the burden of $25 or 25% of the tax increase, while the landlord's burden is $75 or 75%.

No matter what the rent ends up being, the landlord sends the tax payment to the government, but that does not answer the question of who is really responsible for the tax. The landlord may be able to shift some of the tax onto the renter.

It turns out that the elasticities of demand and supply determine who bears the burden. The more inelastic, or price insensitive, the higher the burden.

Tax incidence is the analysis of who bears the burden of a tax. In a moment, we will be working with complicated supply and demand graphs, but the analysis is basically the same as the story of the tenant and the landlord.

Supplier Pays

For most products, the supplier or firm is responsible for collecting the tax when the good is purchased and for sending in the tax payments to the government. This is what is meant by "supplier pays." Of course, we know that who collects and pays the tax is different from the tax incidence because anywhere from 0 to 100% of the tax may be shifted to the consumer.

The elasticities of supply and demand determine how the tax is split between consumer and firm.

STEP Open the Excel workbook *Taxes.xls*, read the *Intro* sheet, then go to the *SupplierPays* sheet.

The sheet has parameters for linear demand and supply curves. Initially, there is no tax so the equilibrium price is \$100/unit and the equilibrium quantity is 125 units. Cell B17 shows that the government collects no revenue and cell E17 shows that there is no deadweight loss (because the market's equilibrium quantity equals the socially optimal quantity).

The price elasticities at the *initial* equilibrium solution are $\epsilon_D = -0.4$ and $\epsilon_S = 1.54$, for demand and supply. The sum of the absolute values is 1.94.

STEP Click on the scroll bar next to cell B14 five times to impose a tax.

A red line appears on the chart and it shifts with each click. Five clicks will set the tax at \$50 and the spreadsheet will look like Figure 17.11.

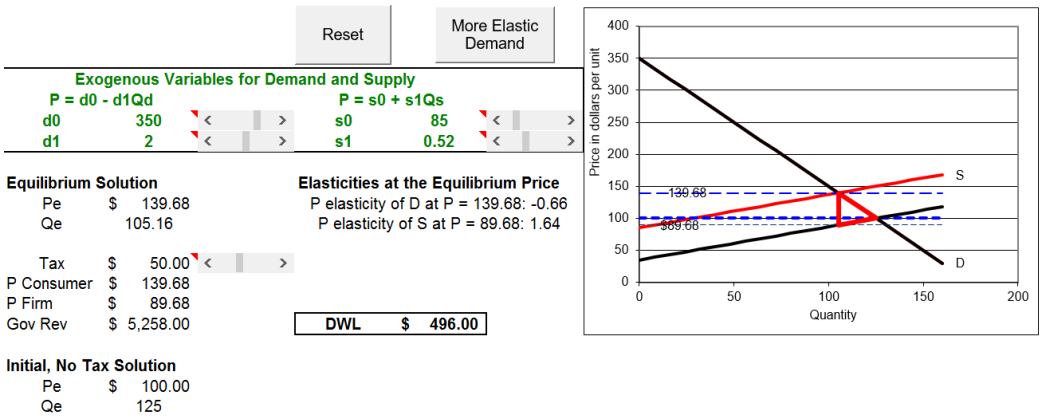


Figure 17.11: Supplier pays a \$50 quantity tax.
Source: Taxes.xls!SupplierPays.

The inverse supply curve has shifted up by \$50/unit because in order for the suppliers to offer a given quantity, they have to receive \$50/unit more than the original supply curve (without the tax). They will not get to keep the extra \$50 per unit—they have to send it to the government.

For example, to offer 125 units at the initial equilibrium solution, firms needed a price of \$100, but now they will need \$150/unit. The value of P is \$150 for $Q = 125$ with the red line in Figure 17.11. Every quantity has the same \$50 increase in price on the red line.

The spreadsheet displays the information we need to compute the tax incidence. We can see that the consumer is bearing the majority of the tax by

looking at the new equilibrium price. The dashed line (and cell B15) shows the new $P_e = 139.68$. We can compute the fraction of the tax borne by the consumer: $\frac{39.68}{50} \approx 79.4\%$. The supplier has managed to pass along all but about one-fifth of the tax to the consumer.

We can also use the absolute values of the *pre-tax* (initial) price elasticities to get the relative burdens for consumer and firm:

$$1 - \frac{0.4}{1.94} \approx 79.4\% \text{ and } 1 - \frac{1.54}{1.94} \approx 20.6\%$$

The *Tax Incidence Formula* to determine the share of the tax burden using demand and supply price elasticities is:

$$1 - \frac{\epsilon_i}{\epsilon_D + \epsilon_S} \text{ for } i = D, S$$

Notice that the formula drops the minus sign for the price elasticity of demand and for the rest of this section, we will mean the absolute value when we refer to the price elasticity of demand.

The elasticity values from the spreadsheet and the *Tax Incidence Formula* make clear that the lower the price elasticity, the higher the tax incidence. As $\epsilon_i \to 0$ (for either D or S), the burden (for D or S) goes to 100%. The consumer is paying four-fifths of tax in Figure 17.11 because demand is much more inelastic than supply at the initial equilibrium price.

We will discuss tax incidence in more detail below, but we turn now to the second, more important issue, the welfare implications of per unit taxes.

With a \$50 quantity tax, the *SupplierPays* sheet shows a deadweight loss of \$496 in cell E17. The deadweight loss can be calculated by finding the difference of the maximum possible surplus minus the surpluses enjoyed by the consumers, producers, and government. This is equivalent to the (red) triangle on the chart, which is also known as a *Harberger triangle*.

We proceed carefully. Consumers' surplus (CS) and producers' surplus (PS) after the tax is imposed have both been reduced by the trapezoidal shapes in Figure 17.12. Clearly, CS has fallen by much more than PS. More importantly, however, is the fact that the deadweight loss (DWL) is not the sum of lost CS and PS because we have introduced a third player—the government. They will get most of CS and PS lost in the form of tax revenue.

The total tax payments of \$5,258 is the area of the rectangle with height \$139.68 − \$89.68 = \$50 and length 105.16 units of output.

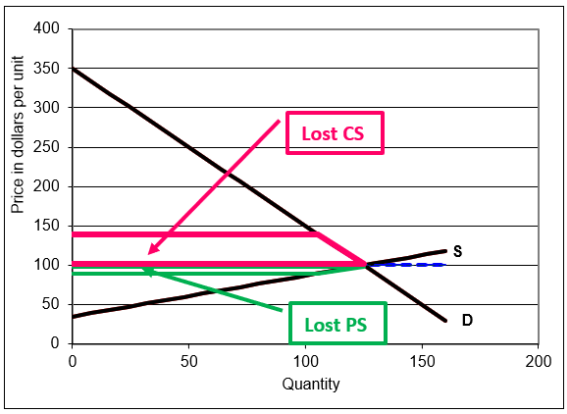


Figure 17.12: Lost *CS* and *PS* from the \$50/unit tax.
Source: Taxes.xls!SupplierPays!AN.

Once we recognize that the tax has lowered *CS* and *PS*, but that part of the surplus is captured by the government, we can see that the deadweight loss is the Harberger (red) triangle in Figure 17.13, with area displayed in cell E17. The surplus in the Harberger triangle vaporizes into thin air, captured by no one.

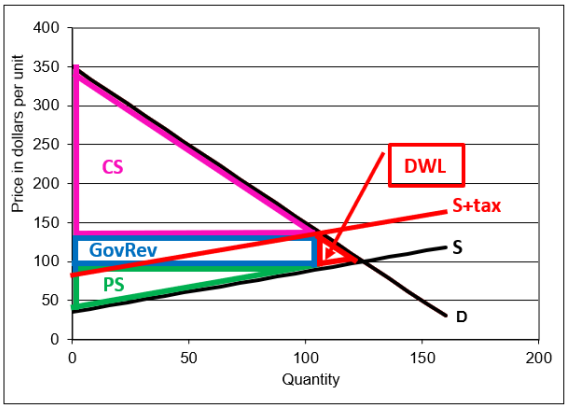


Figure 17.13: *CS*, *PS*, and *GovRev* from the \$50/unit tax.
Source: Taxes.xls!SupplierPays!AN.

The height of the Harberger triangle is the price the consumer pays minus the price received by the firm, which is called the *tax wedge*. This distance is the amount of the tax. When you clicked five times to impose the tax, you

could see the wedge expanding, creating a space between what the consumer pays and the firm receives.

The tax wedge takes surplus from consumers and producers, but this is not a problem. Presumably, the government is building schools, roads, and providing services. As long as someone gets the surplus, partial equilibrium surplus analysis counts it as a successful outcome.

Figure 17.13 shows, however, that the Harberger triangle goes to no one. This is a problem. Deadweight loss is surplus that simply vanishes. It is a loss of surplus that is not recouped by anyone.

The width of the *DWL* triangle is the distance from the new equilibrium quantity after the tax to the original equilibrium quantity. The bigger this distance, the greater is the distortion of the tax in terms of resource allocation.

STEP Click on cell E17 to see its formula. It simply computes the area of the red Harberger triangle.

We summarize and repeat a few key ideas. Deadweight loss is a dollar measure of the distortion caused by the tax—the "market with a tax" scheme is no longer producing the optimal quantity. This is a misallocation of resources. Deadweight loss represents gains from trades that are not being exploited. There is $496 in value that no one is getting. It is simply vaporized and disappears into thin air.

The rectangle formed by the tax times the equilibrium quantity (after the tax is imposed) is a transfer from consumers and producers to the government. This does not count as deadweight loss because someone (the government) is getting it. The key to understanding deadweight loss is that it accrues to no one—it is unclaimed surplus and, therefore, pure waste.

Demander Pays

Suppose that instead of the firm it is the consumer who is responsible for collecting the quantity tax when the good is purchased and for sending in the tax payments to the government. This may seem a little strange at first, but there are cases where this occurs.

For example, if you buy online and the seller does not charge you state and local taxes, you should pay those taxes. At the dawn of the internet, this gave online retailers a big advantage over brick and mortar stores that added sales and other taxes to the total. Few people pay taxes when they are not collected by the seller. Today, almost all online retailers include taxes.

For the purposes of comparing what happens when the buyer or seller pays the tax, forget about administrative costs or the fact that firms are much better tax collectors than consumers. We assume that consumers and firms will both comply and send the correct tax payment to the government even though that is obviously not true.

STEP Go to the *DemanderPays* sheet and impose a $50 tax. Pay attention to the screen as you click. You can watch the tax wedge emerge.

Figure 17.14 shows the result, with the *DWL* triangle displayed.

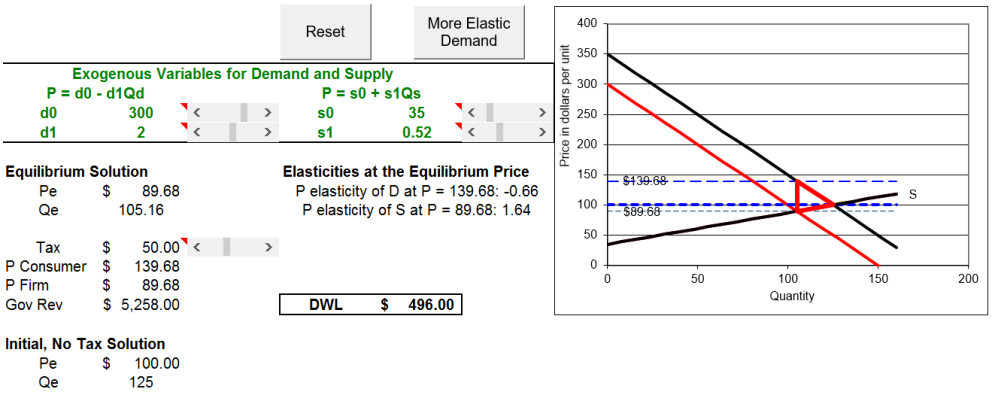


Figure 17.14: Demander pays a $50 quantity tax.
Source: Taxes.xls!DemanderPays.

This time, it is the demand curve that is shifting. Instead of the firm, it is the buyer who must compute the tax and send in the payments. A $50/unit tax will shift the inverse demand curve *down* (not up) by $50 because each consumer is willing to buy any given quantity for $50 less than before since she will have to pay an additional $50 to the government per unit purchased.

As before, a deadweight loss triangle appears when you impose the $50 tax. The tax drives a wedge between the total price the consumer pays and the amount the firm receives. This is the height of the triangle.

The deadweight loss triangle's width is the difference between the initial and new Q_e. The equilibrium quantity is driven down by the tax and, therefore, it no longer equals the socially optimal quantity. The tax causes an inefficient allocation of resources. The deadweight loss of \$496 is a measure of the inefficiency caused by the tax.

The tax incidence can be found by computing the share of the tax paid by the consumer versus the firm. The sellers receive a price of \$89.68 so they bear roughly \$10 of the \$50 tax. The consumer pays the firm \$89.68 and the government \$50 for each unit for a total price of \$139.68. The buyer's share of the tax is about 80%.

The government's revenue is the \$50 tax on each unit sold times the new equilibrium quantity, 105.16. This yields \$5,258 and can be represented as a rectangle in the supply and demand graph.

It is obvious that these numbers are the same as the suppliers pays scenario, but a fun and memorable way to show that it does not matter who pays the government is to toggle back and forth between the two sheets.

STEP Click the *SupplierPays* sheet tab, then click the *DemanderPays* sheet tab. Repeat this several times while keeping your eye on the screen. What do you notice?

The chart is different, of course, and the *d0* and *s0* parameters are different because the demand and supply intercepts do change based on who collects the tax for the government. But the price paid by the consumer, the price received by the firm, government revenue, and, most importantly, equilibrium quantity and deadweight loss are all exactly the same.

There is no doubt about it—tax incidence and deadweight loss do not depend at all on who physically collects and sends tax payments to the government (compliance being equal). If it does not matter if the buyer or seller pays the tax, then what do tax incidence and deadweight loss depend on?

Elasticities Drive Tax Incidence and Deadweight Loss

The relative price elasticities of demand and supply determine both the tax incidence (the distribution of the tax burden) and the deadweight loss (the measure of inefficiency in the allocation of society's resources).

Nothing else matters—certainly not who collects the tax, but also nothing else either. Price elasticities of demand and supply are all you need.

The more inelastic is demand at the initial equilibrium price, given supply, the more the consumer will bear the burden of the tax and the lower the deadweight loss. Likewise, the more inelastic is supply at the initial equilibrium price, given demand, the more the supplier bears the burden of the tax and the lower the deadweight loss.

We return to the apartment rent example to see how supply and demand analysis would work in an extreme case. If you agree to a \$100 increase in rent, your demand for apartments is perfectly inelastic in this price range. The price increase from \$700 to \$800 has no effect on the quantity demanded. In this case, you bear the entire burden of the tax and there is no deadweight loss. The situation is depicted in Figure 17.15.

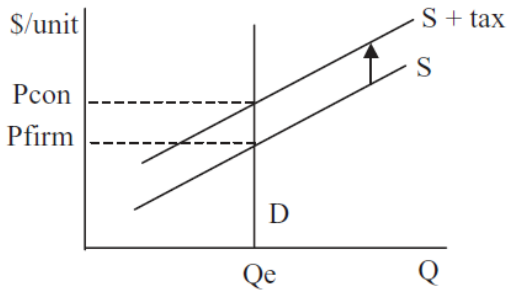


Figure 17.15: Tax effects with perfectly inelastic demand.

If you had to pay the property tax, you would be unable to shift it onto the landlord. In Figure 17.15, D would shift down, but it is a vertical line so it would shift on top of itself. The landlord would get \$700 from you (the initial equilibrium price) and you would pay an additional \$100 to the government.

Our Tax Incidence Formula yields the same result. With perfectly inelastic demand, $\epsilon_D = 0$. Thus, we have:

$$1 - \frac{\epsilon_D}{\epsilon_D + \epsilon_S} = 1 - \frac{0}{0 + \epsilon_S} = 100\%$$

This says the buyer bears the burden of the entire tax. Notice the formula does not have an input for who is writing the check to the government—that does not affect the outcome at all.

The formula also tells us that ϵ_S does not matter at all in the extreme case of perfectly inelastic demand. Any price elasticity of supply greater than zero leaves the buyer bearing the full burden of the tax.

The situation is reversed, of course, for the tax incidence if supply is perfectly inelastic. We would have a vertical S line that shifts up onto itself when the supplier pays the government. This leaves equilibrium price and quantity unchanged so the consumer pays the same amount as before and bears none of the tax burden. Once again, deadweight loss is zero.

Once again, the Tax Incidence Formula gives the same result. With $\epsilon_S = 0$, the ϵ_D in the numerator and denominator cancel and we get zero. This means the consumer bears no burden from a tax on a perfectly inelastically supplied good.

Of course, the main result that relative price elasticities determine tax incidence and deadweight loss applies in general and not just to these extreme cases. We can demonstrate this with the Excel workbook.

STEP To enable comparison, copy the *SupplierPays* sheet by right-clicking the sheet tab and selecting *Move or Copy*.) Select *SupplierPays* so the sheet is inserted before the *SupplierPays* sheet and check the *Create a Copy* box.

Excel inserts a new sheet in the workbook, named *SupplierPays (2)*. We will apply the same \$50 tax with a more elastic demand curve at the initial equilibrium price to see the effect on tax incidence and deadweight loss.

STEP Click the Reset button, then click the More Elastic Demand button in your new sheet.

A new, red inverse demand curve appears that is flatter, yet it goes through the initial equilibrium solution. The button simply sets the intercept and slope to 225 and 1, respectively. The price elasticity of demand at the initial equilibrium solution has risen (in absolute value) to -0.8 (as shown in cell E11).

It is important to not confuse slope and elasticity. The new, red inverse demand curve is more price elastic at $P = 100$ because it is flatter at that point. It is incorrect to say, however, that flatter lines are more elastic as a whole than steeper lines—both the initial and new inverse demand curves have varying elasticity all along the line. Thus, it does not make sense to say

that flatter lines are more elastic. Elasticity refers to a percentage change response at a point. Only at $P = 100$ do we know that elasticity is higher for the flatter, red inverse demand curve.

STEP Click the tax scroll bar five times to impose a \$50 per unit tax.

Figure 17.16 shows that the consumer bears less of the tax burden than before (but still more than the seller) and deadweight loss has risen.

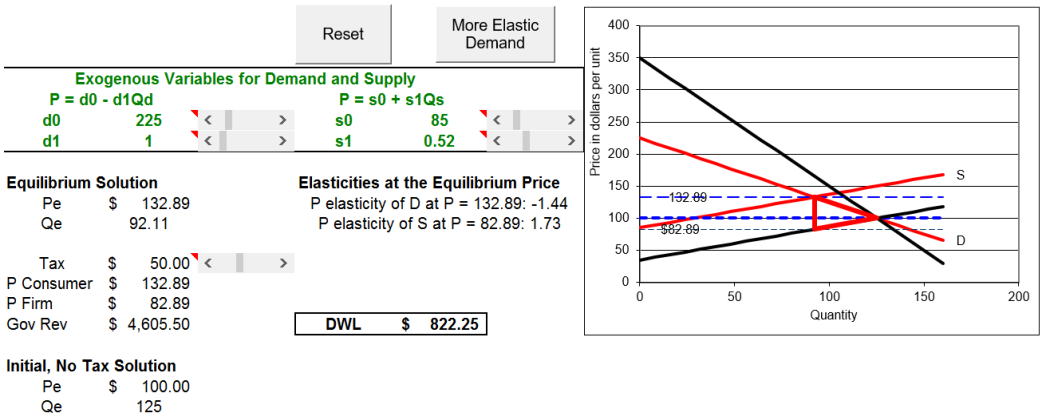


Figure 17.16: Tax effects with a more elastic D.
Source: Taxes.xls!SupplierPays.

With $\epsilon_D = 0.8$ instead of 0.4, ceteris paribus, the tax incidence on the consumer has fallen because the price has risen only to \$132.89 as opposed to \$139.68 on the *SupplierPays* sheet. So, the consumer bears \$32.89 of the \$50 tax or $\frac{32.89}{50} \approx 65.8\%$ of the tax. Notice that firms will now only net \$82.89 per unit instead of \$89.68 when $\epsilon_D = 0.4$. Suppliers tax burden rises to 34.2%.

The Tax Incidence Formula corroborates this result.

$$1 - \frac{\epsilon_D}{\epsilon_D + \epsilon_S} \text{ at } \epsilon_D = 0.4 = 1 - \frac{0.4}{0.4 + 1.54} \approx 79.4\%$$

$$1 - \frac{\epsilon_D}{\epsilon_D + \epsilon_S} \text{ at } \epsilon_D = 0.8 = 1 - \frac{0.8}{0.8 + 1.54} \approx 65.8\%$$

More importantly, deadweight loss has risen after the increase in the price of elasticity of demand from 0.4 to 0.8. Toggle back and forth from the original and new *SupplierPays* sheets to see that deadweight loss increases from \$496 to \$822.25.

While the height of the Harberger triangle has remained the same (the \$50/unit tax), the length has increased because the new equilibrium quantity is farther from the initial $Q_e = 125$.

If you toggle back and forth a few times, you can see how the more elastic demand curve is creating a *DWL* triangle that is longer, but with the same \$50 height. If you keep flattening the inverse demand curve (making sure that it passes through the initial equilibrium solution), the triangle keeps lengthening, but the height stays the same. A perfectly elastic (horizontal) D curve would produce the greatest deadweight loss possible.

STEP After thinking about it a bit, you can verify the claim above by using the control just to the right of the chart. Try all five scenarios.

With *Equal Burden* selected the demand and supply elasticities at $P = 100$ are the same so the \$50 tax is split evenly. The consumer pays \$125/unit and the firm receives \$75/unit.

The bigger drop in equilibrium output with more elastic demand is also responsible for the fall in government revenues. Instead of collecting \$5,258 in tax revenues, the government only gets \$4,605.50. It gets \$50/unit in both scenarios, but equilibrium quantity has fallen to 92.11 units with $\epsilon_D = 0.8$.

But neither the incidence of a tax nor the effect on government revenues is the highest priority issue. The top concern is the misallocation of society's scarce resources caused by taxation. It is this that leads to a theory of optimal taxation.

Optimal Taxation

Figure 17.15 shows why it makes sense to tax inelastically demanded goods. If we could find perfectly inelastically demanded or supplied goods, we would tax them because then we would not distort the allocation of resources.

Our goal is to raise government revenue for needed projects by causing the smallest misallocation of resources. Thus, the optimal tax is the one that has the least deviation of equilibrium output from optimal output, which is equivalent to minimizing deadweight loss.

Clearly, it is better, ceteris paribus, to tax goods with low price elasticities of demand or supply. In the introduction to this section, gasoline, cigarettes, and alcohol were mentioned as goods that carry quantity taxes. It is no surprise that these goods are quite price inelastic at their usual sales prices.

Granted, there may be other reasons to tax these products (and we will see one of them in the section on externalities), but to the extent that government seeks revenue from taxing individual products, it should tax those that will not lead to large deadweight losses.

There is no quantity tax on *Milky Ways*, a scrumptious chocolate candy. Obviously, the government could never generate the same tax revenue from *Milky Ways* as gasoline, but even if it could, with so many substitutes, *Milky Ways* must be very price elastic. A tax on *Milky Ways* would lead to a great fall in equilibrium output. Government revenue would be quite low and deadweight loss very high.

Elasticity Rules

Public Finance (also known as Public Economics) is a subdiscipline of economics that includes the study of government tax policy. The theory of optimal taxation focuses on the best way to tax. The analysis in this section says that quantity taxes should not be applied to goods that are relatively price elastic because the deadweight loss will be high. Instead, by taxing goods with inelastic demand or supply curves, government can raise needed revenue with a minimum of distortion in the allocation of society's resources.

This section also focused on the issue of tax incidence, who really bears the burden of a tax. This is a secondary issue compared to that of the optimal allocation of resources, but there is a surprising key result: It does not matter who collects the tax for the government (ignoring administrative costs and assuming equal compliance) because that party may be able to shift the tax onto someone else. Like deadweight loss, the tax incidence depends only on the elasticities of demand and supply. The more inelastic one of the curves is versus the other, the more that party will bear the burden of the tax. The Tax Incidence Formula sums this up conveniently:

$$1 - \frac{\epsilon_i}{\epsilon_D + \epsilon_S} \text{ for } i = D, S$$

Unfortunately, it is easy to confuse elasticity and slope. Do not fall into the trap of thinking that flat demand and supply curves are elastic and steep ones inelastic. If linear, slope is constant, but elasticity varies—for a linear, inverse demand curve, it rises as you go up and get closer to the vertical axis. Although you should not describe an entire curve as elastic or inelastic, you can correctly infer that where two lines cross, the flatter one is more elastic.

The French economist Frederic Bastiat (1801 - 1850) had a clever way of explaining what economists do. In his final essay, titled "What is Seen and Unseen," Bastiat argues we need to be aware of invisible costs and effects.

Taxes are a good example. It is easy to think that property taxes are paid by property owners, but this is simply not necessarily true. What is seen, a tax payment, is not the whole story. It is amazing, but true, that who pays the tax bill is irrelevant. It is also amazing that price elasticities, which are unseen, completely determine tax incidence and deadweight loss.

Exercises

1. Do we get the same result if we have consumers or firms pay the tax to the government with a perfectly inelastic supply curve? To support your answer, use Word's Drawing Tools to draw graphs. Explain the graphs and the result.

2. Use Word's Drawing Tools to draw a graph where supply is more inelastic than demand at the initial equilibrium price. Apply a quantity tax. Comment on the tax incidence and deadweight loss.

3. In 1937, when Congress set up the Social Security system, it was decided that firms and workers each pay half of the total tax so the tax burden is equally shared. Today, workers and employers each pay 6.2% of wages up to maximum that changes each year. Do you think that by each party paying the same tax the burden is equally shared? Why or why not?

4. Suppose the demand for labor is more elastic than the supply of labor at the equilibrium wage. Use Use Word's Drawing Tools to draw a graph that shows the tax incidence of the Social Security tax.

   Hint: You have to shift both demand and supply by the same amount, and then find the new equilibrium point.

References

The epigraph comes from page 168 of James R. Hines, Jr., "Three Sides of Harberger Triangles," *The Journal of Economic Perspectives*, Vol. 13, No. 2 (Spring, 1999), pp. 167–188, www.jstor.org/stable/2647124. Hines explains that the theory of deadweight loss dates back to Dupuit, Jenkin, and Marshall, but Harberger's papers in the 1950s and 1960s "illustrated the techniques, the usefulness, and the realistic possibility of performing such calculations, and in so doing, ushered in a new generation of applied normative work" (p. 168). For this reason, argues Hines, "Welfare loss triangles are 'Harberger triangles' because Harberger's papers measured them, did so in a consistent manner, and assisted and encouraged a host of others to do likewise" (p. 185).

Arnold Harberger published a number of papers, but perhaps his key contribution was "The Measurement of Waste," *American Economic Review*, Vol. 54, No. 3 (May, 1964), pp. 58–76, www.jstor.org/stable/1818490.

> Marginal cost pricing as a policy is largely without merit. How then can one explain the widespread support that it has enjoyed in the economics profession? I believe it is the result of economists using an approach which I have termed "blackboard economics."
>
> Ronald Coase

17.4 Inefficiency of Monopoly

Partial equilibrium analysis is based on the idea that each good and service with resources allocated via the market system has supply and demand curves. Prices signal quantities demanded and supplied and are pushed toward equilibrium by market forces. The equilibrium quantity is the market's answer to society's resource allocation problem.

If an omniscient, omnipotent social planner, OOSP, were to maximize the consumers' and producers' surplus of an individual good or service, she would explicitly order the production of the socially optimal amount of each good and service.

A critical result from this analysis is that a properly functioning market's equilibrium quantity equals the socially optimal quantity. This is what we mean when we say that a properly functioning market correctly solves society's resource allocation problem. There is no deadweight loss because the correct output is produced.

This section focuses on the following question: What happens if one of the goods is produced by a single seller (instead of the many individual firms that define perfect competition)?

In other words, we explore the welfare effects of monopoly. Our analysis is based on partial equilibrium and uses the tools of consumers' and producers' surplus. We evaluate monopoly by figuring out what a monopolist would produce, and then compare the monopoly output to the socially optimal output.

STEP Open the Excel workbook *MonopolyDWL.xls*, read the *Intro* sheet; then go to the *PC* sheet.

The linear demand and supply curves have the same parameter values used in previous examples. The equilibrium price is $100, which yields an equilib-

rium output of 125 units. Because the socially optimal level of production is also 125 units, the market yields an efficient allocation of resources.

Notice that at the socially optimal and competitive market solution, since supply is the sum of firm's marginal costs, we know that aggregate marginal cost equals demand. This is called *marginal cost pricing* and is indicative of a socially optimal solution. We will see in a moment that monopoly does not share this property.

The Monopoly Solution

Suppose all of the firms that produce a product in a perfectly competitive market were to merge into a giant, single firm. We assume that the cost structure stays exactly the same. In other words, the supply curve, which was the sum of the individual marginal cost curves, now becomes the monopolist's marginal cost curve.

Assuming that the costs of many firms would be the same costs faced by a single firm is a stretch. After all, the monopolist needs only one CEO and one customer service hotline. In other words, there are likely to be economies of scale in administration, distribution, and other areas. We assume this away in our comparison of perfect competition and monopoly. The idea is that the only difference is in the impact on the observed output when we have many firms in competition versus a single firm.

The monopolist will behave differently than the many firms did because there is no competition. Unlike the competitive result, where price is determined by the interaction of many buyers and sellers, the monopolist will choose the profit-maximizing price and quantity.

Chapter 15 explained monopoly profit maximization. What is different in this section is that, after determining the output chosen by the monopolist, we want to evaluate it using the tools of partial equilibrium analysis.

Our path is straightforward: we will solve the monopolist's problem with analytical and numerical methods, then we judge the monopoly outcome.

We know the monopolist will maximize profit by finding that quantity where $MR = MC$. The former is given by the demand curve, but what about MC?

The MC function is given by the supply curve parameters in the PC sheet. Once a monopoly takes over, it does not have a supply curve, but it does have a marginal cost function, which is the same as the supply curve (because of our assumption that there is no difference in costs between a competitive industry and a monopoly).

Thus $MC = 35 + 0.52Q$ and we can derive demand from the demand curve, as we have done before:

$$TR = P(Q)Q = (350 - 2Q)Q = 350Q - 2Q^2$$

$$MR = \frac{dTR}{dQ} = 350 - 4Q$$

As expected, we see that MR has twice the slope of the demand curve.

To find the monopolist's optimal Q, we set $MR = MC$ and solve for Q^*:

$$350 - 4Q^* = 35 + 0.52Q^*$$

$$4.52Q^* = 315$$

$$Q^* \approx 69.69$$

To find P^*, we use the demand curve to compute the highest price obtainable for that quantity.

$$P = 350 - 2Q = 350 - 2[69.69] \approx \$210.62$$

STEP Proceed to the *Monopoly* sheet to use numerical methods.

The graph has been augmented with the MR curve and the supply curve is now labeled MC. The MR curve was always there, but perfectly competitive firms cannot exploit it.

The sheet shows the monopoly price and output in cells B15 and B16 based on the analytical solution. Before we examine the deadweight loss and surplus information, we confirm that numerical methods agree.

When you run Solver, notice that the Solver dialog box is set up to choose that quantity that sets cell B20 to zero. The initial output of 50 units is too low. The fact that $MR - MC$ is \$89 means that the 50th unit of output adds \$89 more in profits and, therefore, more should be produced.

STEP Run Solver to find the Q that sets $MR - MC$ equal to zero.

After running Solver, you should see that cell B20 equals zero and that the Solver solution agrees (not exactly, but practically speaking) with the analytical method. This is not a surprise.

We now arrive at the key moment. How to judge the monopoly solution?

Evaluating Monopoly

We know the monopolized market will have an optimal output of 69.69 units and a price of \$210.62/unit. The evaluation of this outcome is based on computing the consumers' surplus, *CS*, and producers' surplus, *PS*, generated by the monopoly, and then comparing it to the socially optimal result.

The socially optimal result, at $Q = 125$ units, yields \$19,688 of total surplus.

STEP Cell F19 displays \$15,625 of consumers' surplus. Click on the cell to see its formula: = 0.5*(d0_ − P)*Q. *P* and *Q* are named cells for the perfectly competitive solution of 100 and 125, respectively.

Cell F20 has producers' surplus at $Q = 125$. Cell F21 adds *CS* and *PS*. The total surplus of \$19,688 is the maximum surplus possible and it is obtained when 125 units are produced.

Now, consider what happens under monopoly.

STEP Cell I19 shows a dramatic drop in *CS*. Click on the cell to see its formula: =0.5*(d0_-Pm)*Qm. *Pm* and *Qm* are named cells for the monopoly price and output.

The monopolist has lowered output and raised the price, relative to the competitive solution. This has greatly reduced consumer's surplus.

Cell I20 shows producers' surplus. It has more than doubled from what it was when the market was competitive. Its formula is: =(Pm-I18)*Qm + 0.5*(I18-s0_)*Qm. The first part of the formula is a rectangle. The height is the monopoly price minus the *MR* (or *MC* given that they are equal). The width is the monopoly output. A large part of this rectangle—from the monopoly price to the perfectly competitive equilibrium price—used to

belong to the consumers. It has been taken by the monopolist and helps explain why *CS* and *PS* have changed so dramatically.

So, CS has fallen and PS has risen, what is the overall outcome? Cell I21 adds *CS* and *PS* under monopoly. The total surplus of \$15,833 is lower than the maximum possible surplus of \$19,688. The difference, \$3,855 (in cell I23), is the lost surplus due to monopoly. This is also known as the deadweight or welfare loss.

STEP Click the Show DWL in Chart button to see a visual presentation in the graph of the deadweight loss of monopoly. It is a Harberger triangle.

Figure 17.17 is a canonical graph in microeconomics. It shows that the monopoly output is too low (so too few resources are allocated to this market) and the deadweight loss or Harberger triangle is used to indicate the inefficiency generated by monopoly.

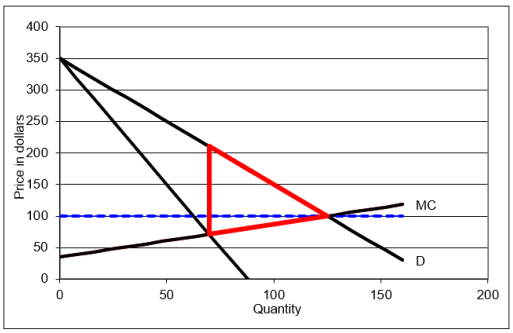


Figure 17.17: The deadweight loss from monopoly.
Source: MonopolyDWL.xls!Monopoly.

Because the monopoly solution does not equal the socially optimal output, we say there is a *market failure*. It is a failure in the sense that resources are not optimally allocated from society's point of view.

Inframarginal thinking can be applied to Figure 17.17. The basic idea is that all of the output in the range from the monopoly solution, roughly 70 units, up to the socially optimal output level of 125 units, exhibits unrealized gains from trade. For example, the marginal cost of producing the 100^{th} unit is $35 + 0.52 \times 100$, which equals \$87. The demand curve tells us that consumers are willing to pay up to \$150 for the 100^{th} unit. Clearly, the 100^{th} unit should

be produced because the additional satisfaction (as measured by willingness to pay) is greater than the additional costs of production.

The monopolist refuses to produce and sell the 100^{th} unit, however, because of an implicit restriction. Monopoly power allows the firm to set the price, but all units must be sold at the same price. Selling the 100^{th} unit at a price of \$150/unit means that all units must be sold at this price. Doing this would lower monopoly profit.

But the partial equilibrium welfare analysis critique of monopoly does not ride on the fact that monopoly forces consumers to pay higher prices than under a competitive market. The real problem with monopoly is that it produces too little output—it produces less than the socially optimal level. This causes too few resources to be allocated to the production of the monopolized good or service. We measure the amount of this inefficiency in resource allocation by the deadweight loss.

Yet another way to frame the inefficiency of monopoly is to focus on the fact that the monopolist produces where $MR = MC$ and this differs from $P = MC$ because MR diverges from the demand curve. A competitive market yields a socially optimal output because output is produced up to the point at which marginal cost equals the price (i.e., marginal cost pricing).

Figure 17.17 makes clear that the monopolist does not conform to marginal cost pricing. $MR = MC$ yields the output that maximizes profits, but $P = MC$ (where demand intersects supply or the aggregate marginal cost curve) is the socially optimal output. The monopolist is not interested in social optimality and, therefore, does not obey marginal cost pricing.

Elasticity Rules Again

In the previous section, we saw that the deadweight loss from a quantity tax depended on the price elasticities of supply and demand. The same holds true for monopoly.

STEP In the *Monopoly* sheet, display the red *DWL* triangle (if needed), and click the D More Elastic button .

Demand is flatter, while going through the same competitive equilibrium

point, $Q = 125$, $P = 100$. Thus, demand is more elastic at this point.

The button is actually a toggle. By clicking it repeatedly, you can switch back and forth from the original, more inelastic demand (price elasticity of -0.4 at $P = 100$) to the more elastic demand (price elasticity of -0.8 at $P = 100$).

STEP Click the D more and less elastic button a few times to convince yourself that the deadweight loss from monopoly is in fact larger when demand is more inelastic at $P = 100$.

While cell E17 shows that DWL is higher when demand is more inelastic at $P = 100$, we can make a graph that clearly shows this.

STEP Copy the *Monopoly* sheet and make the elasticity on the new sheet different than on the original sheet. Copy the chart in one sheet and paste it on top of the chart in the other sheet.

There is no fill in the chart so it is transparent. Your chart should look like Figure 17.18.

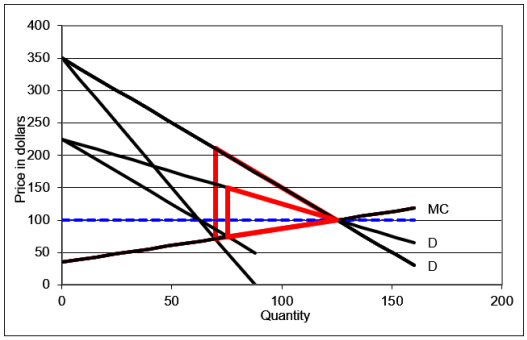


Figure 17.18: Comparing deadweight loss with different price elasticities of demand.

In Figure 17.18, the larger red triangle is the deadweight loss of \$3,855 in the initial case, with a price elasticity of demand of -0.4 at $P = 100$. The smaller red triangle is DWL with more elastic demand of -0.8. The DWL is lower, falling to \$1,870, when demand is more elastic.

Deadweight loss falls when demand is more elastic because the output does not deviate as much from the socially optimal result and the monopoly price is much lower. Hence, the Harberger triangle is both shorter and thinner.

Intuitively, the more inelastic is demand at given price, the greater is the monopoly power. A monopolist who enjoys demand that responds little to price is able to charge very high prices and the gap from marginal cost to demand for the inframarginal units will be large. This increases the deadweight loss from monopoly.

This example shows why economists use deadweight loss to measure inefficiency instead of simply the deviation in output from its optimal value. The monopolist does not change output by much when demand is more inelastic (-0.4), but the fact that consumers are willing to pay a lot more for the inframarginal units drives the large increase in deadweight loss.

Notice that the effect of elasticity on DWL is different than what we obtained for quantity taxes. In that case, more inelastic demand led to lower deadweight loss. The effect is reversed with monopoly, but the principle that elasticity rules remains true.

Monopoly and Price Discrimination

Although we usually assume a monopolist must charge the same price for all units sold, sometimes a seller can charge different prices for the same product. This is known as *price discrimination* and it enables profits to be even greater than when a single price is charged to all customers.

Charging different prices to see a movie in the afternoon versus the evening, different prices for coach versus first-class on a plane or train, and different net tuition to students (in the form of differing amounts of financial aid) are all examples of price discrimination. In each case, the firm is able to increase its profits by separating consumers into different groups and charging them different prices for the same good or service.

Sometimes firms try to slightly change the product so it isn't so obvious that the exact same thing is being sold at different prices. Offering first-class passengers pre-boarding and free drinks on a plane is an example of this. As is the bigger portions of a dinner versus lunch version of a dish at a restaurant. The difference in prices for the first-class and dinner versions of these products is not grounded in higher costs.

What is really going on here is coming entirely from the demand side. Some consumers are willing to pay more and firms are taking advantage of this.

People can get really upset at price discrimination. Dry cleaners can get in hot water when they charge different prices for cleaning men's versus women's clothing that is almost identical. It can be a fun game to spot examples of price discrimination.

There are three requirements for price discrimination to work:

1. Some degree of monopoly power (facing a downward sloping demand curve).

2. The firm must be able to segregate customers into groups (splitting the overall demand curve into subgroup demands).

3. There must be a way to seal the markets to prevent resale from the low-price to the high-price market, which is called *arbitrage*.

Assuming these requirements are met, we can construct a simple example that illustrates the essential logic of price discrimination. To increase profits, the idea is to separate price sensitive from insensitive consumers and then charge insensitive ones more.

STEP From the *Monopoly* sheet, click the Reset button and change cell E8 to 0 (zero).

This makes *MC* constant at \$35/unit and makes it easy to find the optimal solution and deadweight loss.

STEP With *MC* constant at \$35/unit, run Solver to find the monopolist's optimal solution.

Your screen shows that the monopolist will produce 78.75 units of output and charge a price of \$192.50. *CS* under monopoly is \$6,202 and *PS* is \$12,403.

STEP Click the Show DWL in chart button to display the Harberger triangle. Its area of \$6,202 is the *DWL*.

The fact that *CS* equals the *DWL* is not a coincidence. This is a property of linear demand and constant *MC*.

Now, suppose that this monopolist can separate the overall market demand, given by the inverse demand function of $P = 350 - 2Q$, into two separate

markets with two subdemands. For example, the two subdemands could be given by

$$\text{Market 1: } P = 450 - 6Q$$

$$\text{Market 2: } P = 300 - 3Q$$

The coefficients in the two separate markets must be consistent with the coefficients in the overall market inverse demand curve. The intercept and slope are not randomly drawn. If the price is zero, quantity demanded in Market 1 is 75 (= 450/6), while Market 2's quantity demanded would be 100 (= 300/3). The sum of the two is 175, which equals the quantity demanded at $P = 0$ using the overall inverse demand curve. At $P = 300$, Market 1's quantity demanded is 25 and Market 2's is zero, and this sum equals the quantity demanded using the overall demand curve.

How can a monopolist take advantage of the ability to separate the overall market into two sealed, separate subdemands?

The intuitive answer is simple: Instead of charging the same price, $192.50, to all customers, increase the price in the market with demand less sensitive to price and reduce it in the other market. The customers in Market 1 can be charged a higher price than those in Market 2. This will lead to greater profits.

We can see a concrete demonstration of this and figure out exactly what prices we should charge in our example.

STEP Proceed to the *TwoPriceDisc* sheet to see this plan in action.

Unlike the *Monopoly* sheet, there is no need to run Solver. The analytical solution has been entered and all cells and charts will instantly respond to changes in parameter values.

The top of the sheet shows how a perfectly competitive market would behave if there were two separate markets. Marginal cost pricing would result from competition so both markets would have the same price of $35/unit (cells B11 and B15). Market 2 would produce slightly more (B16) than Market 1 (B12), but the sum of the two (B20) would equal the perfectly competitive output of perfect competition for a single market. Thus, the ability to price discriminate, separating a single market demand into two separate, sealed subdemands, has no effect under perfect competition.

The outcome is different for monopoly. We begin by pointing out that the price elasticity of demand, while quite inelastic in both submarkets, is higher in Market 2 at the perfectly competitive price of \$35/unit.

The chart in the sheet is reproduced in Figure 17.19 and helps explain what is going on. This clever display shows the conventional monopoly graph for Market 1 on the right and uses the left side as a mirror for Market 2. Although the x axis shows output as negative on the left side, that is just a consequence of using Excel to draw the chart. Read the output as a positive number.

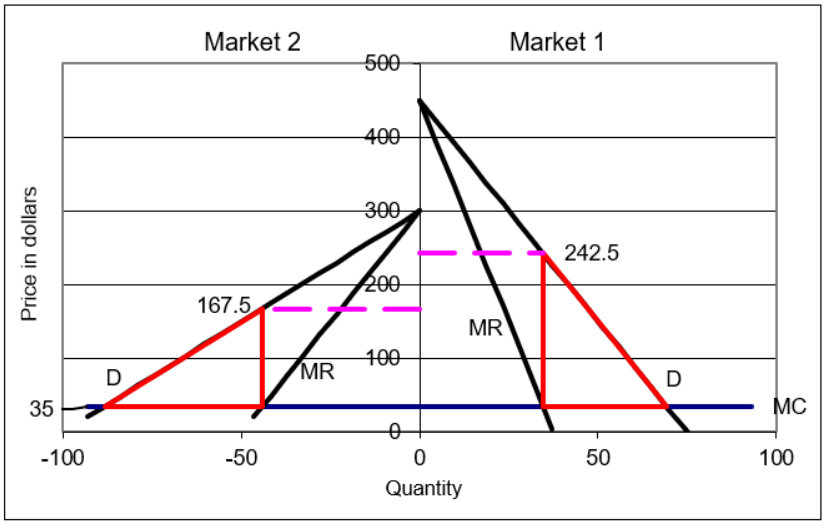


Figure 17.19: Price discrimination.
Source: MonopolyDWL.xls!TwoPriceDisc.

Figure 17.19 shows that the price discriminating monopolist will choose output where $MR = MC$ in each market, then charge the highest price obtainable for that output in each market. The price in each market is indicated by the dashed line and it is clear that price is higher in Market 1. This makes sense because consumers in Market 1 are less price sensitive so the monopolist takes advantage of this to generate higher profits.

STEP To easily compare the results of the single-price monopolist in the *Monopoly* sheet to the price discriminator in the *TwoPriceDisc* sheet, click the Show Single Price Monopoly button and notice that Market 1 is more price inelastic at the single-market monopoly price.

The price discriminator has the same total output, but it splits the single price into two prices. Cell B34 in the *TwoPriceDisc* sheet computes a weighted average of the two prices and it is higher than the single price of $192.50 charged by the conventional monopolist. This enables the two-price monopolist to make greater profits, as shown by the increase in PS from $12,403 to $13,028.

The monopolist will always be able to increase profits if it can split a market and keep the submarkets sealed off from each other, as long as the submarkets have different price elasticities. If so, a monopoly will charge the relatively more inelastic market a higher price and this is the source of the increase in profits.

Profits will continue to rise as markets are ever more finely subdivided. Amazon and other online retailers use your previous buying history, click behavior, and other information to serve up a personal price, just for you. Search for amazon+pricing+algorithm to learn more.

If you never heard about this before and think this is eye-opening or maybe even unfair, think about what colleges and universities do. They require their customers to provide detailed financial information about their ability to pay. They will, naturally, explain this as a benign effort to help the disadvantaged, but you should be glad your grocery store does not do this to you when you walk in the door.

The welfare consequences of price discrimination are not as clear. Comparing cells L38 and H34 shows that *DWL* has increased from $6,202 to $6,514 when the monopolist separated the markets and charged different prices. Of course, the monopolist does not care about deadweight loss; she is focused on maximizing profits. We, however, use *DWL* to evaluate outcomes and we would rather have the single market than the two submarkets exactly because deadweight loss is higher with price discrimination.

Unfortunately, these results do not generalize so we cannot say this will always happen. Higher *DWL* with price discrimination is guaranteed only for linear demand functions. In general, with nonlinear demands, we cannot state with certainty the effects on output and welfare. In other words, it is possible for output to rise and *DWL* to fall with a two-price discriminating monopolist. The effect on output and *DWL* depends on the shapes of the individual market demand curves.

For a concrete scenario of price discrimination improving welfare, consider the following from Scherer, (1970, p. 259):

> It is possible, for instance, that no physician would be attracted to a small town if he were required to charge the same fee to rich patients as to poor. Since profits can be increased by discriminating, the added revenue attainable through discrimination may be sufficient to make the difference between having a service provided and not having it.

Returning to the idea of subdividing the market more finely, there is a special case of price discriminating monopoly power that is a bit mystifying, but does yield a definitive result. The *perfectly price discriminating monopolist* has the ability to charge different prices for different output levels down to each individual consumer. This remarkable power enables the monopolist to sell every unit of output at the highest price the market will bear.

In the *Monopoly* sheet, the first unit goes for \$348, the 100^{th} for \$150, and the 125^{th} is priced at \$100. The perfectly price discriminating monopolist takes every bit of consumers' surplus, making the greatest profit possible, but does produce the socially optimal level of output. Thus, she has no deadweight loss!

Pondering the idea of perfect price discrimination and the fact that we would judge it as a socially optimal outcome cements the idea of surplus and deadweight loss. As long as someone, anyone, even if it is a single monopolist, gets the surplus, we count it as a successful outcome. Deadweight loss is tragic precisely because no one gets it. Deadweight loss vaporizes surplus and it disappears into thin air.

Monopoly Results in Market Failure

Monopoly leads to market failure because, to maximize profits, it restricts output and, therefore, this produces a misallocation of resources. The canonical monopoly graph (see Figure 17.17) has MR splitting off of D so that $MR = MC$ is less than the optimal output where $P = MC$.

While most people do not like monopoly because it charges higher prices than a competitive market, this is not why economists dislike the monopoly outcome. Partial equilibrium supply and demand analysis is based on maximizing consumers' and producer's surplus. The logic of deadweight loss

rides on the idea of waste. The monopolist does not take advantage of infra-marginal sales that would lower its profit, but increase society's total surplus. Any mechanism that generates deadweight loss is said to fail in the sense of not generating an optimal solution.

Another difference in outlook is that economists do not believe monopolists are inherently bad folks. The monopolist, like the perfectly competitive firm and consumer, is optimizing. Monopolies are in a position to improve their individual outcome and they take advantage. According to the economists, put anyone of us in the same position and we do the same thing. Do not blame the monopolist; blame the market structure for the deadweight loss.

We conclude with some advanced and heretical thinking. There is another, radically different view of monopoly that is based on the work of Joseph Schumpeter (1883 - 1950). He argued monopoly was actually a good thing because he had an evolutionary, dynamic view of capitalism. Striving for monopoly drives capitalism and monopolies are toppled by new firms in a process he named *creative destruction*. This oxymoron conveys Schumpeter's vision of capitalism, with entrepreneurs engaged in an epic battle of rising firms slaying established leaders.

Schumpeter's perspective is not that of solving society's resource allocation problem. He considered this static optimization problem to be uninteresting because it did not apply to the real world and it had been solved already. He did not believe that price competition was the real driver of capitalism's success. For Schumpeter, the serious open problem was how and why markets generated so much innovation and growth.

One important difference between mainstream economics and Schumpeter revolves around the government's role. Partial equilibrium analysis says monopolies should be broken up because they generate a misallocation of resources. Schumpeterians reject the need for government to intervene, arguing that dynamic competition will erode monopoly positions through entrepreneurial innovation.

Take an *Industrial Organization* course, an upper-level elective taught in most economics departments around the world, to learn more about monopoly, price discrimination, and Schumpeter's ideas.

Exercises

1. To punish a monopolist, your friend suggests applying a quantity tax on the monopoly's commodity. Is this a good idea? Explain why or why not, using the initial values of the parameters for supply and demand in the *Monopoly* sheet for a concrete example.

2. Another friend suggests a quantity subsidy to eliminate the deadweight loss caused by monopoly. The idea would be to shift down MC via the subsidy until output equaled the socially optimal output. Does this make sense?

3. Consider a monopoly that sells its output in two completely separated and sealed markets. Marginal cost is constant at \$35 per unit.

   Inverse demand in the two markets is given by

   $$P_1 = 200 - 2Q_1$$

   $$P_2 = 300 - 3Q_2$$

   (a) Solve this problem via analytical methods. Report optimal quantity and price in each market. Use Word's Equation Editor as needed.

   (b) Solve this problem with the *TwoPriceDisc* sheet. Enter the appropriate coefficients on the sheet. Take a picture of the results and paste it in a Word doc.

   (c) Which market has a higher price?

   (d) How does the price elasticity of demand in each market affect the price in each market?

   (e) Which market has greater deadweight loss?

   (f) How does the price elasticity of demand affect the deadweight loss?

   (g) The overall market demand is given by $P = 240 - 1.2Q$. How does price discrimination affect welfare loss in this case?

4. Suppose that, in the long run, average cost is decreasing throughout and marginal cost is below average cost, as shown in Figure 17.20. This is called a *natural monopoly*. The profit-maximizing level of output for the monopolist is where $MR = MC$. The socially optimal result is where $P = MC$.

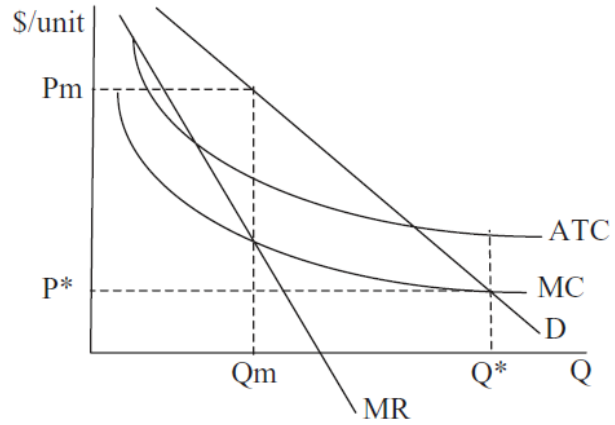


Figure 17.20: Natural monopoly.

(a) What is the problem with using competitive markets to achieve the socially optimal result in this situation?

(b) What government policy could be used to help the market reach the social optimum?

References

The epigraph is from page 19 of Ronald Coase, *The Firm, the Market and the Law* (paperback edition, 1990; originally published in 1988). Coase, a Nobel Prize winner for his work on transactions costs and property rights, criticized economics for its simplified, mathematical models that were completely stripped of real-world nuance and complexity. In "The Marginal Cost Controversy" (originally published in *Economica* in 1946 and reprinted in *The Firm, the Market and the Law*), Coase rejected the standard view that everywhere decreasing average cost (as in Figure 17.20) implied government intervention was the only solution.

Like Schumpeter, Coase has a broader, less mathematical view of economic analysis. Coase says on page 20 of *The Firm, the Market and the Law*,

> Blackboard economics is undoubtedly an exercise requiring great intellectual ability, and it may have a role in developing the skills of an economist, but it misdirects our attention when thinking about economic policy. For this we need to consider the way in which the economic system would work with alternative institutional structures. And this requires a different approach from that used by most modern economists

Visit www.hetwebsite.net/het/profiles/schumpeter.htm for more on Schumpeter. His most accessible work is *Capitalism, Socialism and Democracy*.

F. M. Scherer's *Industrial Market Structure and Economic Performance* came out in 1970 and was an instant classic Industrial Organization textbook. It was by far the most popular text of its day.

Today, almost all IO courses and books begin with a review of perfect competition and its polar opposite monopoly, but most of the course is about the complicated, fascinating, and vast area in between. Once you get past perfect competition and monopoly, the ways in which firms interact, compete, and strategize, is truly amazing.

We apply nonparametric regression models to estimation of demand curves of the type most often used in applied research. From the demand curve estimators we derive estimates of exact consumers' surplus and deadweight loss, which are the most widely used welfare and economic efficiency measures in areas of economics such as public finance.

Jerry A. Hausman and Whitney K. Newey

17.5 Sugar Quota

This section applies the tools of partial equilibrium analysis and deadweight loss to analyze import restrictions on sugar in the United States. Supply and demand analysis is shown to be a flexible, powerful tool.

Before analyzing the US sugar quota through the lens of surplus and deadweight loss, we take a crash course on sugar—production, pricing, and how import quotas on sugar are implemented.

Facts about Sugar

Everyone knows you can buy sugar in any grocery store and pour it into your coffee or use it to bake cookies. But there are many other kinds of white granulated sugars (like confectioners' sugar) and also brown and liquid sugars.

No matter the final form, "All sugar is made by first extracting sugar juice from sugar beet or sugar cane plants" (www.sugar.org/sugar/types/). Cane sugar is grown in warmer areas, whereas beets come from cooler climates. Once refined, you cannot easily tell the difference. Unless you are an expert, sugars from beet versus cane are perfect substitutes.

Some sugars are used only by industrial food manufacturers and not available in the grocery store. Home and commercial users can choose from many other sweeteners, such as high fructose corn syrup, and a long list of artificial sweetener options.

In addition to eating it, sugar can be made into ethanol and used to power a car. Most cars in Brazil are flex-fuel and growing huge quantities of cane has enabled Brazil to greatly reduce oil imports.

Many countries produce sugar. The United States grows both beet and cane sugar, but domestic production does not meet total demand so the United States imports sugar. Figure 17.21 is a subsection of a bigger table that shows sources of US sugar.

Table 1: U.S. sugar: Supply and use, by fiscal year (Oct./Sept.), May 2020

Items	2018/19	2019/20 (estimate)	2020/21 (forecast)
	1,000 Short tons, raw value		
Beginning stocks	2,008	1,783	1,273
Total production	8,999	8,024	9,005
Beet sugar	4,939	4,285	4,965
Cane sugar	4,060	3,740	4,040
Florida	2,005	2,100	2,105
Louisiana	1,907	1,513	1,800
Texas	147	127	135
Hawaii	0	0	0
Total imports	3,070	3,731	3,456
Tariff-rate quota imports	1,541	2,180	1,395
Other program imports	438	350	350
Non-program imports	1,092	1,200	1,710
Mexico	1,000	1,050	1,660
High-duty	91	150	50
Total supply	14,076.75	13,538	13,733
Total exports	35	35	35

Figure 17.21: US sugar sources.
Source: www.ers.usda.gov/webdocs/outlooks/98469/sss-m-381.pdf.

The numbers in the table come in units of short tons, raw value, STRV. A short ton is 2,000 pounds. Raw value means the dry weight of raw sugar. You get 1 ton of refined sugar (the white crystals you buy in the store) from 1.07 tons of raw sugar.

Beets are grown in many states so they are not all listed, but half of US beet production comes from the Red River Valley in Minnesota and North Dakota. The table shows the US domestic sugar industry is split roughly evenly between beet and cane, producing about 4,000 thousand STRVs (or 4,000,000 STRVs) from each crop.

Figure 17.21 makes clear that the United States imports a great deal of sugar, 3,070 thousand STRVs in 2018/19 and approaching 4,000 in 2019/20 (although this estimate was made before the covid 19 pandemic). So, roughly, the United States grows 2/3 of its own sugar and imports the rest.

Figure 17.21 shows that sugar is imported under several categories, the most important of which is the *tariff-rate quota*, TRQ. This is a complicated scheme for controlling the amount of sugar imported from different countries. The details are available at www.ers.usda.gov/topics/crops/sugar-sweeteners/policy.aspx.

A TRQ is a type of import restriction where a split tariff (or tax on imported goods) is employed. There is an extremely low tariff (zero or a nominal charge) applied to imports under a given amount (called the in-quota tariff) and a really high tariff applied to quantities imported beyond the given amount (so little is imported after the in-quota tariff is exhausted).

The TRQ was created in 1990 after multilateral trade agreements forced elimination of traditional quotas. In Europe, the EU Sugar Protocol is similar to the US TRQ system. The U.S. Department of Agriculture (USDA) runs the TRQ. The overall allotment is established by multilateral trade agreements and the USDA decides on the country allocations.

We can look at reports issued by the USDA as Excel spreadsheets to understand the TRQ.

STEP Open the Excel workbook *SugarQuota.xls*, read the *Intro* sheet, then go to the *TRQ* sheet and scroll around.

The data are constantly updated, so the specific numbers are not our chief concern. What matters is that column A has a list of countries and column O has *FY 2020 TRQ Original Allocations*.

As an example, consider the Dominican Republic. As of May 18, 2020, it had used 114,516 STRVs of its 185,335 TRQ allocation. The USDA has given every country in column A an amount that they can import. Beyond the TRQ amount, a hefty tax is applied so imports stop.

Outside of sugar producers and commercial food manufacturers that buy sugar, very few people in the US know or care much about this. In many countries, like the Dominican Republic, however, the US TRQ is a big news. When it is announced, there is intense media coverage and discussion.

If you scroll up and down, you will see that the Dominican Republic has the highest TRQ allocation, even bigger than Brazil, which is obviously a much larger country. What is going on here? In addition to protecting domestic US

sugar producers, the United States uses the TRQ as a major foreign policy lever, using allocations as punishments and rewards for foreign governments.

Now that we know a little about quantities of domestically produced sugar and imports, we turn to the price of sugar.

STEP Proceed to the *Price* sheet to see US and world raw sugar prices.

Figure 17.22 shows that prices have fluctuated over time, but US prices are always higher than world prices. The 1970s produced sharp spikes, followed by a period of calm until another spike during the Great Recession.

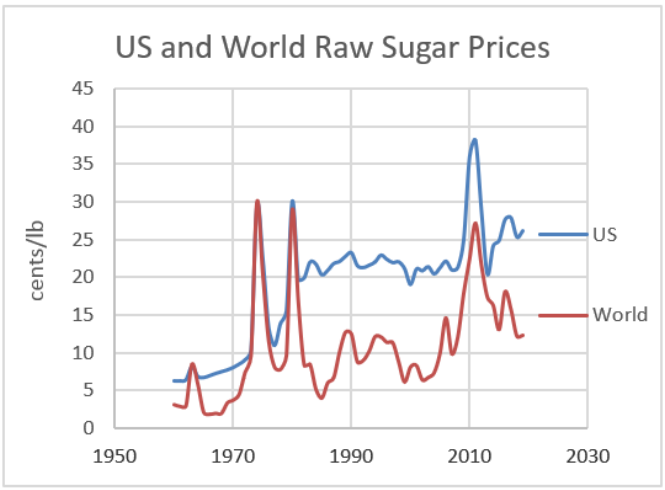


Figure 17.22: Nominal raw sugar prices.
Source: SugarQuota.xls!Price.

Since the TRQ was implemented in 1990, US sugar prices are consistently about 10 cents per pound higher than world prices. That might not sound like much of a difference, but think of it this way: US sugar prices are roughly double what others pay for sugar. If you make ice cream or candy or soft drinks or any one of the many products that uses sugar, doubling costs for this input is a really big deal.

STEP Review the price adjusted for inflation chart in the *Price* sheet.

The real price of sugar had been falling steadily, but it seems to have leveled off more recently. We can expect technological change (especially genetic engineering of cane and beet plants) to lower prices in the future.

We have ended our whirlwind tour of sugar production, the US TRQ system, and prices. Obviously, the sugar quota is causing higher prices for US consumers (including commercial buyers of sugar) and it benefits US producers. But we can say more and evaluate the US sugar quota by applying partial equilibrium analysis.

Supply and Demand for US Sugar

To analyze the effects of the sugar quota, we need estimates of demand and supply curves for sugar in the United States. Because we will work with linear functions, we need intercept and slope parameters for the demand and supply of sugar.

The USDA reports roughly 12,000 thousand STRVs of sugar are bought and sold in the United States for about 25 cents per pound for raw sugar. We assume the market is in equilibrium so we interpret these values as the equilibrium quantity and price.

There is a vast literature on sugar with countless estimates of demand, supply, and elasticities. Since this is an exercise in showing how partial equilibrium analysis works, we will use hypothetical demand and supply functions that are calibrated to the observed values in the US sugar market.

Our linear demand and supply curves are

$$Q_D = 15000 - 120P$$

$$Q_S = 400P$$

At $P = 25$ cents per pound, quantity demanded is 12,000 thousand STRVs (our equilibrium P and Q) and the price elasticity of demand is $\frac{\Delta Q_D}{\Delta P}\frac{P}{Q_D} = \frac{1}{-120}\frac{25}{12000} = -0.25$. That is quite inelastic and conforms with estimates of sugar price elasticities of demand. Although there are substitutes, in many recipes (especially for commercial products), precise amounts of sugar are absolutely required. The price elasticity of supply in our simple model is +1.

The inverse demand and supply curves are

$$P = 125 - \frac{1}{120}Q_D$$

$$P = \frac{1}{400}Q_S$$

You probably did not do this, but computing the quantity supplied from the supply curve with $P = 25$ gives $Q = 10000$. Something is wrong because the quantity demanded does not equal the quantity supplied. For sugar, we need to include imports.

Free Trade

We begin our partial equilibrium analysis of the US sugar quota in Fantasyland—we assume that there is no restriction of any kind on the importation of sugar.

STEP Proceed to the *FreeTrade* sheet to see how the market would work under a regime of no restrictions on imports.

Figure 17.23 reproduces the graph. The demand curve is straightforward, but the supply curve merits special attention.

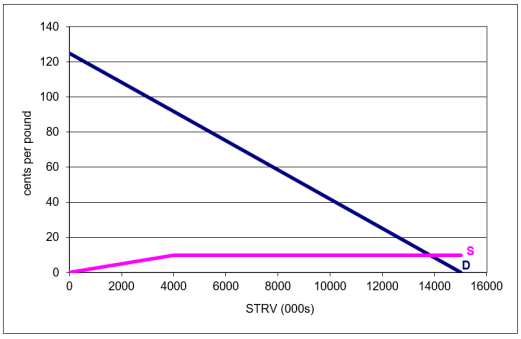


Figure 17.23: Supply and demand with hypothetical free trade.
Source: SugarQuota.xls!FreeTrade.

The first part of the supply curve (from the origin to the kink at $Q = 4000$) is domestically produced US sugar. As long as the price is below the world price of 10 cents per pound, the best, lowest cost US producers will supply the market.

Beyond 4,000 units (measured in thousands of STRVs for consistency with USDA TRQ units), world suppliers take over. It is assumed that the United States has access to as much sugar as it wants at the world raw sugar price of 10 cents per pound. Thus, the market would not continue to use US produced sugar beyond 4,000 units. Instead, supply would come from the perfectly elastic world supply curve.

US consumers (home and industrial buyers) would enjoy a 10 cent per pound price for raw sugar and the equilibrium quantity would be 13,800 thousands STRVs. Over 2/3 of sugar consumed would be imported.

The sum of US consumers' and producers' surplus would be more than $16 billion. In this properly functioning market, this is the maximum possible total surplus.

STEP Click on cells G33 and G34 to see the formulas used to compute *CS* and *PS*.

Notice that many US producers would be driven out of the market because they cannot make sugar at the low world price. Those US producers that remain (selling the first 4,000 units) would earn $400 million in producers' surplus under a free trade regime. As will be clear in a moment, this is an important number to keep in mind.

Incorporating an Import Quota

The TRQ system is too complicated to exactly implement in Excel so we model a simple quota that is easier to understand and acts similarly to the TRQ scheme.

STEP Proceed to the *ImportQuota* sheet to see what happens with an import quota on sugar.

As before, we focus on the supply curve. It is crucial to the analysis.

The *ImportQuota* sheet shows that the supply curve has an upward sloping part, then a flat part, and then it starts sloping up again. The first part is the US domestic supply curve. The lowest cost US firms will supply the market when the price is below the 10 cents per pound world price.

The flat part is the amount of imported sugar allowed. Cell H6 shows this amount is 2,000 units, so the flat segment is 2,000 units long.

The last, rising part of the supply curve is, once again, the domestic US supply curve. Once the quota is filled and no more foreign sugar is allowed into the United States, domestic producers that could not survive in a free market supply sugar.

Notice how the supply curve is pink, indicating it is domestic US sugar, at low and high levels of output. Imports snap the US supply curve, inserting a flat portion of length equal to amount of imports.

Cells I6 and J6 report equilibrium price and quantity (where S and D intersect). Compared to the *FreeTrade* sheet, P_e has risen from 10 to 25 cents per pound, while quantity has fallen from 13,800 to 12,000 thousand STRVs (2,000 of which are imported). Remember, we chose parameter values for the supply and demand curves to match real-world data from the US sugar market.

STEP Move the import slider control left and right to see how the import allotment affects the supply curve.

As you increase the amount of imports, you lengthen the flat segment and push the pink part of the S curve to the right. Decreasing the import allotment does the opposite. The beginning of the supply curve, below 4,000 units remains unchanged.

It is also easy to see how tightening the import allotment increases the equilibrium price and lowers the equilibrium quantity. Relaxing the imports allowed does the reverse.

STEP Enter 9800 in cell H6. This is the same as moving the import slider control all the way to the right.

This mimics the *FreeTrade* sheet. The import allotment is set so high that foreign sugar producers supply all of the US market after the first 4,000 units. Equilibrium price falls back to its free trade level of 10 cents per pound and quantity rises to 13,800 units.

Evaluating an Import Quota

We know import quotas raise prices and lower output, but this is just the outcome of the mechanism. To evaluate import quotas, we use the concepts of surplus and deadweight loss.

STEP Return the import quota to 2000 in cell H6 and then click the *Show CS* checkbox (cell C7).

Cells are displayed in columns A and B that are the source data for the blue consumers' surplus triangle that has been added to the chart. Under the sugar quota, the *CS* no longer extends to the world price of 10 cents per pound and quantity is smaller than the optimal quantity. Consumers lose $3.87 billion in surplus compared to the optimal solution.

STEP Click the *Show US PS* checkbox.

This adds the producers' surplus gained by sugar manufacturers in the United States. Their total *PS* is composed of two separate parts. On the left is a trapezoid and on the right is a triangle. What is in the middle?

STEP Click the *Show Foreign PS* checkbox.

The orange rectangle added to the chart is *PS* that goes to foreign producers. Notice that this is not deadweight loss because someone gets it.

Clearly, producers' surplus is much higher with the quota, rising from a mere $400 million with free trade (which equals the optimal solution) to $2.5 billion for US producers and $3.1 billion for all producers.

The transfer of *CS* to *PS* under the quota system, however, is not without waste. Consumers lost $3.87 billion of surplus and producers gained $2.7 billion. What happened to the rest?

STEP Click the *Show DWL* checkbox.

The red triangle has an area of $1.17 billion. This is the amount of *CS* that was lost during the transfer of surplus from consumers to firms. Figure 17.24 shows the chart with all checkboxes checked.

A leaky bucket is an apt metaphor. While siphoning off billions of dollars from consumers and delivering them to producers, $1.17 billion leaked and was wasted, captured by no one. We can express the leakage as a percentage, $\frac{1.17}{3.8} \approx 30\%$. That is a pretty big hole in the bucket.

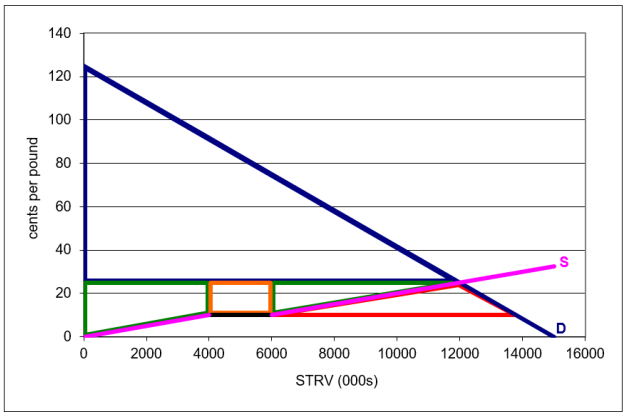


Figure 17.24: Partial equilibrium analysis of a sugar quota.
Source: SugarQuota.xls!ImportQuota.

Notice the geometry in this example. The *DWL* triangle in Figure 17.24 is not the usual bow tie shape (as in the price ceiling, tax, and monopoly applications). In this case, the *DWL* is a triangle under supply and demand. But the interpretation is the same—we are measuring surplus foregone and using this as an indicator of the damage done by the misallocation of resources.

You might wonder why consumers are not up in arms. In fact, commercial sugar buyers do lobby Congress and when prices spike, the TRQ allotments are relaxed. The vast majority of buyers in the supermarket, however, simply have no idea that this is happening. A five pound bag of refined sugar that costs $2 is just another item in the shopping cart.

This is a common problem surrounding import quotas: costs are diffused widely while the benefits are concentrated on a few key players. Thus, although the costs add up to a large number, $3.7 billion in this example, no one individual is impacted enough to object. The handful of US sugar producers, however, have strong incentives to maintain the system to keep their profits. You will see what this means when you answer the last exercise question.

The transfer of surplus, no matter how unfair it may seem, is not the real problem in the eyes of partial equilibrium analysis. The fact that surplus is vaporized and vanishes into thin air so no one gets it—this is the real problem.

It is easy to be confused by the shapes on the graph and concerns that prices are higher and producers are stealing surplus from consumers. None of that really matters. Here is the takeaway: the import quota is causing

a misallocation of resources. The United States is using land, labor, and capital to make sugar when it would be better off buying foreign sugar and using these inputs to make other goods and services.

Comparative Statics

We can explore the effects of changing demand and supply coefficients on the equilibrium price and quantity of sugar, but the natural question to ask is, what is the effect of the import allotment? We are chasing the *import elasticity of price* and the *import elasticity of quantity*. We want to know how responsive price and quantity are to shocking the import allotment.

We can also explore how the surplus and deadweight loss changes. These variables are also endogenous in this model because they are generated by the forces of supply and demand.

We have the initial position. With H6 = 2000, $P_e = 25$ and $Q_e = 12000$.

STEP Set H6 to 3000.

The length of the orange rectangle expands and the rising part of the US supply curve is pushed right. Equilibrium price falls to just over 23 cents per pound and output rises to about 12,231 thousand STRVs. *CS* and foreign *PS* rise. Deadweight loss falls. This is better for US consumers and foreign sugar producers than the initial quota of 2,000 units.

United States *PS*, however, falls. Domestic sugar producers are not happy with this. They prefer a lower import quota.

Elasticities give us more information than the qualitative statements (up or down) made above. We can compute the percentage change in price, quantity, surplus, and deadweight loss for the 50% increase in import (from 2,000 to 3,000 units).

The import elasticity of price $\approx \frac{\frac{23-25}{25}}{\frac{3000-2000}{2000}} = \frac{-0.8}{0.5} = -0.16$. This tells us that equilibrium price is quite unresponsive to the import allotment.

The import elasticity of quantity $\approx \frac{\frac{12231-12000}{12000}}{\frac{3000-2000}{2000}}$ $approx \frac{-0.02}{0.5} = -0.04$ is even smaller. Equilibrium quantity is extremely unresponsive to the import allotment.

These elasticity estimates are for illustration. Our model relies on rudimentary, linear demand and supply curves. The framework, however, is exactly how an economist would model the sugar market and interpret the effects of a sugar quota.

Do as I Say, not as I Do

Rich, developed countries talk a lot about free trade, especially to lesser developed countries, but it is clear that powerful special interests can and do dominate individual markets in the rich countries of the world. The tools of partial equilibrium analysis can be used to (approximately) evaluate the results of protectionist policies.

In the case of the US sugar TRQ program, data provided by the USDA can be used to estimate the size of the deadweight loss. With a total import level of 2,000 thousand STRVs, assuming price elasticities of demand and supply of -0.25 and $+1.0$, the deadweight loss is around one billion dollars. United States consumers bear the brunt of the costs of the TRQ system, while US and foreign producers enjoy much higher profits.

But remember caveat emptor. Partial equilibrium deadweight loss analysis is a rough, back-of-the-envelope calculation. Although progress has been made in estimating deadweight loss (see the references to this chapter), consumers' surplus using demand curves makes interpersonal utility comparisons, violating one of the principles of modern utility theory.

Even more importantly, by focusing on a single market, we ignore the ramifications of the sugar quota on other goods and services. We are not counting lost output of other goods by devoting resources to producing sugar in the United States. We are also not counting health effects of sugar.

Now that you know about the US sugar quota, you can take a break and watch comedian Stephen Colbert's brief segment from 2009: tiny.cc/TRQ. Recall from Figure 17.22 that sugar prices spiked to an all-time high back then.

Exercises

1. Use the *ImportQuota* sheet to figure out what happens if all imports are banned. Explain your procedure and take screenshots as needed. Would you support a ban of all imports? Explain.

2. The deadweight loss estimates in the text are sensitive to the demand and supply curve parameters. Suppose that the inverse supply curve had a slope of 1/100 instead of 1/400. Be sure to change this parameter in both the *FreeTrade* and *ImportQuota* sheets to 1/100. What effect would this have on the TRQ system? Explain your procedure and take screenshots as needed.

3. Search the web for information about how much money US sugar producers contribute to the political campaigns of members of the US Congress. Copy and paste one sentence from a web site that you think shows the influence US sugar producers have on the US Congress. Please document your sentence with a URL and date visited citation.

References

The epigraph is from the abstract of Jerry A. Hausman and Whitney K. Newey, "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss," *Econometrica*, Vol. 63, No. 6 (November, 1995), pp. 1445–1476. www.jstor.org/stable/2171777. This paper has a nice explanation of developments in estimating deadweight loss and an example application to gasoline demand.

Economists know that using ordinary demand curves to measure *CS* and deadweight loss (the Marshallian approach) is a mistake, but some argue the error is small enough to ignore. This paper says the error matters. Pascal Lavergne, Vincent Requillart, and Michel Simioni, "Welfare Losses Due to Market Power: Hicksian Versus Marshallian Measurement," *American Journal of Agricultural Economics*, Vol. 83, No. 1 (February, 2001), pp. 157–165, www.jstor.org/stable/1244307.

When the beekeeper's bees fly into the adjoining apple orchard and pollinate the apple-grower's apple blossoms, they are conferring a positive benefit on the apple-grower that the beekeeper cannot take advantage of directly (i.e., a positive externality).

Eric S. Maskin

17.6 Externality

This section is devoted to explaining the concept of externality, why it causes a market failure, and how the inefficiency in the allocation of resources can be corrected.

The core idea is that externalities cause markets to fail—too much or too little is produced. Society's resources are inefficiently allocated. The reason why markets fail in the presence of externalities is that decision makers (consumers or firms) fail to incorporate the full costs or benefits of an action so they make a poor decision (from society's point of view).

There are three questions to answer:

1. What is an externality?

2. Why do externalities break the market?

3. How can we fix the market?

1. What is an Externality?

An *externality* is a cost or benefit not taken into account by the decision maker. An agent takes an action that impacts others, but she does not incorporate this "external impact" (hence the name externality) into her optimization problem. The decision maker considers only personal or private cost and benefit, not the full or social cost and benefit.

Externalities can arise on the cost or benefit side of an optimization problem. The private costs or benefits are the ones included in the agent's calculations. The external costs or benefits are ignored. The full or total costs or benefits are called social costs or benefits.

We can better understand externalities by looking at examples. The key is always that the optimizing agent is not considering all of the costs and benefits. Costs are imposed, but not felt by the agent or benefits are conferred on others, but not captured by the agent. This leads to a privately optimal solution that diverges from the socially optimal solution and produces a misallocation of resources.

A classic example of an externality is industrial pollution. When the cost of pollution is not taken into account by the firm, this is called a *negative production externality*. A steel firm deciding how much steel to produce factors into its choice of output level the revenue from making steel and a whole series of costs: labor, raw materials, and equipment. The costs that are counted are private costs.

If the firm pollutes the air through a smokestack, but does not have to pay for polluting the air, this is an external cost. Social costs include private costs and external costs. It is a negative externality because costs are imposed on others that are not taken into account by the decision maker. It is a production externality because the decision is made by a firm deciding how much to produce.

A college education is another classic example of a situation where the decision maker fails to consider the total picture. It is often used to explain a *positive consumption externality* because there are benefits to education that are not taken into account by the student.

The choice variable is how many years of schooling to acquire beyond high school. The costs are huge—out-of-pocket costs of a 4-year college degree include tuition and books, but opportunity costs are even greater. The benefits are also quite large, including access to better jobs, higher pay, and greater quality of life. These private benefits are considered when high school students decide whether or not to go to college so they are not part of the externality.

But society benefits from education also. College-educated people have lower unemployment rates, smoke less, and are more likely to vote. These social benefits are ignored by individuals making a decision about whether or not to acquire a college education. It is a positive externality because benefits flow to others that are not taken into account by the decision maker. It is a consumption externality because the decision is made by a consumer deciding how much to purchase.

Many studies attempt to estimate the gap between the social rate of return and private rate of return to a college degree. Social rates of return to education are several percentage points higher than the private return. This gap is an estimate of the external value generated by education.

Externalities are everywhere. Some are easy to spot, like the loud music your next door neighbor plays (a negative consumption externality). To the extent that you ignore the impact on others, your decision about which shirt to wear contains an externality.

But externalities can be subtle also. Consider an army with soldiers that were drafted into service. The externality is that the government does not take into account the full cost of acquiring its soldiers. This externality disappears with a volunteer army because the military has to pay enough to entice people to join.

Externalities are all about impacts on others so it is easy to see why they are also known as *spillover effects*. Remember, the private costs and benefits are counted by the decision maker, but the external effects are not.

2. Why Do Externalities Break the Market?

Recall Figure 17.6, reproduced below as Figure 17.25 for your convenience.

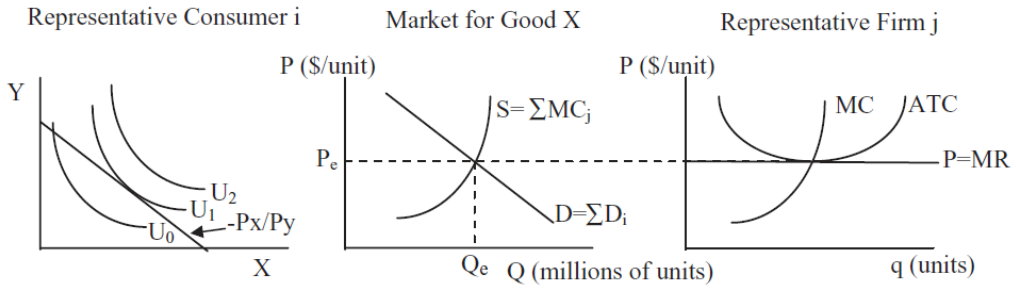


Figure 17.25: An overall view of supply and demand.

This figure has three canonical graphs: the Theory of Consumer Behavior on the left, the Theory of the Firm on the right, and supply and demand in the middle. It says that the equilibrium solution is found at the intersection of supply and demand, which come from the firm and consumer graphs.

We can show that the equilibrium quantity equals the quantity that would maximize consumers' and producers' surplus. Price controls (such as ceilings or floors), taxes, and monopoly all generate market failures, defined as quantities that do not maximize *CS* and *PS*.

We can add externalities to this list. Negative externalities are costs not taken into account and they produce too much output, while positive externalities do the reverse.

Look carefully at Figure 17.25. For the market system to yield a socially desirable outcome, supply and demand must reflect the full costs and benefits of the product. But this is precisely what is not happening if an externality is present. There are positive or negative spillover effects that result in a market equilibrium that is sub-optimal.

Suppose we have a situation where producers do not take into account the costs of pollution created as a by-product of manufacturing. Then the *MC* curve in Figure 17.25 is not incorporating the full costs of production and the supply (which is the sum of individual *MC* curves) is also too low.

There is a *marginal social cost*, *MSC*, curve that does include all costs and it does yield the socially optimal solution. Figure 17.26 shows the canonical graph of a negative externality in production. It is easy to see that the *marginal private cost*, *MPC*, which firms use to decide how much to produce to maximize profits, is too low. This produces an equilibrium output that is too high.

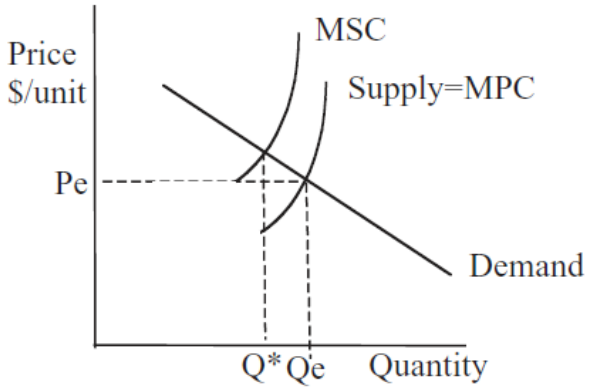


Figure 17.26: A broken market with a negative production externality.

Q^* in Figure 17.26 shows the optimal output for society. The socially optimal level of output is based on the full, social cost of production. Q_e shows the (broken) market's output. The market's equilibrium output is based only on the private cost of production so it is too high.

To sum up, a negative production externality means that firms fail to include all costs and, therefore, $MPC < MSC$, and, therefore, $Q_e < Q^*$. This is why externalities cause market failure.

We can use Excel to create a simple spreadsheet that demonstrates the concepts of externality and market failure.

STEP Open the *Externality.xls* workbook, read the *Intro* sheet, then proceed to the *Externalities* sheet.

Let's take a quick tour of the screen. On the left are the total and marginal graphs for a single firm. We ignore the average cost curves (ATC and AVC) because we are not interested in this firm's profit position. All we care about is how much it will produce. The cost function is a simple quadratic and the market price is \$40/unit so the revenue function is $40q$.

On the right is the conventional supply and demand graph. Notice that the y axes of the individual and market graphs are the same. The x axes, however, are different. There are 1,000 firms and, combined, they produce tens of thousands of units of output.

Initially, this firm is producing 10 units of output. What would you advise this firm to do? Why?

STEP Use the firm's scroll bar control to adjust its output level.

To maximize profits, this firm will choose output where $MR = MC$. This output level will generate the maximum difference between the total revenue and total cost curves in the top graph.

The problem is easily solved via analytical methods.

$$\max_q \pi = 40q - (200 + q^2)$$

$$\frac{d\pi}{dq} = 40 - 2q = 0 \rightarrow q^* = 20$$

Both analytically and with Excel, we can see that the firm will produce 20 units when equilibrium price in the market is \$40/unit. When all 1,000 firms do this we get a market equilibrium output of 20,000 units. This is the socially optimal allocation of resources to this product.

STEP To implement the externality, slide the *Set Externality* control all the way to the right (so the red lines and curve are above the black ones in the three graphs).

The red objects are not labeled. What do they represent?

STEP Insert text boxes to label the red curve in the top graph, the red line in the bottom graph, and the red line above the supply curve.

The correct labels must include the word *social*. The red line in the top graph is the *total social cost*, *TSC*, and its marginal counterpart is the *marginal social cost*, *MSC*. The divergence between the red social cost and the black private cost signals the presence of an externality. The distance between the curves are costs not taken into account by the firm.

In the market graph, the red line is *MSC*, by which we mean the sum of the indivdual marginal social costs. Like in the individual graph, divergence between supply and MSC is a clear marker of the presence of an externality.

Note that neither the firm's profit-maximizing output level nor the market's equilibrium solution changes in the presence of the externality. We have imposed an added cost, yet the firms and market do not respond because the cost is ignored.

The dashed line from the intersection of *MSC* and demand is the socially optimal level of output. An omniscient, omnipotent social planner, OOSP, would incorporate the full costs of production in determining the optimal solution to society's resource allocation problem. OOSP would choose output at the intersection of *D* and *MSC*.

We could measure the inefficiency caused by the externality by the deadweight loss. This would be the area of the triangle shown in Figure 17.27.

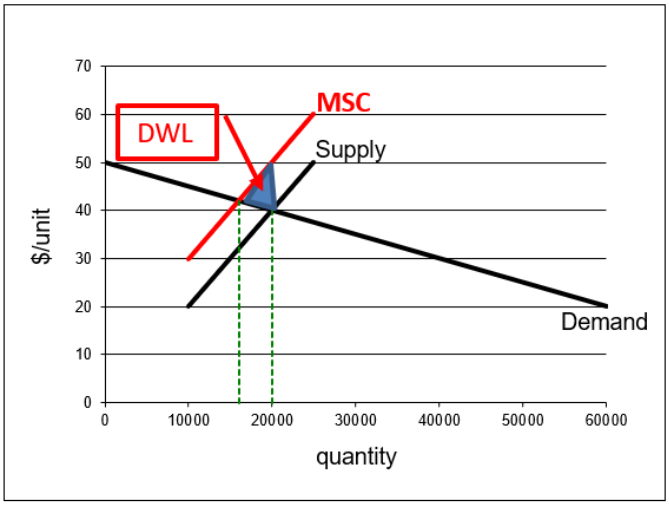


Figure 17.27: Deadweight loss from a negative externality.
Source: Externality.xls!Externalities.

The market in the presence of a negative externality has produced too much output. Units beyond 16,000 have greater marginal social cost than marginal benefit (as given by the demand curve) and should not be produced. The market produces an extra 4,000 units because it ignores the external costs of production.

3. How Can We Fix the Market?

Externalities break the market because costs and benefits are not fully incorporated into the agent's optimization problem. There are two possible solutions: government regulation and more property rights.

There are several regulatory approaches the government can take to fix the market failure caused by externality. They are united by the use of authority to correct the equilibrium output level so that it equals the socially optimal output.

Perhaps the most obvious regulatory fix is a strict limit on production, for example, a quota on pollution. If firms are allowed to pollute only a certain amount, they cannot produce as much as they want.

This is known as *command and control*, a term borrowed from the military, where top down decision making is the norm.

But this approach suffers from a serious drawback. It requires massive amounts of information to set the total amount of pollution and output.

Furthermore, if everyone is forced to reduce pollution by, say 20%, this does not take advantage of the fact that some firms can reduce pollution more cheaply than others. In other words, the government not only has to determine the total amount of pollution and output, it has to tell each individual firm exactly what and how to produce.

Command and control has long been used in environmental regulation. In the case of pollution, the Environmental Protection Agency (EPA) still uses effluent restrictions, but the EPA has moved toward other regulatory strategies.

Another government focused approach to fixing a market failure brought on by externality allows firms to decide how much to produce, but uses taxes and subsidies to incentivize decision makers to choose the socially optimal outcome.

This is based on the work of Arthur C. Pigou (rhymes with zoo, 1877 - 1959). He was a student of Marshall's and in 1908 he was appointed to Marshall's chair in economics at the University of Cambridge. Pigou argued that whenever private and social costs or benefits diverged, the government could offer incentives to align individual optimal solutions with socially optimal levels of output. Thus, today we call this solution a *Pigovian tax or subsidy.*

By imposing a Pigovian tax on polluting firms, producers are forced to consider the full costs of production in a roundabout way—the tax takes the place of the external cost.

The Pigovian tax shifts the supply curve up so that, if properly calibrated, the amount of the tax reflects the external cost not taken into account. Figure 21.28 shows how a Pigovian tax fixes the market failure caused by externality. Notice that the *Supply + Tax* curve equals the *MSC*. This enables the market equilibrium solution to equal the socially optimal solution.

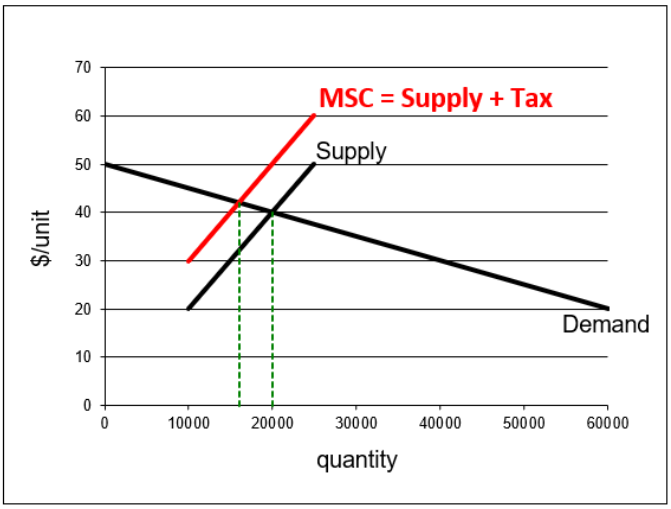


Figure 17.28: Pigovian tax correcting the inefficiency from an externality. *Source: Externality.xls!Externalities.*

The Excel workbook *Externality.xls* enables you to correct the externality with a Pigovian tax.

STEP With an externality in place, click the Set Pigovian Tax/Subsidy scroll bar to fix the inefficiency.

With every click, the market supply curve shifts up because you are imposing additional tax. A Pigovian tax works like a regular tax—it shifts the supply curve up. Obviously, you want to set the tax so that the black supply curve is coincident with and covers the red *MSC* curve.

The Pigovian tax fixes the inefficiency caused by the negative externality when the amount of the tax takes the place of the divergence between marginal social and private cost. You know you have the right tax when the market's equilibrium output equals the socially optimal level of output at 16,000 units.

Unlike regular taxes, which are applied to generate revenue for the government and cause the equilibrium quantity to be less than the optimal quantity, Pigovian taxes are actually applied to correct a market failure. They do generate revenue, but the primary purpose of a Pigovian tax is to change the market's equilibrium output to allocate resources optimally.

Pigou's approach dominated economics for many years. Then, in 1960, Ronald Coase (1910 - 2013), who spent most of his long career at the University of Chicago, offered an ingenious alternative: Define property rights over all resources (such as clean air) to internalize the externality. It took some time, but Coase's approach caught on and would win him the 1991 Nobel award in economics.

In essence, Coase cures the "market failure" by creating more markets. Market failure is in quotation marks because the argument is that it is *not* a market failure since we do not have complete property rights over all resources. A little intellectual history will help clear this up.

Frank Knight (at the University of Chicago) disagreed with Pigou in an article way back in 1924. Pigou used too much traffic as an example of a market failure in his influential book, *The Economics of Welfare*, in 1920. On page 194, Pigou explained that individual drivers would fail to take into account the additional congestion they caused when deciding whether to take one road versus another. Thus, the drivers would distribute themselves inefficiently. He pointed out that the government could impose a toll, a tax to use the road, to fix this market failure (Pigou used the phrase *laissez-faire* and it would not be until the 1950s that "market failure" was coined).

In his 1924 paper, Knight replied that, far from this being a market failure, the problem created by the externality was that there was a missing market! He said Pigou's logic was error free. It is true that drivers following their own self-interest would produce too much congestion. It is true that this decentralized system failed and a corrective tax would fix it. But, said Knight, while decentralized, this is not a market system because nobody owns the roads. Not all decentralized systems are automatically market systems.

Knight maintained that you cannot blame the market system for a lack of property rights. In Knight's view, a properly functioning market system would force firms to pay for all of the resources used. A negative externality meant that firms would treat some resources as free and it is no surprise that they would overuse those resources.

Pigou removed the traffic congestion externality example from the next edition of his book. He left, however, the overall framework of corrective taxes and subsidies intact and it became part of the paradigm of economics. For decades, students learned that corrective taxes and subsidies could and should be used to fix inefficient levels of equilibrium output.

In 1960, Coase wrote his most famous article (and perhaps the most often cited article in the history of economics), "The Problem of Social Cost." He explained how more property rights would enable markets to cure externalities. For a negative spillover like pollution, instead of command and control or a government tax, Coase advocated establishing property rights to clean air and letting the market work its magic. Firms would no longer treat the air as free if they had to pay to use it.

There is no Excel implementation of Coase's solution. The idea is simply that unpriced resources be priced. This happens when unowned resources are assigned owners. This creates a market, buyers and sellers, for the resource. This directly internalizes the externality.

Coase has said that the property rights solution was influenced by Knight. They were colleagues at the University of Chicago for many years. Knight is known as the father of Chicago School economics and an impact on the work of many social scientists at Chicago and around the world.

A theorem bears Coase's name and a brief explanation of its content is in order. The *Coase Theorem* arises out of the idea that more finely delineated property rights enable the market to solve the problem of externality. The word *theorem* is loosely used here and Coase never claimed to have found or in any sense proved the Coase Theorem.

Coase showed that by settling property rights disputes, courts played a key role in enabling markets to work. Before the court ruled, trade would be impossible because there was disagreement over ownership. These high transactions costs would prevent negotiation.

Once the court ruled, there would be a clear potential buyer and seller. Coase argued that it was not important who won the case because the resource would end up with whoever valued it more. By giving one party the property right, the court established ownership and enabled the resource to be traded. If the winner valued the resource more, the loser would be unwilling to buy it. If the winner valued it less, the loser would buy the resource. Either way, said Coase, once the judge ruled, the resource would end up at its most highly valued use. This idea is now known as the Coase Theorem.

So, in the case of pollution, perhaps homeowners would sue the polluting firm. The court would rule and, either way, once the property right was es-

tablished, the market would begin to function. Assuming the polluting firm values the property more, it will buy the right to pollute if it loses and will not sell the right if it wins the court case. Either way, it incorporates the cost of pollution because it has to purchase the right if it loses or it recognizes the opportunity cost of having the asset if it wins.

Coase criticized Pigovian taxes and subsidies as a way to fix inefficiency in the allocation of resources by a market system. Coase saw Pigou's approach as hopelessly idealistic and impossible to implement in the real world. It is easy to draw Figure 17.27 and a snap to show that the correct tax or subsidy enables the market to hit the socially optimal output as in Figure 17.28.

Unfortunately, this blackboard economics (as Coase derisively called it) is easy to draw and teach, but almost impossible to implement. The government regulator will know neither the demand nor the supply functions, and changes over time imply constant tweaking of optimal taxes or subsidies.

Economists think of Coase and Pigou as locking horns and often cast the issue as free market versus regulation. It is clear, however, that Coase and Pigou share some common ground. They both seek to maximize the value of output; they want to optimally allocate resources.

Both offer solutions that work well in theory, but can prove difficult to implement. Once we recognize that neither approach is perfect, we can begin the difficult task of deciding which approach is better in a particular situation.

The EPA and Acid Rain

Although Pigovian taxes remain a staple of economics, in recent years, market-based strategies relying on Coase's logic have gained popularity.

For example, *cap and trade* works by creating a total amount of allowable pollution and creating a market where firms can buy and sell rights to pollute. This forces firms to take into account the full costs of their production decisions. They must buy a permit in order to pollute and this forces them to internalize the externality.

The EPA's sulfur dioxide (SO_2) cap and trade program is aimed at decreasing pollutants that cause acid rain, www.epa.gov/airmarkets/allowance-markets. Instead of command and control or taxes, the EPA sets a total emissions con-

straint, or bubble, then allows firms in the bubble to buy and sell pollution permits. This scheme is equivalent to setting up a market for pollution.

There are many details to be worked out when setting up a market. For example, the government can give each firm an initial allocation of permits or they can auction off the permits.

Some environmentalists remain strongly opposed to market-based solutions to pollution abatement. They see such programs as "licenses to pollute." But the market's ability to price resources correctly and enable socially optimal resource allocation is a powerful factor in favor of the market.

Other countries (including such different places as Europe, Costa Rica, and China) have started emissions trading programs. The idea of creating a market for pollution to correct the market failure caused by externality is most definitely a real, practical solution that continues to grow in popularity.

Externalities, Market Failure, and Corrective Action

Externalities are costs or benefits not taken into account by the decision maker. Externalities cause inefficiency because the equilibrium level of output does not equal the socially optimal level of output. As usual, we can measure the inefficiency in the allocation of resources caused by an externality by computing the deadweight loss.

The inefficiency caused by externality can be corrected by command and control, but this approach requires micromanagement by government regulators. Pigovian taxes and subsidies are a type of government regulation that allows individual agents to decide what to do. A firm, for example, would decide how much to pollute and produce, but they have to pay tax. The Pigovian tax is optimized to push the market to the socially optimal output.

The Pigovian approach is definitely at play in the area of education. Truancy laws and other absolute requirements concerning schooling are an example of command and control. Government support of higher education through student grant and loan programs are Pigovian subsidies. The idea is to help students capture the full benefit of a college education and ensure that private decision making is socially optimal.

Another repair relies on market-based solutions to the inefficiency created by externality. Instead of taxing or subsidizing buyers or sellers, property rights

for unowned and unpriced resources are established and then the market is left to work its magic. Cap and trade is an example of this approach.

Coase is credited with the idea of fixing inefficient market outcomes with property rights, but Knight definitely had an influence. Knight's criticism of Pigou's toll road example is long forgotten, but it contained the seed of the logical argument that a Pigovian market failure is no such thing because not all decentralized systems are market systems.

Mechanism design is a new subfield in economics where we consciously design a game and then let agents play to reach a desired result. This is totally different than the evolution of the market system. Adam Smith did not draw up a blueprint for a market-based society. It happened organically. But now that we know how it works, we are trying to design institutions that give desirable results.

Exercises

1. Give an example of a positive externality in consumption.
2. Analyze the welfare effects of a positive externality in consumption. Use Word's Drawing Tools to support your answer with a demand and supply graph.
3. In each case that follows, describe the regulatory strategy to correct the market failure caused by a positive externality in consumption.
   (a) Command and control
   (b) Pigou
   (c) Coase

References

The epigraph is from the first page of Eric S. Maskin, "The Invisible Hand and Externalities," *The American Economic Review*, Vol. 84, No. 2, Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association (May, 1994), pp. 333–337, www.jstor.org/stable/2117854.

"The Problem of Social Cost" was published by Ronald Coase in the fledgling journal of a new field, *Journal of Law and Economics*, Vol. 3 (October, 1960), pp. 1–44, www.jstor.org/stable/724810.

The transcript of a conversation about Coase's article and the Chicago School in general is in Edmund W. Kitch, "The Fire of Truth: A Remembrance of Law and Economics at Chicago, 1932–1970," *Journal of Law and Economics*, Vol. 26, No. 1 (April, 1983), pp. 163–234, www.jstor.org/stable/725189. George Stigler describes the presentation by Coase to a "collection of superb theorists" as "one of the most exciting intellectual experiences of my life:"

> My recollection is that Ronald didn't persuade us. But he refused to yield to all our erroneous arguments. Milton would hit him from one side, then from another, then from another. Then to our horror, Milton missed him and hit us. At the end of that evening the vote had changed. There were twenty-one votes for Ronald and no votes for Pigou. (p. 221)

That was no typo on Coase's lifespan in the text, 1910 - 2013. And he worked to the end. In 2009, at the age of 99, he published a book, with Ning Wang, How China Became Capitalist.

Arthur Cecil Pigou's *The Economics of Welfare* was first published in 1920 and is freely available online at the Library of Economics and Liberty, www.econlib.org/library/NPDBooks/Pigou/pgEW.html.

Frank H. Knight's criticism of Pigou's traffic congestion example is in "Some Fallacies in the Interpretation of Social Cost," *The Quarterly Journal of Economics*, Vol. 38, No. 4 (August, 1924), pp. 582–606, www.jstor.org/stable/1884592.

> Lack of legal sanctions means that loyal members of the cartel must exact penalties against deviants in the market place. Unless such disciplinary actions (mainly price cuts) can be localized, every member of the cartel, loyalist and defector alike, suffers. That is a very severe (if little remarked) limitation on the efficiency of cartels.
>
> Oliver E. Williamson

17.7 Cartels and Deadweight Loss

We know that the equilibrium output of a competitive market equals the output that maximizes consumers' and producers' surplus. We also know that monopoly produces too little output and the resulting deadweight loss is a measure of the inefficiency of monopoly. But competition and monopoly mark opposite ends of a spectrum that includes a wide range of other market structures.

A cartel is a type of market structure in which a group of firms cooperate to control output and price. Perhaps the most famous international cartel is the *Organization of the Petroleum Exporting Countries*, OPEC. Cartels are not monopolies because there are several independent firms in the syndicate or trust, but they hope to act like a monopolist, restricting output and raising price, to earn monopoly profits. Cartels are inherently unstable because it is in the interest of each member to cheat and sell more than the agreed amount.

This section explores the welfare properties of a specific type of cartel. The application is based on the workings of the Norwegian cement cartel as explained by Röller and Steen (2006). Analyzing the cartel involves solving a two-stage game and the cartel result is compared to monopoly and non-cooperative, Cournot competition. This material is advanced and it is recommended that the chapter on Game Theory be completed before proceeding.

A Brief History of Norwegian Cement

Cement output in Norway (and in other countries that use the metric system) is measured in tonnes (pronounced tons). This is not simply a foreign spelling for a ton. A ton is 2,000 pounds. A tonne, sometimes called a metric ton, is 1,000 kilograms. Given there are roughly 2.2 kilos in a pound, a tonne is about 2200 pounds. Thus, a tonne is bigger than a ton.

Figure 17.29 shows that production rose dramatically during the second half of the 1960s, greatly outpacing demand. This excess output was sold at a loss in other countries. A balance between production and consumption was restored by the early 1980s.

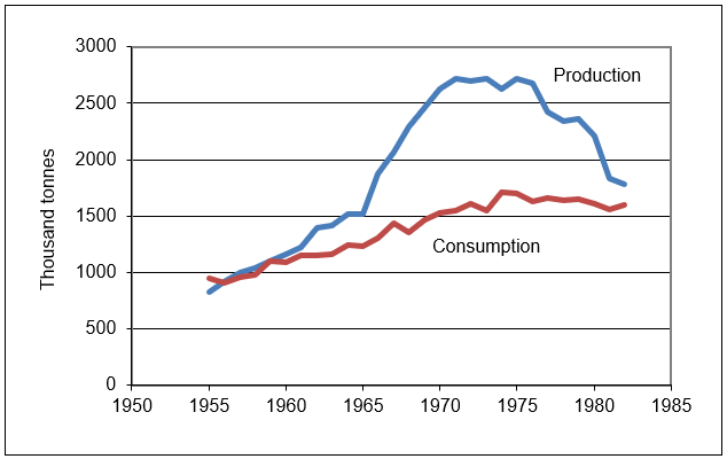


Figure 17.29: Norwegian cement production and consumption, 1955-1982. *Source: CartelDWL.xls!Data.*

Production rocketed because of the sharing rule adopted by the Norwegian cement industry. A *sharing rule* determines how the monopoly rent is to be distributed among the firms in the cartel. Each firm's share of the domestic market was based on its fraction of total industry capacity. We will see that this gives each firm an incentive to expand plant capacity and led to the explosion in output shown in Figure 17.29.

In 1968, the three producers in the cement industry abandoned the cartel market structure and merged to form a monopoly. By then, however, plant capacity had been expanded and it took years to reduce output.

Röller and Steen explain that there are few empirical studies of cartels because they are illegal in many places (including the United States) so obtaining data is difficult. Such is not the case in Norway. "Given the legality of the Norwegian cement cartel, we have a large amount of primary data allowing us to do a complete welfare analysis." (Röoller and Steen, 2006, p. 321.

Monopoly Review

STEP Open the Excel workbook *CartelDWL.xls*, read the *Intro* sheet, then go to the *Monopoly* sheet.

Given the linear inverse demand curve and constant marginal cost, finding the monopolist's profit-maximizing solution is easy.

STEP Use the scroll bar under the chart to find Q^*. As you change the quantity, you can see the corresponding price in the chart and in cell B11. You can also see the producers' surplus (also known as profits) change in cell B19 as you set Q.

You can choose Q^* by watching cell B19, but you could also find Q^* by choosing the intersection of MR and MC.

Excel's Solver offers yet another alternative to finding the profit-maximizing level of output.

STEP Run Solver and configure the Solver dialog box to solve the monopolist's profit maximization problem.

Finally, click on cells B18, B19, and B21 to show the consumers' surplus (CS), producers' surplus (PS), and deadweight loss (DWL) from the monopoly solution in the chart.

Having found the monopoly solution, we turn to output (and price) under a noncooperative, Cournot environment.

Cournot Review

STEP Proceed to the *CournotFirm* sheet.

Chapter 16 on game theory presented the material reviewed here, which assumes a basic understanding of the Cournot model and Nash equilibrium.

Instead of a single firm, there are three firms making a homogeneous product. They do not collude or combine forces. Instead, they compete. Unlike

perfect competition, however, there are so few producers that they impact each other's decision making. If one firm decides to produce a lot, this will lower the price for all three firms.

How will an individual firm decide how much output to make? The core idea is that each firm will make profit-maximizing output decisions based on conjectures about what the other firms will do. The output level at which each firm's decision is consistent with the output chosen by the other firms is the solution, called a Nash equilibrium.

The *CournotFirm* sheet opens with cell B10 set equal to zero. This means that Firm 1 is exploring what its best option is if the other firms produce nothing.

STEP Use the scroll bar under the chart to find the profit-maximizing output for the conjecture that the other firms produce nothing.

If the other firms decide to produce zero output, Firm 1 will produce 2.3 million units of output. But this is not an equilibrium solution because the other firms would not choose to produce zero units of output when this firm produced 2.3 million tonnes. How much would the other firms produce?

STEP Click the Set Exo Q button to copy Firm 1's optimal solution (in cell B15) to the conjectured output in cell B10.

Notice how the chart shows new red D and MR curves. These are the residual demand and residual marginal revenues curves for Firm 2, given that Firm 1 produces 2.3 million and Firm 3 produces nothing.

STEP Use Excel's Solver to find the profit-maximizing output for the conjecture that the other firms produce 2.3 million units.

You should find that Firm 2 will produce 1,150,000 units when the other two firms produce 2.3 million. We have stumbled upon the Nash equilibrium solution! If each firm produces 1.15 million units, then none of them will regret its output decision. In other words, each firm's optimizing decision (1.15 million) is consistent with the conjectured output (2.3 million).

Notice that the Nash equilibrium is not Firm 1 = 2.3 million, Firm 2 = 1.15 million, and Firm 3 = 0. Both Firms 1 and 2 would regret their decisions

and would opt for different output choices. It should be clear, however, that if each one makes 1.15 million, then none of the firms would regret or wish to change its chosen output level.

The Cournot solution can be found via iteration (which was easy in this example) or by analytical methods (see work starting in cell A28). The reduced form for the industry's Nash equilibrium output in this Cournot model (linear demand and cost function and n firms) is:

$$Q_e = \frac{n}{n+1}\frac{d_0 - MC}{d_1}$$

Price, of course, is simply read from the inverse demand curve.

STEP Proceed to the *Cournot* sheet to see the welfare implications of the Cournot solution. Click on cell B14 to see that the formula for the Nash equilibrium has been entered.

Notice that the Cournot output level is between the perfectly competitive ($D = MC$) and monopoly ($MR = MC$) output levels.

STEP Click on cells B18, B19, and B21 to highlight *CS*, *PS*, and *DWL* in the chart.

Once again, notice that the *DWL* for the Cournot solution is between the monopoly (highest *DWL*) and perfect competition with many firms (no *DWL*) extremes.

STEP Increase the number of firms in cell B10 to 5, 10, and 20.

As n rises, *DWL* falls because as n rises, we are approaching the ideal solution of competition with many firms. Thus, perfect competition is simply an n-firm Cournot model with an infinite number of firms. You can confirm that at $n = 1$, the monopoly solution is found.

Having covered the monopoly and competitive Cournot models, you are ready to tackle yet another market structure: the cartel.

Cartel Behavior

Suppose an industry, made up of several firms, organized into a *cartel*. In other words, the firms would join forces and cooperate in making decisions.

The cartel would decide the total domestic output and price for the product. In addition, the cartel would have to determine how much each firm would produce. This is called the sharing rule.

Different sharing rules yield different results. Suppose that the sharing rule applied is that each firm's output reflects its share of total industry capacity. There are no limits on each firm's capacity and any output not sold domestically could be exported at the world price.

Although each firm chooses capacity first and then the cartel chooses total output (and price), we solve the two-step optimization problem recursively. This means we start at the second stage, then work backwards to the first stage.

Stage 2: Choosing Total Domestic Output (and Price)

STEP Proceed to the *CartelStage2* sheet.

The information is laid out as in the *Monopoly* sheet, but there are additional variables. The world price (below marginal cost) has been added in cell F8 and to the chart. Individual firm parameters start in row 26. The three firms have chosen their capacities (cells B30:B32), determining total capacity (B28) and shares of domestic output (C30:C32).

STEP Use the scroll bar under the chart to explore different quantities of domestic output. This is the cartel's key choice variable. It can choose anywhere from no output to the vertical, total capacity, line (which is determined by the firm's capacity decisions in stage 1 and is now an exogenous variable to the cartel).

STEP Click on cell B19, which is the *PS* and also the profit generated by a given output level, to highlight the *PS* in the chart. The formula and the chart reveal that *PS* has two parts: = (P − s0_)*Q − (s0_ − R_)*(B28 − Q).

The first part is a rectangle with height from *MC* to price and width from zero to the chosen output. This would be *PS* under monopoly.

But the cartel has a second component to *PS*. This is the smaller rectangle on the chart and it is subtracted from the bigger rectangle. This second part is the excess output that is exported and sold at the world price. It

is subtracted from profits because the world price is below MC. Thus, these units are sold at a loss.

STEP Use the scroll bar to find the cartel's Q^*. Notice that you can find the optimal output by keeping an eye on PS (in cell B19) or by setting $MR = R$. You can also use Excel's Solver to find the optimal output.

Cell B13 shows the optimal output and your cell B12 should equal this solution. The cartel will produce 3,150,000 units and charge $1,725 per unit. This is a higher output (and lower price) than the monopoly solution.

R is a key variable. It plays the role of MC in the cartel's optimization problem. What effect does changing R have on Q^* and P^*? What welfare effect does changing R have? We can answer these questions with Excel.

STEP Change R to 500 in cell F8. Solve the cartel's optimization problem again.

You should see that optimal domestic quantity is lower and price is higher.

STEP With the new optimal solution for $R = 500$ in B12 ($Q^* = 2.8$ million), click the Compare Surplus button. It displays the initial CS, PS, and DWL values (for $R = 150$) and computes the difference between the new and initial values.

As R rises, CS falls and PS rises. Total DWL is bigger by $136 million, with both parts of DWL (the traditional triangle that represents domestic DWL and the export loss) rising.

STEP Click the Reset button (or reset R to 150).

We conclude our analysis of the cartel's first stage of the optimization problem by examining the effect on the individual firms. Cells D30:G32 show how the sharing rule is applied to determine how much each firm produces, given the cartel's total domestic output decision. The blue text color means these variables are endogenous—they are determined by the cartel's domestic output decision.

STEP Adjust Q via the scroll bar under the chart and keep your eye on cells D30:G32. As Q changes, so do the individual firm variables in blue.

Because the firms have equal capacities, each sells a third of the domestic output and exports the rest. Domestic and export sales for each firm are displayed.

STEP Enter 3,150,000 in cell B12 (the value of Q^* at the initial values of the exogenous variables) to see the PS earned by each firm at the cartel's optimal output.

From the cartel's point of view, the individual firm capacities are given. But would profit-maximizing firms choose these particular capacities? This question is at the heart of the first stage of the cartel's two-stage optimization problem.

Stage 1: Choosing Capacity

Now that we know how the cartel is going to decide how much domestic output to produce and the sharing rule, we can tackle the question facing each firm: How much capacity?

At any point in time, firms have a given maximum total production, or capacity, determined by factory size. To increase capacity, firms must expand factory size and this takes time.

Notice that the marginal cost of cement production is different from the marginal cost of capacity. The former is assumed to be low and it does not play a role in this analysis. In fact, it is assumed that firms always produce up to capacity.

The capacities of each firm and hence total capacity are given to the cartel but are chosen by each firm. Each firm would pick that capacity that would maximize its profits.

The profit function has revenue from two sources: domestically sold output at price P (chosen by the cartel) and the excess output that is exported and sold at the world price, R. The cost of capacity function is linear, with constant marginal cost.

STEP Proceed to the *CartelStage1* sheet and click on cell B11 to see that the formula reflects the firm's profit function.

Cells B19:B23 have the exogenous variables. Each firm chooses capacity (q_i) to maximize profits.

The sheet opens with the firm having a capacity level of 1,200,000 units, the same as the other two firms, so the total industry capacity is 3,600,000 units.

STEP Click on the scroll bar (next to B27) to increase capacity.

Notice that the larger the chosen capacity, the greater the share of the domestic sales (Q), which is chosen by the cartel, and thus domestic revenues (B13) rise. As capacity increases, exports also rise (because only a share of the firm's output is sold domestically) and this hurts because the world price is below marginal cost. Of course, increasing output is going to increase costs because the firm has to build a bigger plant.

Given these trade-offs, what level of capacity should this firm select?

STEP Keep your eye on cells F27:H27 as you adjust the scroll bar to select the profit-maximizing output.

As usual, you can equate *MR* to *MC* to find the optimal solution.

STEP Check your work by using Excel's Solver.

The optimal capacity, 1,342,758 units, differs from the original 1.2 million units. This means that the optimizing firm would choose to make 1,342,758 units when the other two firms make a total of 2.4 million.

STEP Copy the optimal capacity in cell B27 and paste it in cell K9 (or enter 1,342,758 units in cell K9).

We are not done yet because if this firm wants to make 1,342,758 units, it stands to reason that the other firms (with identical cost structures) will also want to do this.

STEP Return to the *CartelStage2* sheet, select cell B30, and paste (or type in) 1,342,758.

Notice that cells B31 and B32 change to the value of cell B30. Cell B28, *Total Capacity*, is now higher and, thus, the vertical line in the chart has shifted right.

We do not need to run Solver again because the cartel's optimal output and price combination in the domestic market is unaffected by the total industry capacity. The extra output is simply exported and sold at the world price.

STEP Return to the *CartelStage1* sheet and notice that *MR* no longer equals *MC*. Click on cell B20 to see that it has a formula.

Cell B20, *Other Capacity*, has changed because the other two firms have selected different capacities.

STEP Copy cell B20, select cell J10, and *Paste Values*, then run Solver. Copy the new optimal Q (cell B27) and paste it in cell K10.

Notice that we still do not have an internally consistent solution between the two optimization problems. The firm capacity optimal solution is different from the total capacity used by the cartel. We must iterate.

STEP Do these three steps:

1. In the *CartelStage2* sheet, select cell B30, and paste the value of optimal capacity.

2. Return to the *CartelStage1* sheet and copy cell B20, select cell J11, and *Paste Values*.

3. Run Solver to find the new optimal solution. Copy the optimal Q (cell B27) and paste it in cell K11.

We still do not have a situation in which the optimal capacity decision of Firm 1 agrees with the total capacity parameter used by the cartel.

STEP Fill in the Stage 1 and Stage 2 Consistency table. You will need to iterate, repeating the process of solving for Firm 1's optimal capacity, pasting that result in the CartelStage1 sheet, then returning to the CartelStage2 sheet to see if the two solutions coincide (the three steps above).

STEP When you have finished completing the table, click the Show Data button.

This reveals results in columns L, M, and N that are based on your iterations. It also shows the Nash equilibrium solution for q_i*. As with our work in the

Cournot model earlier, there is an analytical solution to each firm's optimal and consistent capacity and we entered it in cell K19.

Figure 17.30 shows what your screen should look like. The total capacity, the vertical line in the *CartelStage2* chart, is driven to an equilibrium value of 3,891,176 units. The total capacity line bounces right and left until settling down at a value that is consistent with the optimal solution to the individual firm's profit maximization problem. In equilibrium, each firm will have a capacity of 1,297,059 units. This is consistent in the sense that each firm would choose this capacity if it knew the sharing rule adopted by the cartel.

Stage 1 and Stage 2 Consistency

Iteration	Other Capacity	qi*	Starting Total Capacity	Ending Total Capacity	Difference
1	2,400,000	1,342,758	3,600,000	3,742,758	(142,758)
2	2,685,515	1,273,616	4,028,273	3,959,131	69,142
3	2,547,232	1,308,620	3,820,848	3,855,852	(35,004)
4	2,617,240	1,291,240	3,925,860	3,908,480	17,380
5	2,582,480	1,299,959	3,873,719	3,882,438	(8,719)
6	2,599,917	1,295,607	3,899,876	3,895,524	4,352
7	2,591,213	1,297,784	3,886,820	3,888,997	(2,178)
8	2,595,569	1,296,696	3,893,353	3,892,265	1,088
9	2,593,392	1,297,240	3,890,088	3,890,632	(544)

Nash eq q_i	1,297,059
Nash eq Total Q	3,891,176

Figure 17.30: Nash Equilibrium capacity.
Source: CartelDWL.xls!CartelStage1.

Given the demand curve parameters, marginal cost, and the world price, we know the cartel's profit-maximizing domestic output and price. Because we know the equilibrium solution to each firm's capacity decision, we can compute the total output produced and export loss. Thus, we can compute *CS*, *PS*, and *DWL*.

STEP Copy cell K19 from the *CartelStage1* sheet and *Paste Values* in cell B30 of the *CartelStage2* sheet.

STEP Click on cells B18, B19, and B21 to display the *CS*, *PS*,and *DWL* generated by the cartel solution.

Cartel Model Summary

Determining the cartel's output is not easy. One has to solve a two-stage game. The cartel's sharing rule means that each profit-maximizing firm is willing to trade off export losses in order to get a share of high-priced domestic output.

The vertical total capacity line in the *CartelStage2* chart is actually an equilibrium solution to the first stage of the game. There is only one value of total capacity that is internally consistent with individual firm capacity decisions.

The cartel game-theoretic model also can be solved via analytical methods. The mathematics is not easy, but if you are interested in seeing the solution, click the Show Analytical button near cell M5 of the *CartelStage1* sheet.

Having determined the output and price solutions to each of the three market structures, we are ready for the welfare analysis.

Comparing Monopoly, Cournot, and Cartel Solutions

STEP Proceed to the *Compare* sheet.

Given the parameter values (in the shaded cells), the table displays the output, price, *CS*, *PS*, and *DWL* associated with perfect competition, monopoly, cartel (with the sharing rule), and Cournot market structures.

Cells B18:B21 are connected to the market structure currently displayed on the chart. Initially, the perfectly competitive result is displayed. *DWL* will be computed against this standard.

STEP Click the *Monopoly* option.

Cell range B18:B21 is updated and the chart displays the monopoly result. Notice that the monopolist ignores the world price and does not export cement. She maximizes profits by choosing output where $MR = MC$.

Compared to perfect competition (in cells B10:B14), monopoly generates much lower *CS*, higher *PS*, and a substantial *DWL*.

STEP Click the *Cartel* option.

The chart displays the total capacity vertical line and the exports are highlighted. We can compare the cartel to the monopoly and PC results by looking at the cells in columns B, C, and D, in rows 10 to 14.

Note that for the cartel option, cell D13 shows the value of profits for the cartel. This is domestic *PS* less export loss. Cell B19, also labeled *PS*, shows domestic producers' surplus (and leaves out the export loss). This is confusing, but it allows separation of the two sources of total *DWL*, domestic *DWL*, given in B21, and export loss, shown in cell B22, and ensures that domestic *DWL* plus total surplus will sum to total surplus in the perfect competition case. Total *DWL*, the sum of domestic *DWL* and the export loss, is reported in cell D14.

Compared to perfect competition, the cartel generates lower output and higher prices, but it is better than monopoly. Cells G10:G14 show what happens when you move from cartel to monopoly.

STEP Click on cells G10 to G14 to see their formulas.

If the Norwegian cement industry merged to monopoly from a cartel, we would see the following: Output falls, price rises, *CS* falls, *PS* rises, and *DWL* rises.

The increase in *DWL* would enable to us to judge such a move as a failure in terms of resource allocation in the Norwegian economy.

STEP Click the *Cournot* option.

Comparing cartel and monopoly to perfect competition is not particularly useful, because we are not going to get a perfectly competitive cement industry. There are only three firms. If we had competition, it would be Cournot competition. The three firms would not collude, but they would behave strategically.

If the industry went from cartel to Cournot, cells F10:F14 show what would happen. As with cells G10:G14, these cells report the difference from the cartel to the Cournot market structure. Notice that output rises, price falls, *CS* rises, *PS* rises, and *DWL* falls.

Of these effects, *PS* rising is surprising, but remember that under Cournot, the export losses are eliminated.

This completes the theoretical welfare analysis. The results are clear: To maximize surplus, the Norwegians should have moved from a cartel to Cournot competition. Of the three market structures, Cournot has the lowest *DWL*.

There is, however, one important issue left unresolved: These results apply only to the parameter values on the sheet. We do not know the intercept or slope of the Norwegian demand curve for cement, nor do we know *R* or *MC*. We need to get these parameter values, and then do the analysis based on these real-world parameter values.

Welfare Analysis for 1968

STEP In the *Compare* sheet, scroll to the right of the graph and click the Show Actual button (over cell N1).

After clicking the button, a new sheet appears, populated with key parameters for 1968, the last year of the cartel.

Figure 17.31 shows the results for the various market structures for the estimated demand curve for 1968. The conclusion is clear—Cournot is the best of the three feasible market structures. It produces the highest output, lowest price, highest *CS*, and lowest *DWL*.

Exogenous Variables for Demand and Marginal Cost						
P = d0 - d1Qd			MC = s0		World Price	
d0	953.7203057		s0	288.6596342	(R)	
d1	0.000252144				235.0740945	
	Perfect Competition	Monopoly	Cartel	Cournot	Cartel to Cournot	Cartel to Monopoly
Q (domestic)	2,637,622	1,318,811	1,425,071	1,978,217	553,146	(106,260)
P	kr 288.66	kr 621.19	kr 594.40	kr 454.92	(kr 139.47)	kr 26.79
CS (millions)	kr 877	kr 219	kr 256	kr 493	kr 237.332	(kr 36.758)
PS (millions)	kr 0	kr 439	kr 391	kr 329	(kr 61.745)	kr 47.891
DWL (millions)	kr 0	kr 219	kr 230	kr 55	(kr 175.587)	(kr 11.133)

Figure 17.31: Welfare analysis.
Source: CartelDWL.xls!CompareActual.

Figure 17.31 also makes clear why the industry went to monopoly instead of Cournot after the cartel collapsed (under the weight of overproduction

and export losses). PS would rise when moving from Cartel to monopoly (by 47,891,000 kroner), but fall (by 61,745,000 kroner) if the industry had adopted a noncooperative Cournot arrangement. Thus, it is clear that the cement industry chose to maximize its own PS instead of $CS + PS$. This is not surprising.

In fact, Röller and Steen build an even stronger case by exploring the welfare effects over several years. Scroll to column AE and read the text box if you are interested.

STEP Click the *Monopoly* option to display the monopoly solution in the graph.

The monopolist would choose output where $MR = MC$ and charge the highest price possible for that level of output. Monopoly profit in 1968 would have been 439 million kroner. Consumer surplus would be much smaller than under perfect competition and Norway would suffer a deadweight loss from monopoly of 219 million kroner.

But the Norwegians did not have a monopoly before 1968, they had the cement cartel.

STEP Click the *Cartel* option.

The cartel chooses output where $MR = R$, allocates the domestic output to the three firms based on capacity shares, and exports the excess output.

Notice, however, that Röller and Steen do not use the predicted capacity based on the demand curve parameters. Instead, they use actual exports. The story here is that capacity takes time to build. The cartel puts persistent pressure on expansion, but the firms do not actually reach their goal of vast capacity because the cartel collapses.

STEP You can check the theoretical cartel solution for the estimated parameters by simply copying the range A5:F8 from the *CompareActual* sheet and pasting in the same range in the *Compare* sheet. Click Yes if prompted to replace the destination cells. You may need to click the *Cartel* option to refresh the screen.

Figure 17.32 shows the result. Capacity is huge and export losses are staggering. This is the capacity that would have been installed in the long run

under the cartel. Röller and Steen do not use this capacity value. Instead, they use actual exports, based on the actual capacity in 1968.

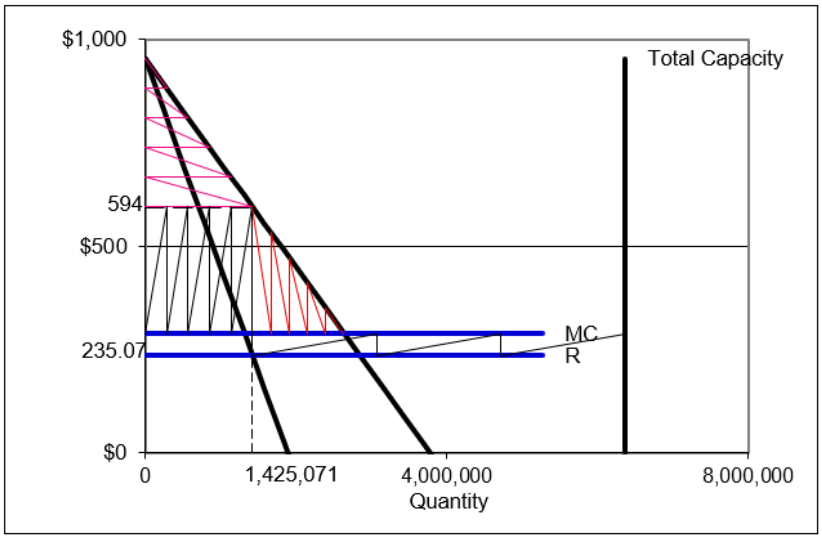


Figure 17.32: Cartel results with capacity determined theoretically. *Source: CartelDWL.xls!Compare* using estimated parameters for 1968.

STEP Return to the *CompareActual* sheet. Focus on columns F and G.

We know the firms merged to monopoly and the cartel to monopoly column (G) shows the welfare implications of this move for just 1968. As expected, output falls and price rises, *CS* falls and *PS* rises. The net welfare effect can be computed as the sum of the changes in *CS* and *PS*, which is an 11 million kroner increase (in cell G15).

Alternatively, the net welfare effect can be determined by looking at the reduction in *DWL* in cell G14. Because *DWL* falls as we move from cartel to monopoly, this number is negative. But notice that the absolute values are the same.

Our standard models tell us that merger to monopoly is the worst possible outcome—monopoly generates the greatest *DWL* of any market structure. However, because of the sharing rule, welfare actually increases when the cartel merges to monopoly because monopoly does not suffer export loss.

STEP Compare the values in Table 3 for the cartel to monopoly in 1968 to the values in column G.

The slight differences are due to rounding and precision differences.

Although monopoly beats the cartel, this is a poor argument for supporting monopolization. After all, the cartel could have dissolved into a noncooperative, Cournot competition. We must examine the welfare effects of this move and compare it to moving to monopoly to find the better option.

STEP Compare the red circled value of the change in *PS* when moving from *Cartel* to *Cournot* in Table 3 to cell F13. These numbers should be the same, but they are not.

Röller and Steen made a mistake in computing the net welfare effect for the move from cartel to Cournot in Table 3. They report the change in domestic *PS* in the table, not the change in total *PS*, which includes the export loss. As a result, the net welfare effect for cartel to Cournot in Table 3 is also incorrect. By failing to include the export loss in the reported *PS*, they underestimated the welfare gain from adopting a Cournot noncooperative market structure.

This error does not change Röller and Steen's conclusion. In fact, if anything, their results are strengthened once the export loss is accounted for. The loss in *PS* that the cement industry undergoes in moving to Cournot competition is not as bad as Table 3 suggests because of the elimination of the export loss. The true net change in welfare is some 45 million kroner higher than Table 3 estimates.

Consequences of Using Actual Versus Theoretical Total Capacity

Now that we understand how net welfare effects for 1968 are computed, we turn to the issue of how the export loss is measured.

Cell D20, the export loss in 1968, is based on actual exports—the difference between actual capacity (total production) and domestic output.

Figure 17.32 and your *Compare* sheet show that at the Nash equilibrium, long run capacity is much higher than the actual capacity (based on actual total production). How does this impact the analysis? This is an important question with a surprising answer.

STEP Compare the formulas in cells G16 and G17. Both display the same number, but the formulas are different.

G16 computes the net welfare gain from going to Cournot instead of monopoly (from the cartel, of course) by taking *DWL* from cartel to Cournot minus the *DWL* from cartel to monopoly. Cournot beats monopoly by about 165 million kroner.

G17 computes the same net welfare gain, but does so by subtracting the net welfare effect from going to monopoly from the net welfare effect from going to Cournot. Once again, the move to Cournot beats the move to monopoly by roughly 165 million kroner.

STEP Copy the two cells, G16:G17, and go to the *Compare* sheet, pasting these cells in the same range.

The result is surprising—the superiority of Cournot over the cartel remains exactly the same, even though the *Compare* sheet is using theoretical, long run total capacity and the export losses are huge.

If you compare the values in columns F and G in both sheets, you will find that for both the move to monopoly and the move to Cournot, the change in PS and the change in net welfare are much higher if the theoretical capacity is used. This makes sense because the export loss is much greater.

However, the relative improvement in Cournot over monopoly remains the same because both Cournot and monopoly avoid export losses. Thus, the size of the export loss does not matter.

Had Röller and Steen used the theoretical, long run total capacity level based on the estimated parameters in 1968, their qualitative and quantitative conclusion regarding the superiority of Cournot over monopoly would remain completely unaffected.

Lessons from the Norwegian Cement Cartel

Röller and Steen (2006) evaluate the effectiveness of the (legal) cement cartel in Norway over the period 1955 to 1968. They solve monopoly, Cournot, and cartel models and compare the results. They find that because of the sharing rule adopted by the cartel, consumers actually did better (in terms of con-

sumer surplus) than they would have if the industry had been monopolized. Producers, on the other hand, lose in the domestic market with the cartel compared to a monopoly. Producers suffer an additional export loss under the cartel and this leads to a key result: The merger to monopoly that occurred in 1968 actually improved net welfare relative to the cartel outcome. This is certainly a surprise, given that we expect monopoly to be the worst market structure. The authors point out, however, that simply breaking up the cartel and allowing Cournot competition would have improved welfare even more.

The fact that Röller and Steen used actual exports instead of estimated exports makes no difference to their final conclusion that Cournot competition would have been the first-best choice. The reason it does not matter is that both monopoly and Cournot competition result in the elimination of the export loss, so in comparing a move to either Cournot competition or monopoly, the actual size of the export loss is irrelevant.

Röller and Steen (2006) give an excellent example of how economists use *CS*, *PS*, and *DWL* in policy analysis. It also enables deeper understanding of game theory by examining the two-stage game played by members of the cartel.

This section is certainly not typical of an Intermediate Micro course, but it offers advanced students a chance to see a sophisticated application of welfare analysis.

Exercises

Suppose the inverse demand curve is $P = 1000 - 0.5Q$, marginal cost is constant at \$100 per unit, and the world price is \$50. Enter these parameter values in the *Compare* sheet and answer the questions below. Enter the demand slope as a positive number, 0.5, and click one of the market structure options to refresh the chart.

The math theory prep section showed two surprising results. First, consumers' and producers' surplus under the cement cartel do not depend on the marginal cost of capacity. Second, as the number of firms in the cartel rises, the likelihood a merger to monopoly will be welfare enhancing rises.

To answer the questions that follow, taking pictures is helpful. You can select cells (e.g., A1:M25) and copy as a picture, then paste.

1. Increase *MC* from 100 to 200 and determine the impact on the cartel's *Q*, *P*, *CS*, *PS*, *DWL*, and export loss. What happens to each of these variables as *MC* rises?

   Be sure to click the *Perfect Competition* and then *Cartel* option button to refresh the data below the buttons.

2. Which changes, if any, in the variables are surprising? Why?

3. At what value of *MC* will there be no exports? Take a picture of this situation and paste it in your Word document.

4. Increase the number of firms from 3 to 5 (with *MC* at the no export loss value). What effect does this have on the cartel's *Q, P, CS, PS, DWL*, and export loss?

5. What can you conclude about the effect of the number of firms on *PS* from a merger to monopoly (from the cartel)?

References

The epigraph is from page 278 of Oliver E. Williamson, *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* (1985). Williamson applies the standard tools of economic reasoning (optimization and comparative statics) to transactions and argues that the institutions we observe are the evolutionary product of selection based on optimization.

This application is based entirely on the excellent paper by Lars-Hendrik Röller and Frode Steen, "On the Workings of a Cartel: Evidence from the Norwegian Cement Industry," *American Economic Review*, Vol. 96, No. 1 (January, 2006), pp. 321–338, www.jstor.org/stable/30034368. Special thanks to Frode Steen for making additional data available.

> The presence of people in the market who are willing to offer inferior goods tends to drive the market out of existence—as in the case of our automobile "lemons." It is this possibility that represents the major costs of dishonesty—for dishonest dealings tend to drive honest dealings out of the market.
>
> George A. Akerlof

17.8 Signaling Theory

We all want to live in a world in which every buyer and seller is always completely honest, dependable, and trustworthy. In such a world, no one would lie, cheat, or steal. No one would misrepresent a product or hide a defect to make a sale, and the buyer would always alert the cashier when receiving too much change. Even politicians and children would always tell the truth.

Plainly, we do not live in such a world. Cigarette manufacturers swear under oath that their products are safe and that there is no proof that tobacco causes lung cancer. Management lies to labor about the true profitability of the firm and the size of the wage increase that the firm can really afford. It seems that we live in the midst of lies and deceit. Few can be trusted and few trust us.

This then is the problem: How can we make our world—the one full of distrust and scams—more like the world we all agree is better—the one in which individuals are sincere and open? How can we get people to tell the truth?

Three Ways to Handle Dishonesty

We review utopian and authoritarian solutions to fighting dishonesty, and then focus on a third way that most people rarely consider.

If somehow it were possible to create a perfectly honest person, we could attain our goal of living in an honest world. People could be counted on, with no doubt or reservation whatsoever, to be completely clear and forthright. This is the *utopian solution.*

Karl Marx believed private property, money, and the capitalist system created an all-encompassing greed that generated fraud, deception, and a variety of other reprehensible individual behaviors. For Marx, the solution to the

problem was quite simple: Replace vicious capitalism with its superior evolutionary offspring, communism, and replace the money-hungry *homo economicus* with the noble *new socialist man*.

Although seemingly hopelessly idealistic, in certain cases, reliance on people's good qualities is, in fact, possible. We all have close friends and family whom we can trust to be sincere and truthful. In our daily lives, however, we deal with countless strangers, and we cannot rely on personal relationships to ensure honest behavior. In a modern society that incorporates the actions and decisions of millions of individuals, it is simply impractical to expect trustworthiness from everyone.

To protect against dishonesty, many people think immediately of monitoring. This second approach can be called the *authoritarian solution*.

If a store owner thinks customers are going to steal, valuable merchandise can be put under a glass counter, security cameras installed, and guards can watch the customers. If the government knows that citizens will cheat on their taxes, a sample of tax returns will be audited carefully to check for full compliance and severe penalties will be imposed on those caught cheating.

In general, the authoritarian approach to solving the problem of dishonesty requires a powerful judge who can check the truthfulness of statements and punish those who are caught violating the rules. Monitoring and punishment can work well when it is clear what constitutes a lie, and it is easy to observe the dishonest behavior.

Unfortunately, in many cases, it is quite difficult to determine dishonest behavior because there are shades of deceitfulness, ambiguities in truthfulness, and inherent uncertainty in the world. For example, if I sell you an expensive product, promising that it is of high quality, and then it breaks, am I a liar? It may very well be a high-quality good that just happened to break. Of course, I may have known that it was really shoddy merchandise and I just tricked you. How can you know which case is true?

In addition to that rather large subset of cases in which detecting dishonesty is nearly impossible, every application of the authoritarian approach suffers from a much larger drawback. To be effective, the powerful judge must be able to monitor individuals, including investigating alleged wrongdoing, determining guilt, and meting out punishment accordingly. This raises a serious concern: Who watches the watcher?

The inescapable paradox is that the stronger the authority, the more it will be able to control the individual, but also the more dangerous it becomes to the individual. Secret police, neighbors spying on friends, and severe control of individual behavior via strict rules and regulations seem the destiny of authoritarian schemes to coerce honesty from unwilling individuals.

There is little doubt that the authoritarian approach to the problem of dishonesty is the most common solution contemplated and applied. Faced with severe cheating, our first instinct is to call the referee and demand that force be applied to ensure truthfulness. There is, however, another alternative—one that does not suffer from the dangers inherent in the authoritarian solution.

Transforming humans to remove the driving force of self-interest or imposing authoritarian control to repress behavior driven by greed is like swimming against a powerful tide. The third approach is completely different. It is based on accepting self-interest and greed as immutable forces, but using them to get desired behavior. We can harness the power of self-interest in favor of our desired end. Individuals are free to decide to lie or not, but lying leaves them worse off. If honesty is the best choice from a self-interested point of view, then honesty is what we will get. This is the key idea underlying signaling theory.

An Economic Model of Used Cars

Suppose that there are only two kinds of used cars: high-quality A cars and low-quality B cars (called *lemons* in the United States). To keep things simple, suppose that there are equal numbers of each and that the high-quality A car is worth $10,000 while the low-quality B car is worth only $5,000.

The seller knows whether his or her car is of low or high quality, but the buyer does not. This is called *asymmetric information* because one party has knowledge and the other does not. The general problem of honesty, in this case, is reduced to figuring out a way to get sellers to tell the truth about the quality of the cars they are selling.

It is important to emphasize that, as illustrated in Figure 17.33, the buyer has no easy way to tell the cars apart. The underlying distribution of cars is on the left, and is known to the seller, but what the buyer actually sees is on the right.

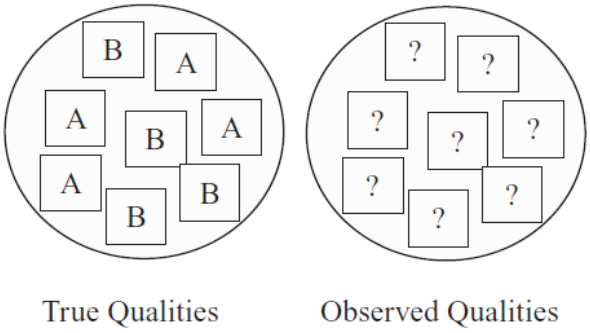


Figure 17.33: The problem of asymmetric information.

In a world where buyers cannot tell the difference between low- and high-quality cars and there are equal numbers of each type, buyers would expect to get a car worth \$7,500 on average. Half of the time they would get a \$10,000 car and the other half a \$5,000 car. Thus, on average, a used car would be worth \$7,500 and this is the amount buyers would be willing to pay for a used car.

Whereas sellers of low-quality cars would be quite happy getting \$7,500 for their low-quality cars, sellers of high-quality cars would be upset. After all, owners of A cars have a product worth \$10,000.

They might try to convince buyers to pay \$10,000 by making claims about the high quality of the car. Declarations about high quality, however, are likely to be ignored because the buyer has no way of knowing if the seller is telling the truth. After all, the seller might actually have a low-quality car worth \$5,000 and is lying to make more money. The buyer would worry that the seller's self-interest would dominate any desire to be honest.

The frustrated sellers of high-quality used cars simply leave the market. This phenomenon is an example of *Gresham's Law*, "bad money drives out good." It was first stated in the 16th century, when monarchs would debase coinage (by adding filler) to get more coins out of a given amount of gold. People would exchange the less valuable coins (bad money) and hoard the pure gold ones (good money). With more bad money in circulation, prices would rise.

Applied to the used car market, the low-quality used cars can be seen as driving out the high-quality cars. Left alone, we would not expect to see high-quality used cars for sale. In fact, that is not what happens—high-quality used cars are sold. How?

Instead of fixing the problem of dishonesty (lying about the quality of the car) by attempting to correct the unethical behavior of the sellers of low-quality used cars (whose dishonesty is causing the trouble here) or imposing authoritarian control over the used car sellers, an alternative scheme has evolved that has certain appealing properties—not the least of which is that car sellers truthfully reveal the qualities of their cars without any central, controlling authority.

Before explaining signaling theory, it is worth pointing out that what is happening here is actually an externality problem. The low-quality sellers fail to take into account the full cost of their lying and, therefore, they lie too much. No individual seller is aware, or would care, that his or her lying is contributing to the elimination of high-quality goods.

Another point that merits attention is that no one designed the system you are about to see. It emerged out of the interaction of buyers and sellers. Probably, some seller of a high-quality car got the idea and, when it worked, it was imitated, but you are about to meet another example, like supply and demand, of a decentralized system.

Signaling Theory

Developed by Spence (1973), the idea behind signaling theory is simple: when we cannot directly observe quality, we use a substitute that is observable (a signal) to enable the market to function. The signal is like a stoplight, green means go and red means stop. The signal will sort the combined low- and high-quality cars into separate markets.

Buyers cannot directly observe the quality of the car, but there are other observable characteristics bundled with the car and seller. *Indices* are attributes that cannot be changed, such as the age of the seller. *Signals*, on the other hand, are observable markers that can be acquired.

The signal, however, must have some special properties to be effective. The signal must be correlated with the underlying, unobservable characteristic. It must be something the A car owner is willing to do, but the B car owner is not, so that it is not immediately copied by unscrupulous sellers of low-quality cars.

In the case of used cars, a common signal is a warranty. Suppose that high-quality cars will have low warranty costs to the seller because they are unlikely to break, but the sellers of low-quality cars would face high warranty costs for their cars that will probably require many repairs.

We have gone as far we can in abstract terms and we are ready to see an Excel implementation of the signaling model.

STEP Open the Excel workbook *SignalingTheory.xls*, read the *Intro* sheet, then go to the *Optimizing* sheet.

The cost of the warranty to the sellers of A and B cars is depicted in Figure 17.34. With no warranty at all (the car is sold "as is"), at a warranty level of zero, a seller has no warranty costs—if something breaks after the car is sold, it is the buyer's problem.

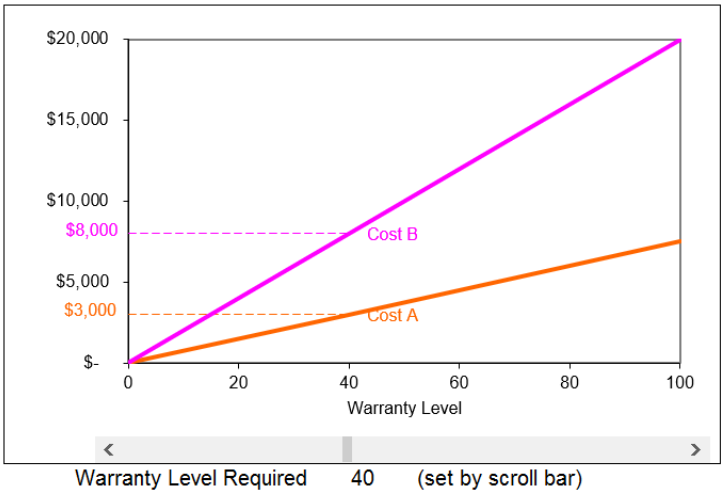


Figure 17.34: Seller's cost of warranty for each type of car.
Source: SignalingTheory.xls!Optimizing.

As the amount of warranty coverage offered by the seller increases, however, costs rise. The seller of the B car's costs rise faster so the gap between the two seller's warranty costs expands.

At a warranty level of 40 (this might be repairs covered by the seller for the first 12 months or 12,000 miles), in Figure 17.34, sellers of high-quality cars expect to incur costs of about $3,000, whereas the sellers of low-quality cars will pay around $8,000 for repairs.

The warranty cost functions are determined by the slopes in cells C6 and C7. It is easy to see that a seller's warranty cost is simply the slope parameter times the warranty level.

Now, suppose there was a warranty level, which is set at 40, initially. Buyers are willing to believe anyone who claims that their cars are high quality and pay the $10,000 price if and only if the car comes with a warranty level of 40.

So the warranty is the signal and any seller who acquires it will sell a car for $10,000. It seems like everyone will offer the warranty, right? Not so fast.

STEP Click the Show Price button.

Excel adds a price function to the chart. It is simply two horizontal lines with a break at a warranty level of 40. The hollow and solid dots mark the discontinuity. The solid dot means the endpoint is included and the hollow dot indicates it is not. Thus, any warranty level from zero up to the signal level (the hollow dot) means the car sells for $5,000. As soon as the signal level is reached, the price jumps to $10,000.

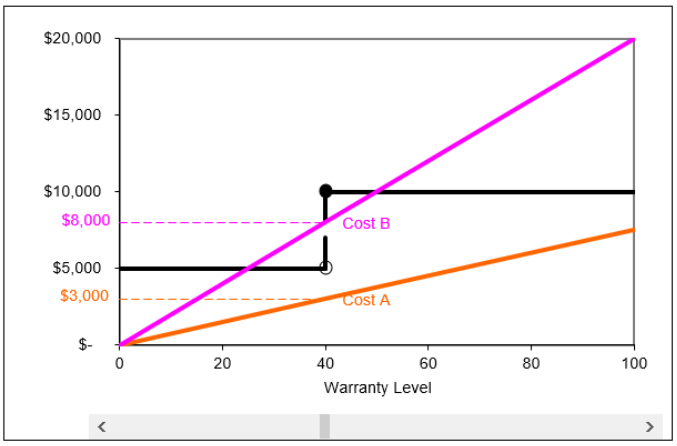
Anyone buying a car with a warranty level below 40 will be willing to pay, at most, $5,000 because it is assumed that the car is of low quality. Even if the car is actually a high-quality car, if it fails to come with the warranty level for high-quality cars, no buyer will pay $10,000 for it because the claim that the car is of high quality is unbelievable without the warranty. On the other hand, a buyer would be willing to pay $10,000 for any car with a warranty level of 40, even if it is actually a low-quality car.

It is now up to the sellers of used cars to make a decision of whether or not to lie. Sellers of low-quality used cars can claim that their cars are high quality and thereby receive the $10,000 high-quality price.

They will not misrepresent the quality of the car, however, because they would end up worse off. Their individual self-interest will drive them to tell the truth.

STEP Click the Show Net Gain button to see why low-quality sellers will not lie.

Figure 17.35 shows what is on your computer screen. We use data from the graph to create a table below that explains how the two sellers will behave.



Warranty Level Required 40 (set by scroll bar)

Decision Making
Net Gain Calculation

	Warranty Offered? No	Yes	Choice
A	$ 5,000	$ 7,000	Yes
B	$ 5,000	$ 2,000	No

Figure 17.35: Understanding why sellers will not lie.
Source: SignalingTheory.xls!Optimizing.

All sellers seek to maximize the net gain, or profit, from the sale of their goods and services. Sellers of used cars would not look simply at the fact that they can make $10,000 by offering a warranty level of 40. This decision-making strategy completely ignores the cost of the warranty. Instead, sellers must compare the net gain, price minus cost of the warranty, to arrive at an optimal decision concerning the warranty level.

The table below the graph contains each type of seller's net gain from selling a car with no warranty versus selling the same car with warranty level of 40. Read the table horizontally—for each type of seller, compare the net gain without and with the warranty, and choose the higher number.

It is clear that sellers of high-quality used cars will offer the warranty level and make $7,000 in profit because that beats the $5,000 net gain if no warranty is chosen. The sellers of low-quality used cars will choose to forgo the warranty and walk away with $5,000 because that is superior to the $2,000 net gain from choosing to lie and offering the warranty.

This is a rather remarkable result. To restate the outcome, the sellers of low-quality used cars will voluntarily and honestly admit that their used cars are of low quality and only worth $5,000. The sellers of low-quality used cars will not lie to the buyers. Is this because they suddenly were overcome by their conscience? No. They are the same fallible, less than perfectly honest people before and after the warranty scheme. Are they telling the truth because an authority figure is watching them, ready to punish liars? No. No one is watching them.

The sellers of low-quality used cars can lie if they so wish. They will not lie, however, because it is not in their self-interest. They end up worse off if they lie in this situation. The warranty scheme has managed to successfully separate or sort the two qualities of cars into their respective groups. This result is called a *separating equilibrium*.

Figure 17.36 shows that the warranty acts as a screen, separating the true car qualities into two distinct groups, Xs and Ys, from which it easy to tell which cars are high quality and which are not.

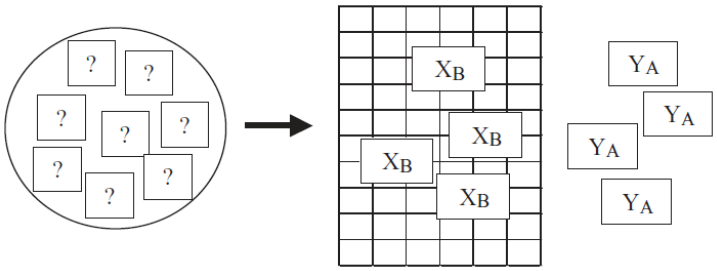


Figure 17.36: Warranty as a screen.

In essence, two markets for cars are created, one for low- and the other for high-quality cars, each with their own prices. Sellers of low-quality cars, although they are able to do so, will not lie and enter the high-quality car market because the price of admission is too high. Lying is not profit maximizing; therefore, sellers will not lie.

Let's repeat a key idea: no individual or organization runs this scheme. No one sets the warranty level and no one sets the price of the cars. The whole system bubbles up from the interaction of the two kinds of sellers and the buyers. Adam Smith would have called it an example of the *invisible hand* of

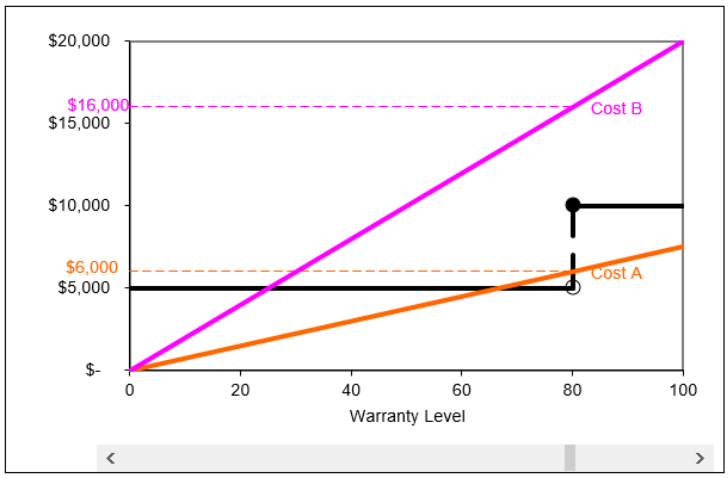
the market; Friedrich Hayek would have described it as a *spontaneous order*; and modern day mathematicians would speak of *self-organizing systems*. It is all the same thing—individual interaction generating a quite agreeable systemwide result. To see how the equilibrating forces operate in this model, we examine how the signaling scheme can break down.

Signaling Failures and Equilibrium

One way that a signal can fail is if it is set too high.

STEP Use the scroll bar to set a high warranty level like 80 or so.

In this case, as shown in Figure 17.37 and your computer screen, not even the sellers of high-quality cars find it in their self-interest to offer the warranty level that brings the $10,000 price. The signal has failed to separate the two qualities of cars.



Warranty Level Required 80 (set by scroll bar)

Decision Making
Net Gain Calculation

	Warranty Offered? No	Yes	Choice
A	$ 5,000	$ 4,000	No
B	$ 5,000	$ (6,000)	No

Figure 17.37: Signaling failure from a warranty level set too high. *Source: SignalingTheory.xls!Optimizing.*

On the other hand, if the signal is set too low, sellers of B cars will find it in their self-interest to lie and claim their cars are actually high quality. They

will choose the warranty level that brings the $10,000 price.

STEP To see this, use the scroll bar to set a low warranty level, 20 or less.

Your screen should show that both sellers opt to acquire the signal. The low-quality seller will lie and claim that the car is of high quality because the net gain from lying (cell H27) is greater than the net gain from telling the truth (cell G27). Once again, this signal has failed.

When the signal is too high, the holes in the screen are too small and no one can get through. If the signal is too low, the holes are too large and everybody passes through. In a separating equilibrium, the level of the signal is such that the two types are sorted and grouped together so they are easily identifiable.

The fact that signals can be observed as failing provides the key to understanding how the system can settle down to a result that effectively solves the problem without central control. If the signal is too low, self-interested sellers of high-quality cars will offer higher warranty levels in order to block their lying brethren from diluting their market. The sellers of high-quality cars want to distance themselves from low-quality sellers.

If the signal is too high, no one will take it and buyers will lose the means by which to identify the two qualities of cars. The market will collapse so pressure will push the level down.

The forces inherent in the system, self-interested behavior by the interacting agents, will conspire to generate an equilibrium signal level that effectively sorts the two qualities of cars. The process works just like supply and demand—pressure in disequilibrium pushes the signal in one direction or another until it equilibrates.

STEP Play around with the warranty level to reveal the range for which it effectively separates the two qualities of cars.

You already know 80 is too high and 20 is too low. Look at the chart to help you see what must be true for the signal to succeed. When you are ready to check your answer, click the Range button.

Other Applications of Signaling Theory

We have barely scratched the surface of signaling theory. There are many situations in which one party to a transaction has available information that the other party lacks and this asymmetric information puts honesty in peril.

Consider the job market (which was Spence's original example). Faced with many job applicants, all claiming to be high-productivity A workers, the firm might insist on a signal, a college degree, to back the claims made by job applicants.

Suppose that low-productivity workers are also likely to be weaker students, and that it is more costly for them to acquire the educational signal. As in the used car case, the successful screen will separate the two worker groups into their respective low- and high-productivity categories. The signal will elicit honest responses from low-productivity workers because lying requires a college degree to be believed and this is not in their best interest.

Additional applications of signaling include insurance, legal bargaining, and firm entry models. In both health and life insurance, asymmetric information is critical. The insurance company does not know the health status of the applicant. If the price of the insurance depends on the applicant's health, just saying they are healthy is not enough for the insurance company to believe it.

In a lawsuit, where the plaintiff seeks damages from the defendant, asymmetric information means neither party knows the other's true intentions and beliefs. They can signal the strength of their case by demanding a high pre-trial settlement.

Firm entry models use signaling to convey the degree of confidence and strength of incumbent firms to potential newcomers. Incumbents can signal or make reliable claims about their low costs and ability to compete by charging low pre-entry prices.

In these cases, an incentive mechanism has developed that accepts self-interest among buyers and sellers as a powerful, immutable, driving force. Instead of fighting self-interest by removing or suppressing it, the incentive mechanism uses self-interest to reach the desired end.

The Economics of Honesty

Dishonesty exacts a large cost on society. For lesser developed countries, corruption is a severe obstacle to economic growth. Getting people to be truthful is a serious, critically important goal.

The primary solutions to the problem of dishonesty have centered on utopian and authoritarian approaches. The former seeks to perfect human behavior; the latter to directly control it. A third, somewhat counterintuitive, alternative exists that relies on self-interest to yield an agreeable systemwide result.

This third alternative is marked by individuals following their self-interest. When geese fly in a V-shaped pattern over thousands of miles, they do so not under the guidance of an authoritarian drill sergeant or master goose who tells each bird where to fly, but because they obey a simple rule that says, "If there are no birds around, fly; if a bird is in front, fly just off its wing because it is easier." This minimizes the effort for each bird and produces a pattern which no bird intended.

Likewise, modern society is composed of millions of individual agents whose interaction establishes a systemwide pattern. Unsatisfactory results can be changed via transmuting the motivating forces of each agent, imposing decisions on each agent, or changing the incentives faced by each agent. The last option is rarely considered, but may be the most effective and best of the three.

Signaling theory says that by making honesty the best policy—for the selfish, greedy individual—we will get honesty. Sellers reveal the truth because lying leaves them worse off than telling the truth. This is the economics of honesty.

To be sure, signaling requires rules and institutional support. If the seller of low-quality used cars knows that he can renege on warranties or other contracts because the court system is nonexistent or corrupt, then signaling will be useless.

There is, however, a world of difference between an authoritarian approach that relies on a central power to coerce honesty and the system that evolves out of the interaction of the buyers and sellers given appropriately supporting institutions. The decentralized system avoids the question of "Who watches the watcher?" because there is no dominant, central power. And in the end, this may be its most significant advantage.

Exercises

1. Suppose a firm is trying to determine whether an applicant is of low or high ability and it believes people with long fingernails have higher ability. Would fingernail length be an effective signal? Draw a graph to support your answer.

2. Draw a graph that shows how education as a signal could be used to separate low- and high-ability job applicants. Explain how education as a signal works.

3. Draw a graph in which education as a signal fails because the signal level is set too high. Explain why the signal fails.

4. College education as a signal clashes with *human capital theory*, which says that educated workers earn more because they were made more productive by their education. What does signaling theory say about the value of education? In other words, according to signaling, why are educated workers paid more?

5. Why has it been difficult to determine with data whether human capital or signaling theory is right about college education?

References

The epigraph is from page 495 of George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," The *Quarterly Journal of Economics*, Vol. 84, No. 3 (August, 1970), pp. 488–500, www.jstor.org/stable/1879431. This is the paper that Michael Spence described as "quite electrifying" in his 1991 Nobel acceptance lecture (available at www.nobelprize.org). The Nobel Prize was shared that year by Akerlof, Spence, and Joseph E. Stiglitz "for their analyses of markets with asymmetric information."

Akerlof's paper led to an exchange concerning the empirical validity of the claim that lemons drove out high-quality used pickup trucks. Eric W. Bond, "A Direct Test of the 'Lemons' Model: The Market for Used Pickup Trucks," *The American Economic Review*, Vol. 72, No. 4 (September, 1982), pp. 838–840, www.jstor.org/stable/1810022 found no evidence for the claim.

Michael Pratt and George Hoffer, "Test of the Lemons Model: Comment," *The American Economic Review*, Vol. 74, No. 4 (September, 1984), pp.

798–800, www.jstor.org/stable/1805151 conduct a "finer test" and "conclude that the market for used pickup trucks is a lemons market."

In a reply, Eric W. Bond, "Test of the 'Lemons' Model: Reply," The *American Economic Review*, Vol. 74, No. 4 (September, 1984), pp. 801–804, www.jstor.org/stable/1805152 said that "Pratt and Hoffer find used trucks to be of lower quality not because they have a 'finer' test, but because they fail to adjust for observable quality differences and include trucks that are more than 10 years old." Bond believes there is no lemons effect for used pickups because institutions have arisen to counteract the effects of asymmetric information.

This debate and projects involving new data to test signaling would make an excellent thesis topic.

The signaling model is laid out in Michael Spence, "Job Market Signaling," The *Quarterly Journal of Economics*, Vol. 87, No. 3 (August, 1973), pp. 355–374, www.jstor.org/stable/1882010. This article was based on his doctoral dissertation and published as a book titled *Market Signaling* in 1974.

George Selgin, "Gresham's Law," available at eh.net/encyclopedia/greshams-law/, offers a short explanation of the history and application of this concept.

Chapter 18

General Equilibrium

The Edgeworth Box

General Equilibrium Market Allocation

Pareto Optimality

General Equilibrium Monopoly

"[Irma Adelman] was an early proponent of simulation models. In addition to work with input-output and linear-programming models, she was one of the pioneers in developing computable general equilibrium (CGE) models and applying them to developing countries, especially for analysis of income distribution."

Distinguished Fellow Citation for Irma Adelman

18.1 The Edgeworth Box

We have become quite familiar with society's resource allocation problem. We have used partial equilibrium analysis to focus on a single commodity, exploring how supply and demand determine an equilibrium quantity that is the market's answer to the resource allocation question.

We know all about consumers' and producers' surplus, market failure, and deadweight loss. We have repeatedly drawn supply and demand graphs and emphasized comparison of equilibrium to socially optimal output.

But the focus on a single commodity is limiting. In fact, the market system uses supply and demand for each good or service to answer the fundamental production and distribution questions. In other words, there are many interacting markets (one for each commodity) simultaneously in operation.

If we monopolize one commodity, we cause a misallocation of resources in the monopolized market (too little is produced). Partial equilibrium analysis stops there. But the low output and high price in the monopolized market reverberates throughout the economy. After all, resources that would have gone into that market are going to go somewhere else and the high price in the monopolized commodity will shift demand curves for substitutes and complements of that good.

General equilibrium analysis attempts to account for supply and demand in *all* markets at once. As you can imagine, it is much more difficult than partial equilibrium analysis, but it is also superior because the entire resource allocation question is under consideration.

This book focuses on general equilibrium *theory*, but as the epigraph to this chapter explains, computable general equilibrium models are used to estimate the general equilibrium effects of tax policies, monopoly power, and other events. Economists have always been aware of the limitations of par-

tial equilibrium analysis, but it was not until the development of modern computers that these complicated models could be solved and applied.

Before beginning our study of general equilibrium theory, two observations are in order.

1. Society can decide which goods and services are handled by the market. Society may decide that human organs or votes may not be legally bought and sold. Different market-based societies may choose different lists of commodities to be allocated by the market. We call a society *market based* if individual resource owners make decisions about how to allocate the inputs they manage, even if particular commodities are regulated or entire sectors of the economy (such as education or health) are not privately owned.

2. A complete general equilibrium analysis of the market system is beyond our scope. There are three parts, of which this book covers only the first one.

   (a) Pure exchange: Assume each consumer has endowments of already produced goods and allow trade to occur.

   (b) Production: Allow goods to be produced from inputs.

   (c) Combine pure exchange and production in a general equilibrium analysis.

We focus solely on pure exchange and ignore the next two stages. This means we will not complete a true general equilibrium analysis of the market system. Emphasizing only the problem of pure exchange enables you to see the core concepts of general equilibrium, including the Edgeworth Box graph, without overwhelming complexity.

Even limiting ourselves to a situation where all products are already made requires serious investment of intellectual capital. As we will see, the Edgeworth Box is a clever graph, but it takes some practice to read it.

Our work on pure exchange will enable us to come full circle and return to the beginning—consumers decide what to buy and sell based on the optimal solution to an Endowment Model. As you work on the model and recall ideas and terminology, you will further cement truly fundamental knowledge.

Constructing the Edgeworth Box

The canonical graph used to depict a pure exchange economy is called the *Edgeworth Box*. It is also commonly referred to as the Edgeworth-Bowley Box. It turns out that both names are wrong. Blaug (1996, p. 523), discussing something called the *Ricardo Effect*, points out an interesting thing about names:

> Whether it really is in Ricardo is a nice question. The fact that the Ricardo Effect is hard to find in Ricardo exemplifies a general rule. According to R. K. Merton, 'eponymy' is the "the practice of affixing the name of the scientist to all or part of what he has found" but it is a striking fact that the outcome of eponymy is almost always to hang the right label on the wrong person. Thus, Thomas Gresham never stated Gresham's Law. Jean Baptiste Say only stated Say's Law after James Mill had stated it for him. Robert Giffen never stated Giffen's Paradox. Francis Edgeworth never drew the Edgeworth Box. Ernst Engel never drew an Engel's curve. Walras never stated Walras' Law. Irving Fisher did not invent the Ideal Index Number and actually pleaded (in vain) that it should not be named after him. Arthur Bowley did not enunciate Bowley's Law. Arthur Pigou did not state the Pigou Effect—and so on. Indeed S. M. Stigler has advanced "Stigler's Law of Eponymy: No scientific discovery is named after its original discoverer," a law which is confirmed as soon as it is stated (see *Transactions of the New York Academy of Sciences*, Series 11, 39, 1980). Nevertheless, there are also counter-examples in economics to Stigler's Law, such as Pareto-optimality and the Wicksell Effect.

If it was not Edgeworth, then who created the canonical graph of general equilibrium analysis? According to Tarascio (1972), it was Vilfredo Pareto (pronounced pa-ray-toe) who should be credited with inventing the graph that we call the Edgeworth Box. Because no one has ever heard of the Pareto Box, we will continue to call it the Edgeworth Box, but now you know the truth behind the name.

The Edgeworth Box is a graph that is constructed by putting together the consumer choice problem graphs from two consumers. It ends up looking like a box; hence its name. While most books just draw a box, we can use Excel to see exactly how you build an Edgeworth Box.

STEP Open the Excel workbook *EdgeworthBox.xls* and read the *Intro* sheet, then go to the *A* sheet to see consumer A's optimization problem.

Take the time to look over the sheet. The goal is to maximize satisfaction, given by a Cobb-Douglas utility function that faithfully reflects the consumer's preferences. The budget constraint's slope is $-\frac{p_1}{p_2}$ and at the initial endowment (35,10), the MRS is less than the price ratio.

You know you do not need to run Solver because at $25{,}16\frac{2}{3}$ (the actual values on the sheet are Solver's false precision) the equimarginal condition is met and the consumer is reaching the highest attainable indifference curve.

At the given prices, the sheet shows that A will maximize utility, subject to the budget constraint, by selling 10 units of x_1 and buying $6\frac{2}{3}$ units of x_2. These are the *net demands* for x_1 and x_2.

STEP Proceed to the *B* sheet to see consumer B's optimal solution.

Notice that B has a different initial endowment (5,30) than A, but the rest of the optimization problem is the same. Given the same prices faced by consumer A, consumer B optimizes by buying 20 units of x_1 and selling $13\frac{1}{3}$ units of x_2.

Figure 18.1 has Endowment Model graphs for the two consumers. We can see that they make different decisions about what to buy and sell. A moves up the constraint (selling x_1 and buying x_2), while B does the reverse.

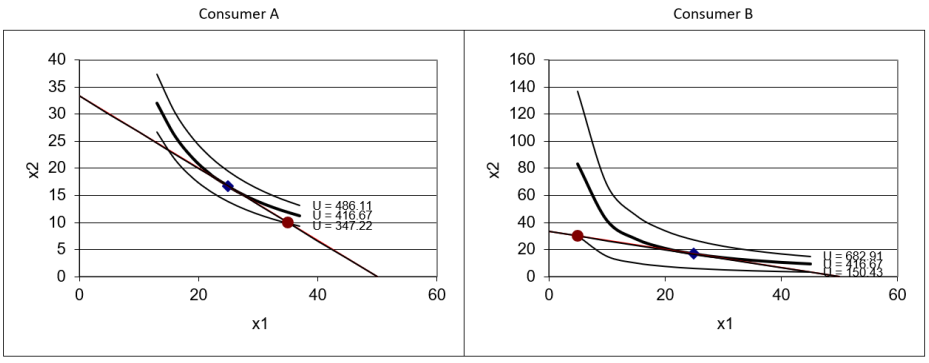


Figure 18.1: Preparing to build the Edgeworth Box.
Source: EdgeworthBox.xls!A and B.

Figure 18.1 shows the two consumers side by side and that helps us see what they are both doing, but it does not show how their plans for buying and selling match up. This is the key to the Edgeworth Box. We want to be able to instantly see if the two consumer's optimal decisions mesh.

The crucial step in understanding the Edgeworth Box is the next one: Flip consumer B's graph, as shown in Figure 18.2. Sheet *B* in Edgeworth-Box.xls shows how to do this.

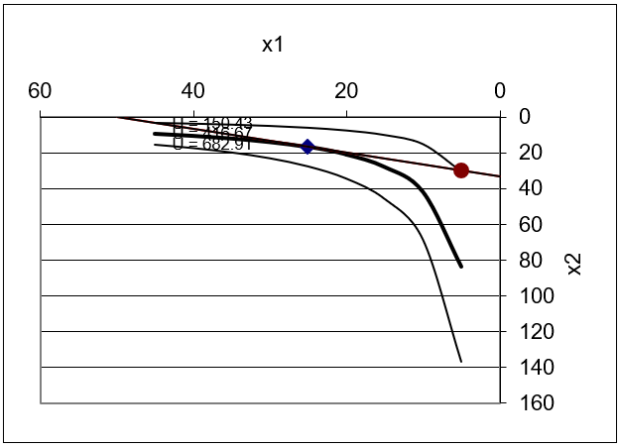


Figure 18.2: Flipping B's graph.
Source: EdgeworthBox.xls!B.

STEP Follow the instructions in column F of sheet *B* to replicate Figure 18.2.

Actually flipping B's graph will help you remember that B's decisions about buying and selling are always read from the perspective of the northeast (top right) corner of the Edgeworth Box.

The last step in constructing the Edgeworth Box is to join A's graph with B's flipped graph. The result of this operation is a graph that looks like a box.

STEP Proceed to the *EdgeworthBox* sheet for your first look at an Edgeworth Box. You may need to scroll down a bit to see it.

How is this chart created? By following the instructions above and taking advantage of Excel's ability to make transparent objects.

STEP Click on the graph to select it, and then drag the graph to the right.

It comes apart! Clearly, the Edgeworth Box is simply two separate graphs superimposed on top of each other. The top graph has no fill, so it is transparent.

STEP Click the Make Box button to put the box back together. The button simply lines up the two graphs precisely to make it easy to create the box.

STEP Scroll back up to see the organization of the sheet.

Let's take a tour of the sheet. The two consumers' optimization problems are represented in columns A and B and columns M and N. In the middle (columns G and H), market information is displayed. Cells H16 and H17 contain the prices of the two goods.

The price of good x_2, called the *numeraire*, has been set equal to 1 and p_1 is expressed as $\frac{p_1}{p_2}$. Instead of $p_1 = 2$ and $p_2 = 3$, we can focus on $\frac{p_1}{p_2} = \frac{2}{3}$ as the relative price. With many goods, a single one is chosen (think gold) as the numeraire and everything is priced relative to that good. In the next chapter, we will see how prices respond to supply and demand.

Properties of the Edgeworth Box

The Edgeworth Box has properties and conventions that will be helpful in our future work. Here are a few of them.

1. The sides of the box give the total amounts of the two goods available. Total $x_1 = 40$ units and total $x_2 = 40$ units so this box is a square.

2. If there is more total x_1 than x_2, then the box is wider than it is tall (if the same axis scale is used for both goods). The first exercise question asks what it means if the box is tall and skinny.

3. Since consumers face the same prices, one budget line is shared for both consumers.

4. The slope of the budget line is the price ratio, $\frac{p_1}{p_2}$, and that is what matters, not the individual prices themselves. By convention, we normalize the problem and set $p_2 = 1$, and call x_2 the numeraire.

5. Net demands for x_1 and x_2 for both A and B can be read from the box. This requires careful attention because it is easy to be tricked. Remember to read B's decisions about buying and selling from the top right corner.

6. The Edgeworth Box has enough information to figure out how prices will change and where the equilibrium solution lies. The next section shows how.

Edgeworth Box Basics

This section introduced the canonical graph of general equilibrium theory. It is unlikely that you have seen this graph before so we are proceeding slowly. Figure 18.3 shows the chart the *EdgeworthBox* sheet.

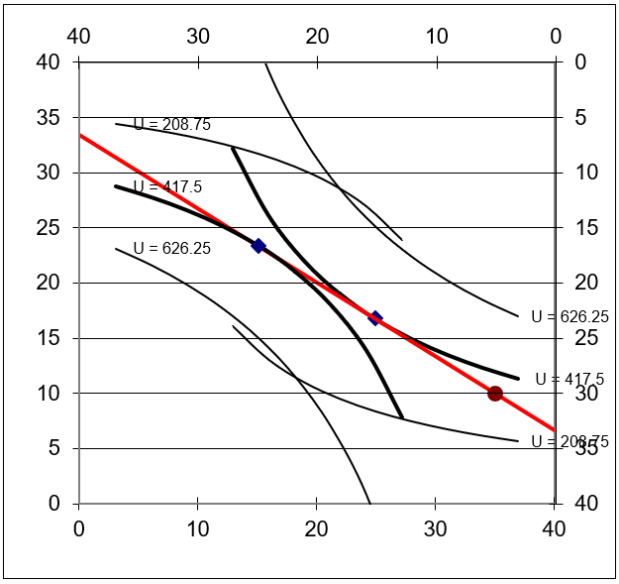


Figure 18.3: An Edgeworth Box in disequilibrium.
Source: EdgeworthBox.xls!EdgeworthBox.

The Edgeworth Box simultaneously displays the optimization problems of two consumers. A's view is the usual x–y axis configuration with the origin in the lower left corner of the graph. B's graph has been flipped so

the origin is at the top right corner. Thus, x_1 rises as you move to the left on the top of the box and x_2 rises as you move down the right side of the box.

If you drew an Edgeworth Box on a piece of paper or are reading this on a laptop or tablet, you could literally rotate the paper or device so that B was the usual configuration at the bottom left and A's axes were at the top and right. This would not change anything substantive.

In the next section, we will use the Edgeworth Box to see how both markets equilibrate simultaneously. This is the hallmark of general equilibrium analysis. Figure 18.3 is not in equilibrium. There are forces that will make the red budget line swing.

The Edgeworth Box will also be used to explain the concept of Pareto optimality and the idea of economic efficiency in a general equilibrium setting. Although it does not have the widespread recognition of supply and demand, the Edgeworth Box is a truly foundational graph in general equilibrium theory. It is important to grasp how it is constructed and read to be able to understand future concepts that rely on the Edgeworth Box.

Exercises

1. Suppose an Edgeworth Box was very tall and very skinny. What would that tell you?

2. Use Word's Drawing Tools to draw an Edgeworth Box that is the same as the *EdgeworthBox* sheet except B's utility function is $U = minx_1, x_2$. Draw three representative indifference curves for B.

   Hint: Return to the Theory of Consumer Behavior to find out what the indifference curves look like for this utility function.

3. Click the Reset button in the *EdgeworthBox* sheet and set cB in cell M21 to 0.1. Click the Take a Picture button and paste the graph in your Word document.

4. Explain B's buy/sell decision for each good.

5. How does B's buy/sell decision make sense given that B has so little of x_1 and so much of x_2?

References

The epigraph is from "Irma Adelman: Distinguished Fellow 2003," *The American Economic Review*, Vol. 94, No. 3 (June, 2004), www.jstor.org/stable/i369727. Advances in computers have enabled real-world, empirical applications of general equilibrium analysis. Computable general equilibrium models (CGEs) are used to find equilibrium solutions with many agents and commodities. The effects of taxes and other shocks are simulated and evaluated.

The history of how computers have been used in ever more sophisticated economic models is a story of determination and grit. See Irma Adelman, "The Research for the Paper on the Dynamics of the Klein-Goldberger Model," *Journal of Economic and Social Measurement*, Vol.32 (2007), pp. 29–33, content.iospress.com/articles/journal-of-economic-and-social-measurement/jem00269, to learn how Adelman and her physicist husband, Frank Adelman, used an IBM 650 mainframe computer in 1958 to produce one of her most famous articles, "The Dynamic Properties of the Klein-Goldberger Model," *Econometrica*, Vol. 27, No. 4 (October, 1959), pp. 596–625, www.jstor.org/stable/1909353. This was the first attempt to solve an econometric model with an electronic computer. Adelman (2007) also says that the work was "I believe, a first application of Monte Carlo techniques in economics." (p. 32)

In an introduction to Adelman's description of how the model was estimated, Renfro describes the IBM 650 and how incredibly impressive it was that Adelman managed to use it to estimate the model. In addition to covering her den with pieces of paper indicating the contents of each memory register at each step in the computation and having to pay more than a month of her salary for 1 hour of computing time, Renfro (p. 24) points out that the work had to be done at night. "Throughout the entire mainframe era, those who needed to get something done quickly worked through the night. Computers in those days had multiple users; this was the time of day that provided the best turnaround, when only the most serious were awake." See Charles G. Renfro, "Introduction," *Journal of Economic and Social Measurement*, Vol. 32 (2007), pp. 23–28, content.iospress.com/articles/journal-of-economic-and-social-measurement/jem00271.

Agent-based computational economics (ACE) is related to CGE. To learn more about "growing economies from the bottom up," visit www2.econ.iastate.edu/tesfatsi/ace.htm.

On the claim that it should be called the *Pareto Box*, see Vincent Tarascio, "A Correction on the Geneology of the So-Called Edgeworth-Bowley Diagram," *Economic Inquiry*, Vol. 10 (1972), pp. 193–197, onlinelibrary.wiley.com/doi/10.1111/j.1465-7295.1972.tb01599.x.

Economic Theory in Retrospect is a classic book on the history of economic thought (the intellectual history of the discipline) is Mark Blaug (1962 originally published, 5$^{\text{th}}$ edition, 1996).

> Without Pareto, the Theory of General Equilibrium, of which Walras was without question the real founder, would never have acquired the fame which it has now, nor indeed would it have been possible to speak of the Lausanne School.
>
> Umberto Ricci

18.2 General Equilibrium Market Allocation

Partial equilibrium analysis relies on supply and demand for a particular commodity to explain how the market establishes an equilibrium output that is society's answer to the resource allocation question. The figure X traced out by supply and demand lines is perhaps the most basic and well known picture in economics.

Compared to the easy, familiar supply and demand graph, general equilibrium analysis labors and struggles with a new graph, the Edgeworth Box, that is confusing when first encountered. It is busy, with many elements, and requires the user to change persepective to read it. As you work on mastering the Edgeworth Box, remember this: the equilibration process in an Edgeworth Box is based on the same logic used in supply and demand analysis.

We will leverage knowledge of supply and demand to explain how general equilibrium works and to learn how to read the Edgeworth Box.

Tatonnement: The Equilibration Process

Introductory economics students know that shortages cause prices to rise and surpluses push prices downward. In a supply and demand graph, the price is displayed as a horizontal line that falls when it is above the intersection and rises when it is below.

In the Edgeworth Box, there are two markets simultaneously equilibrating. The prices of the two goods are displayed by a single line, which is the budget constraint faced by the two consumers. The slope of the price line, also known as the *price vector*, is $-\frac{p_1}{p_2}$.

Just like supply and demand, shortages and surpluses push prices up and down. In the Edgeworth Box, this translates to the price vector swinging.

Remember that we are considering the special case of a pure exchange economy. All products have been produced and individuals are trading from their initial endowments. Prices are determined competitively by the interaction of all buyers and sellers—every consumer takes prices as given.

A two-dimensional Edgeworth Box allows for only two consumers. A third consumer would make it a cube and, beyond that, we run out of dimensions and cannot draw the object (although it exists). Our two-consumer, toy model version implements price-taking behavior by supposing that there is an *auctioneer* who shouts out prices. Our consumers take these prices as given and use them to make buy and sell decisions.

Although each commodity has a price, in general equilibrium analysis, only relative prices matter. We can arbitrarily take one good and set its price to 1. This makes that good the numeraire.

Our two consumers hear the prices and make optimizing decisions based on those prices. If the buy and sell decisions do not match, the prices are adjusted by the auctioneer. No trades are actually made until all markets are in equilibrium.

As prices are called out by the auctioneer, the price vector rotates around the initial endowment, swinging to and fro. It becomes more vertical as $\frac{p_1}{p_2}$ rises and flatter if $\frac{p_1}{p_2}$ falls. We mean, of course, rising and falling in absolute value.

At any moment, the consumers can compute the optimal amounts of each good to buy and sell. If the amounts each wants to buy and sell are not mutually compatible, then the price vector swings toward the equilibrium price vector.

The word *tâtonnement* (pronounced ta-tone-mon) was used by the French economist Leon Walras (1834 - 1910) (pronounced Val-rasse) to describe the equilibration process. Google translates it as groping. Walras visualized the market groping, feeling, working its way through an iterative process that converged to a position of rest. In the technical literature of general equilibrium theory, the word tatonnement (without the circumflex) is accepted without italics.

You may have noticed that the terminology of general equilibrium analysis has a decidedly French-language flavor to it. Walras, the father of general

equilibrium theory (and described by Schumpeter as "the greatest economist ever") was French. His successor at the School of Lausanne was Vilfredo Pareto (1848 - 1923), a native Italian with a background in math and engineering, who invented the concept of Pareto optimality (and is the actual originator of the Edgeworth Box).

In the second half of the 19th century, continental European economists were at the leading edge of general equilibrium theory and mathematical economics. This strong mathematical tradition continues today. French-born Gerard Debreu and Maurice Allais have won Nobel Prizes in Economics for their work in general equilibrium theory.

We will use Excel to implement a concrete problem with actual prices, surpluses, and shortages to see how the Walrasian model works.

STEP Open the Excel workbook *EdgeworthBoxGE.xls*, read the *Intro* sheet, then go to the *EdgeworthBox1* sheet.

We review the display, piece by piece. It is worth going slowly and being careful. There is a lot going on and the details matter.

Consumer A's optimization problem is in columns A and B. No need to run Solver—cells B11 and B12 contain A's optimal reduced-form expression. With a price vector with slope $-\frac{2}{3}$, consumer A would like to sell 10 units of good 1 and buy $6\frac{2}{3}$ units of good 2.

Columns M and N display consumer B's optimization problem. Like A, we have entered the reduced-form formulas for B's optimal consumption of the two goods. At the initial prices, consumer B wants to buy 20 units of good 1 and sell $13\frac{1}{3}$ units of good 2.

This information is all we need to know that the p_1 relative price in cell H16 is not an equilibrium, or market clearing, price. After all, A wants to sell more x_1 than B wants to buy and vice versa for x_2.

Thus, no trades will be made at these prices and the Walrasian auctioneer will call out new prices as the search for equilibrium goes on.

We can also use the Edgeworth Box to reach this same conclusion about the plans not matching at the initial relative price of -0.67.

STEP Scroll down to see the Edgeworth Box.

Figure 18.4 reproduces a portion of what is on your screen, augmented with arrows and dashed lines to help explain what is going on.

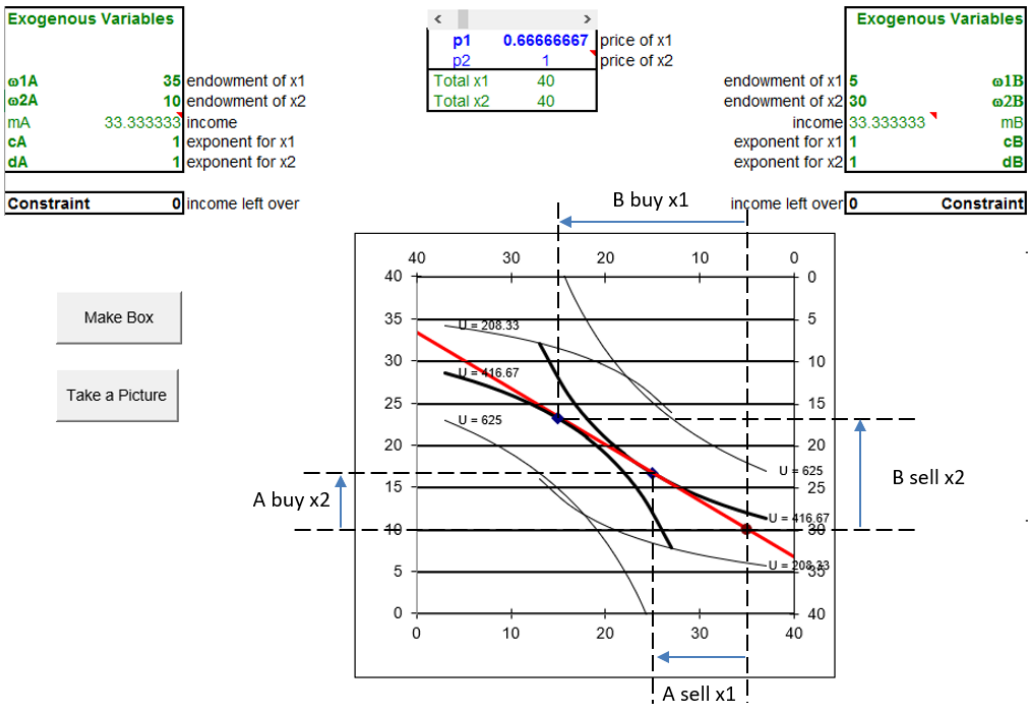


Figure 18.4: An Edgeworth Box in disequilibrium.
Source: EdgeworthBoxGE.xls!EdgeworthBox1.

We begin with A, which is easier than B. In Figure 18.4, arrows along the bottom and left sides of the box indicate what A wants to do: sell x_1 and buy x_2. It is natural to read the dashed lines from A's optimal solution and see that left on the x axis means sell, while up on the y axis means buy.

Reading B is trickier. B also has arrows, but they run the reverse of the usual because we read B's graph from the northeast corner. B wants to buy x_1 and sell x_2.

The direction of the arrow indicates buying or selling. Although one wants to buy and the other sell, the length of the arrows in Figure 18.4 show that the plans do not match. The length of the arrows indicate the *amounts* to be bought and sold. If the lengths are not equal, we are not in equilibrium.

We review the buy and sell decisions of B more carefully, to make sure there is no confusion. B wants to buy 20 units of good 1. From her initial endowment of 5 units, she wants to move *left* along the top axis, which means acquiring *more* x_1, until she ends up with 25 units. On the other hand, she wants to sell $13\frac{1}{3}$ units of good 2, moving *up* the right axis —which means she is *reducing* her desired amount of x_2.

If you get in the habit of drawing dashed lines on an Edgeworth Box, either on a piece of paper or by inserting dashed line shapes in Excel or Word, *from the optimal solution of A and B*, you greatly increase your chances of reading the graph correctly. Those dashed lines are a visual cue that remind you to read A from the bottom left and B from the top right.

STEP Scroll down below the Edgeworth Box to see two supply and demand graphs.

These are the partial equilibrium markets for the two goods. Good 1 shows a shortage, with price below the intersection of supply and demand. Good 2 has demand and supply reversed from the usual display because the price on the y axis is p_1/p_2. There is a surplus of x_2 at $p_1/p_2 = \frac{2}{3}$.

Both markets adjust simultaneously. We know there is upward pressure on p_1 from the shortage and downward pressure on p_2 from the surplus. This will make the price ratio rise and the price vector will become steeper.

STEP Use the scroll bar (over cells G15 and H15) to see how price changes affect the box. Set the price ratio to 1.5.

The spreadsheet does most of the hard work for you. A's and B's optimal solutions are instantly calculated. The market position cells immediately reflect the position of markets for each good at the new prices (where good 1 is one and a half times as expensive as good 2).

The Edgeworth Box is a live graph that reflects the new price vector. It shows that we have overshot the equilibrium price vector because we now have a surplus of good 1 and a shortage of good 2.

STEP Practice reading the Edgeworth Box. With $\frac{p_1}{p_2} = 1.5$, use the graph to read the amounts that A and B want to buy and sell. Compute the surplus and shortage of each good from the box alone.

Verify (using the cells in the Market Position part of the sheet) that your answers are correct. Look at the graphs below the Edgeworth Box to make sure you understand that the Edgeworth Box conveys the same information about the position of each market.

STEP Play with the price vector, adjusting the scroll bar to set different price ratios and interpreting how the consumers will respond to each price ratio by using the Edgeworth Box.

As you rotate the price vector, you are the Walrasian auctioneer. You are calling out prices and the two consumers are reacting to them. The more you practice reading the Edgeworth Box, the more comfortable you will get with it.

As you adjust the price ratio, the price vector swings to and fro. It always rotates around the initial endowment (which would change if and only if any of the four initial endowment parameter values change). The tatonnement process is how the market responds to shortages and surpluses by changing prices in such a way that the surpluses and shortages are reduced, until they are completely eliminated.

There is, of course, no auctioneer in the real world, but price pressure from surpluses and shortages are quite real. Our model captures these pressures by the fiction of the auctioneer changing prices in response to disquilibrium in the two markets.

General Equilibrium

You have seen how shortages and surpluses push the price line to and fro, swinging around the initial endowment point.

We know that equilibrium means *no tendency to change*. We apply this definition of equilibrium to this particular model: when $\frac{p_1}{p_2}$ has no tendency to change, we know we have settled to the equilibrium solution. The equilibrium solution generated by the market tells us how much x_1 and x_2 each consumer will end up with if the market is used and how much each consumer wants to buy and sell of each good.

STEP Use the scroll bar to find the equilibrium price vector.

The equilibrium solution in a General Equilibrium Pure Exchange Model is a canonical economics graph that is reproduced as Figure 18.5. If your screen does not look like this graph, set the price ratio to 1.

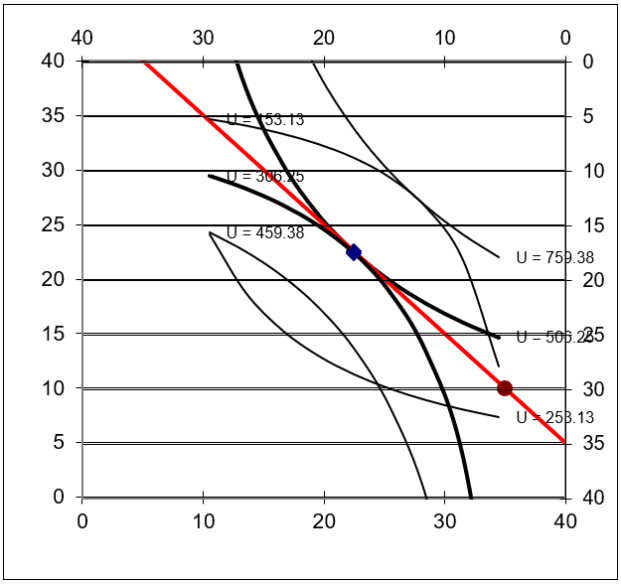


Figure 18.5: The canonical graph of general equilibrium.
Source: EdgeworthBoxGE.xls!EdgeworthBox1 with $\frac{p_1}{p_2} = 1$.

As Figure 18.5 clearly shows, when the equilibrium position is reached, the optimal solution of both consumers lies on the same point. This eliminates all shortages and surpluses (as shown in the supply and demand graphs below the Edgeworth Box) so the price ratio has no tendency to change.

The single point in the Edgeworth Box represents a mutually compatible solution for both consumers and is the hallmark of a general equilibrium solution. The single point is akin to the intersection of supply and demand in a partial equilibrium analysis.

Our general equilibrium model shows how the market is an allocation mechanism. It will redistribute the initial endowments of the two consumers by using prices until it settles down to a position where plans match and forces in the model are in balance.

Notice, however, that the two consumers don't get equal amounts of the two goods. Why does A end up with more? Because A started out richer. At the equilibrium price vector, the market values A's endowment at $45 and

B's at $35. General equilibrium theory does not ask why A is richer. It takes the initial endowment as given.

Walras' Law

Leon Walras is the father of General Equilibrium Theory. The law that bears his name states the following: *The value of aggregate excess demand is identically zero.*

Using Walras' Law, we can deduce the following logical result: If n–1 markets are in equilibrium, then the last market must be in equilibrium.

A concrete demonstration of Walras' Law is the best way to understand what it means.

STEP With $p_1 = 1$ (at the equilibrium solution), change p_2 (cell H17) to 2. Find the equilibrium p_1.

The equilibrium p_1 is now 2. This shows that, no matter the value of p_2, the equilibrium solution will be found when $\frac{p_1}{p_2}$ equals one.

Thus, it looks like there are two endogenous variables here, p_1 and p_2, but there is really only one endogenous variable, $\frac{p_1}{p_2}$. This is the idea behind Walras' Law and why we can find equilibrium in both markets by varying only p_1.

STEP Click the Reset button. Scroll right to cell V5 and click the Show Walras' Law button to reveal calculations that demonstrate Walras' Law in action.

Although the two markets are not in equilibrium, the sum of the value of aggregate net demands in cell Y11 is zero. Look at the cell formulas in row 11 to see how they are computed.

STEP Change p_1 (via the scroll bar) and notice that no matter the price, the sum of the value of aggregate net demand is always zero.
A direct implication of Walras' Law is that in a general equilibrium system with n goods, we do not have to find n prices. If $n - 1$ markets are in equilibrium, the last one automatically has to be in equilibrium.

This is why we actually have only a single endogenous variable, the price ratio, in the two-good case. All that matters is the relative price, not the two individual prices. With n goods, one good would be the numeraire (historically, gold has played that role) and all other goods would be valued in terms of the numeraire.

Comparative Statics with the Edgeworth Box

Having found the initial equilibrium solution, we could pursue a variety of comparative statics experiments, shocking an exogenous variable and tracking how the equilibrium solution (of various endogenous variables) responds.

STEP Click the Reset button and then set cA (cell B21) to 2. What happened to A's indifference curves and optimal solution?

With steeper indifference curves (since A likes good 1 more than before), A's new tangency point is quite close to the initial endowment. This means A wants to sell little x_1. You can scroll down to see how the partial equilibrium graphs have changed—the chart of x_1 confirms we have a big shortage.

STEP Where is the new equilibrium solution? If you decide to use Solver to answer this question, please make the target cell H15 because that is the cell that the scroll bar is affecting. This way you will not destroy the formula in cell H16.

You should find a new equilibrium solution at a relative price ratio of about 1.53. Approximately 7.3 units of good 1 will be traded and 11.8 units of good 2 will be exchanged.

Two Advanced Ideas

In a mathematical sense, General Equilibrium Theory is perhaps the most abstract and sophisticated area of economics. Two questions that have been studied intensively involve existence and uniqueness.

The question of the existence of an equilibrium solution was posed by Walras himself. The issue, loosely stated, is that we cannot be sure that a general equilibrium system with thousands or millions of individual goods has a place where the entire system is at rest. In fact, from an intuitive point of view,

given the huge number of products, consumers, and firms in a real-world economy, we might doubt that an equilibrium solution exists at all.

Walras and other early theorists thought that if the number of endogenous variables (unknowns) equaled the number of equations, then a solution was guaranteed. This is not so. Existence proofs in the 1950s utilized *fixed-point theorems* to prove rigorously the conditions under which an equilibrium solution was guaranteed to exist. Brouwer and Kakutani fixed point theorems are examples of this approach.

Closely tied to existence is the problem of the uniqueness of a general equilibrium solution. Even if an equilibrium solution is proved (in a rigorous mathematical sense) to exist, the worry is that there may be multiple equilibria in a general equilibrium system. Research has focused on what assumptions must be invoked to guarantee a single equilibrium solution.

Existence and uniqueness proofs are well beyond the scope of this book. They rely on topology and advanced mathematical concepts. This is another way of saying that our presentation of the Edgework Box and general equilibrium in a pure exchange economy is introductory and rudimentary. General Equilibrium Theory is a vast ocean and we are paddling near the shore.

Market Allocation in an Edgeworth Box

The canonical supply and demand graph is used in partial equilibrium analysis to find the equilibrium solution. General equilibrium uses the Edgeworth Box to do the same thing.

It appears cumbersome and tedious at first, but, in fact, it is an ingenious graphical device. By representing two consumers simultaneously, while sharing a common budget constraint (given that they face identical prices), the box enables one to quickly see whether the two-good, pure exchange economy is in equilibrium. It also reveals how prices must change as the system finds its way to equilibrium via the tatonnement process.

Whether a pure exchange economy is in a general equilibrium can be determined in an instant by seeing whether the optimal solutions of the two consumers are compatible—that is, if there is a single point where the two consumers want to be, given the existing price ratio.

But what about the final, equilibrium allocation generated by the market—what are its properties? This is a fundamental question that leads to the famous Pareto optimality conditions and the First Fundamental Theorem of Welfare Economics. It is explained in the next section.

Although we have used numerical methods (implementing the problem in Excel) to analyze and find the general equilibrium solution, you should be aware that there are analytical approaches also. We could write down demands for goods by each consumer and impose the equilibrium condition that $Q_D = Q_S$ in each market. This would enable solution of the equilibrium price vector with the aid of algebra (and, as soon as we left the simple world of two or three goods, linear algebra).

Exercises

1. Use Word's Drawing Tools to draw your own Edgeworth Box. Place the initial endowment so that A has more x_2 than x_1.

2. Add a price vector to your box in the previous question that generates a shortage of x_1. Draw arrows along the bottom and top x_1 axes to show the amount of x_1 each consumer wants to buy or sell.

3. Use Word's Drawing Tools to draw a supply and demand graph for x_1. Include a horizontal line in the graph that shows the current price of x_1.

4. Add the equilibrium price vector to your Edgeworth Box graph in question 1. Explain why this price vector is the equilibrium solution.

   Hint: Add indifference curves to your graph to support your explanation.

References

The epigraph is from page 11 of Umberto Ricci, "Pareto and Pure Economics," *The Review of Economic Studies*, Vol. 1, No. 1 (October, 1933), pp. 3–21, www.jstor.org/stable/2967433. You can learn more about Walras, Pareto, and the Lausanne School by visiting the History of Economic Thought web site at www.hetwebsite.net/het/.

Perhaps no area of economics is as mathematically sophisticated and intense as General Equilibrium Theory. There has always been disagreement among

economists regarding the use and necessity of mathematics in economics. Pareto sneered at the literary economists and the use of math as a weapon continues today.

Akerlof says that economists only value "hard" and ignore "soft" questions so the discipline stifles research into issues that cannot be answered with formal tools and models. See George Akerlof (2020),"Sins of Omission and the Practice of Economics," https://doi.org/10.1257/jel.20191573, 58(2), 405–418, doi.org/10.1257/jel.20191573.

Roy Weintraub traces the influence of math in economics in *How Economics Became a Mathematical Science*, published in 2002. For the connection between economics and physics, see Phil Mirowski, *More Heat than Light*, published in 1989.

> Except during short intervals of time, people are always governed by an elite. I use the word elite (It. *aristorocrazia*) in its etymological sense, meaning the strongest, the most energetic, and most capable—for good as well as evil. However, due to an important physiological law, elites do not last. Hence, the history of man is the history of the continuous replacement of certain elites: as one ascends, another declines.
>
> Vilfredo Pareto

18.3 Pareto Optimality

Evaluating the welfare effects with general equilibrium is the same as with partial equilibrium. First we determine the equilibrium solution, then we find the optimal solution, and last we compare the equilibrium to the optimal solution.

The previous section used an Edgeworth Box with a price vector to find the initial equilibrium solution. We know that shortages and surpluses swing the price line to and fro until it settles down where the plans of the two consumers are mutually compatible.

In this chapter, we use the Edgeworth Box to display the optimal solution. The price vector is removed because prices play no role in determining the optimal solution. Just as with partial equilibrium, we logically separate the equilibrium from the optimal solution. If the two agree, then we know we have a good result.

Optimality

STEP Open the Excel workbook *EdgeworthBoxParetoOpt.xls*, read the *Intro* sheet, then go to the *EdgeworthBox* sheet.

The workbook is quite similar to the EdgeworthBox sheet from the previous section, except there is no price or market position information. We are not interested in markets right now. We are focused on determining the optimal solution.

An omniscient, omnipotent social planner, OOSP, is charged with determining the optimal allocation, given the initial endowment.

With OOSP's special powers, we can reallocate the initial endowment as we see fit. Each point in the box is an allocation, distributing the total amounts of the two goods to A and B. We can arbitrarily give and take from one person to the other, choosing any point in the box. What should we do?

At first glance, it might seem that we would want to solve an optimization problem like this:

$$\begin{aligned} &\max U_A(x_{1A}, x_{2A}) + U_B(x_{1B}, x_{2B}) \\ &\text{s.t. } x_{1A} + x_{1B} = \text{Total } x_1 \text{ and } x_{2A} + x_{2B} = \text{Total } x_2 \end{aligned}$$

In other words, we could give consumers A and B the amounts of goods 1 and 2 that maximize the sum of the individual utilities subject to the total goods available.

This strategy suffers from a serious problem: *We cannot make interpersonal utility comparisons.* This brings us full circle to work we did at the very beginning of this book in the Theory of Consumer Behavior. Utility is ordinal, not cardinal. Monotonic transformations (that keep rankings intact) of utility are allowed. Utility has no meaning in terms of its units.

Thus, an optimization problem that aggregates individual utilities is invalid. It makes no sense to say that the utility of A is added to the utility of B to get a total utility. There are no common units with which to measure and add utility. You might as well say that you added three cars and four pencils and got seven carpencils.

There is, however, a way to judge and evaluate different allocations of goods to A and B. This is Pareto's great contribution to welfare economics.

Pareto developed logical rules that enable us to get around the limitations of utility. His basic idea was that you can compare two allocations in terms of better or worse so you can make statements about one allocation compared with another. He invented a new vocabulary for his rules and today we use his name when we work with these rules.

Pareto knew we cannot add utility, but we might be able to compare two allocations and declare which one is better. We proceed by example, using the Excel workbook and Figure 18.6.

From the initial endowment point in Figure 18.6, suppose we consider the point (30,15) for A and (10,25) for B.

Figure 18.6 reproduces what is on your screen. The two thicker indifference curves going through the initial endowment are the starting point. They represent the benchmark satisfaction to which we will compare other allocations.

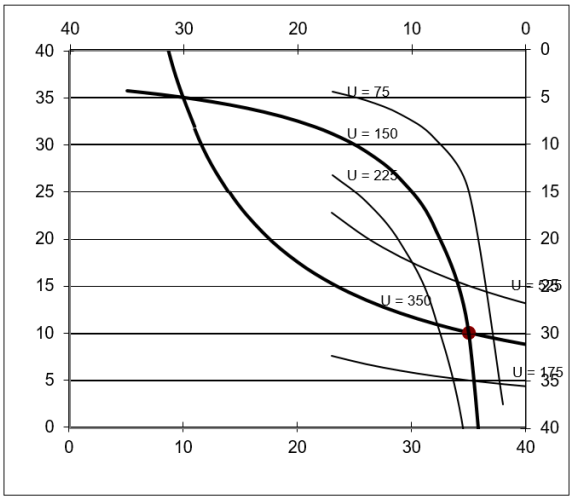


Figure 18.6: Edgeworth Box for Pareto criteria.
Source: EdgeworthBoxParetoOpt.xls!EdgeworthBox1.

From the initial endowment point in Figure 18.6, suppose we consider the combination of 30,15 for A and 10,25 for B.

STEP Click the (30,15) button. A red point appears at that coordinate in the box along with a text box.

Is A better off at the new point compared with the initial endowment? How about B?

As the text box explains, although the indifference curves for A and B are not drawn through the red point, we know they exist because the indifference map is dense—there is an indifference curve through every point in the box. If we draw an indifference curve for A through that point and it lies above the indifference curve that goes through the initial endowment, we know that A prefers 30,15 to the initial endowment.

In fact, indifference curves extend beyond the box in a northeast direction for A and southwest for B. The box just shows the total amounts available for exchange.

The same argument we made for A can be made for B. The only trick for B is to remember that you interpret the box from the top, right corner and B's satisfaction increases as the indifference curves move farther away from the northeast corner in a southwesterly direction.

Because both A and B are better off at 30,15 than the initial endowment, we know that the 30,15 allocation is *Pareto Superior* to the initial endowment. We can also flip the statement to say that the initial endowment is *Pareto Inferior* to point 30,15.

Pareto Superior means that it is possible to make at least one person better off without making anyone else worse off. We make no claims as to how much better off. We do not use the units of utility at all. This is similar to how we first discussed satisfaction in the Theory of Consumer Behavior. We asked consumers to simply choose between one bundle and another. The same logic is being used here.

Consider another point that is 30,10 for A and 10,30 for B.

STEP Click the (30,10) button.

As before, a red dot is placed on the chart and a text box appears. We want to compare the red dot to the initial endowment. Is A better off? How about B?

Because the point 30,10 is better for B, but worse for A, then this allocation is *Pareto Noncomparable* to the initial endowment because at least one person is made worse off. As soon as at least one person is made worse off, it is removed as a candidate for evaluation.

We certainly cannot evaluate these points by saying B's utility goes up by more than A's falls because utility is only ordinal. According to Pareto, we can never trade off a small decrease in satisfaction for one person for a large gain in satisfaction for one or many people because you cannot add up utility.

Now that we understand Pareto Superior and Pareto Noncomparable points, we can shade in all of the points that are Pareto Superior to the initial endowment. This is called the *lens* for reasons that will be obvious in a moment.

STEP Click the Show Lens button.

Every point in the space between and including the two indifference curves going through the initial endowment is shaded red, representing the area of Pareto Superior points. With usually shaped indifference curves, this is a lens-shaped object.

We return to the first point, 30,15. It is, of course, inside the lens so it is Pareto Superior to the initial endowment, but does it have any points that are Pareto Superior to it?

STEP Click the Hide Lens button and then the 30,15 button.

The 30,15 point, like the initial endowment, has a whole set of points that are Pareto Superior to it. These points also form a lens, albeit smaller that the lens formed by the Pareto Superior points to the initial endowment, that stretch from the point 30,15 to where the two indifference curves intersect again.

Clearly, whenever indifference curves from A and B cross at a point, such as the initial endowment or 30,15, we can find Pareto Superior points in a lens from that starting point. What happens when the indifference curves are tangent?

STEP Click the Hide Lens (if needed) and Show B Curve buttons.

A red dot is shown on an indifference curve for B that is tangent to A's highest displayed indifference curve. We will call this point of tangency between the indifference curves point PO1. This point PO1 is obviously Pareto Superior to the initial endowment since it is inside the lens.

But there is something special about PO1. It has a property that contains Pareto's key idea: Does PO1 have any Pareto Superior points to it? No, it does not. Movement in any direction from point PO1 lowers someone's satisfaction. There is no lens from point PO1.

Thus, we say that PO1 is a *Pareto Optimal* point—one that has no Pareto Superior points to it. You cannot make someone better off without hurting someone else. Pareto Optimal points are where we want to be!

It is important to note that there are an infinite number of Pareto Optimal points. Wherever the indifference curves are tangent, we are at a Pareto Optimal point.

The set of all Pareto Optimal points is called the *contract curve*. A minimalist version of a contract curve for an unknown (but well-behaved) pair of utility functions is displayed in Figure 18.7. A few indifference curves are displayed, but you should understand that every point on the contract curve is a point of tangency between two indifference curves. The sides of the box are not labeled, but you know how to read an Edgeworth Box.

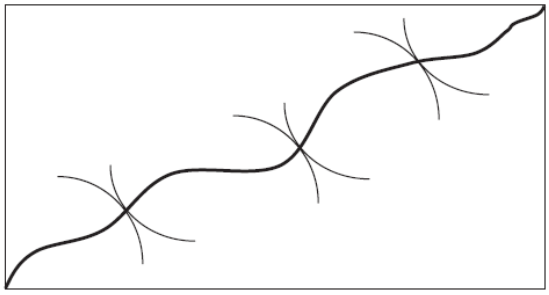


Figure 18.7: The contract curve.

Pareto Optimal points are especially desirable because they ensure that there is no way to improve the allocation without harming someone. In other words, given the limitations of ordinal utility, we can say that we have wrung out as much gain as possible if we are at a Pareto Optimal point. Thus, from any given initial endowment, OOSP would want to reallocate the two goods so that the allocation is on the contract curve.

One drawback of the Paretian framework is that there are many Pareto Optimal points when starting from an arbitrary, non-Pareto Optimal point. There is no way to choose between Pareto Optimal points.

Mathematically, it should be clear that Pareto Optimal points occur only when $MRS_A = MRS_B$. When this condition holds, the two indifference curves are tangent. This means we have a Pareto Optimal point and we are on the contract curve.

Pareto Optimality with Solver

One way to find Pareto Optimal points is to solve an optimization problem. It is not the silly, nonsensical "sum the utilities" objective function, however.

STEP From the *EdgeworthBox* sheet, open Solver.

Your Solver dialog box should look like Figure 18.8. Notice the $UtilityB = Initial\_UtilityB$ constraint. We are going to maximize A's utility without harming B. The constraint requires that B's utility be the same as the initial utility. Thus, B will be indifferent between the new allocation and the initial endowment.

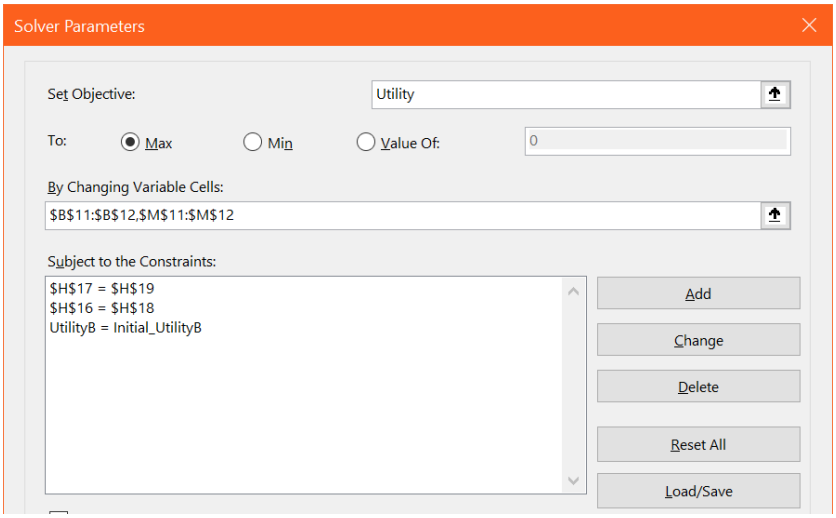


Figure 18.8: Solver parameters dialog box.

STEP Click Solve to find an optimal solution to this problem.

Scroll down (if needed) to see the Edgeworth Box. We are at the top most (from A's point of view) Pareto Optimal point. This point is on the contract curve.

What if we ran the same analysis, but maximized B's utility subject to maintaining A's utility constant? This is yet another Pareto Optimal point.

Some students want to make claims about points in the middle of the contract curve in the lens as being somehow better than the two extreme points, but the Pareto analysis does not allow for such distinctions.

The Contract Curve with Excel

STEP Proceed to the *ContractCurve* sheet.

It is set up just like the *EdgeworthBox* sheet, except A's Initial Endowment cells (B18 and B19) have a formula, =ROUND(randomnv()*38+1,0).

This formula allows you to generate random initial endowments, then you can use Excel's Solver to find a point on the contract curve from that initial endowment. You can use the "max A's utility keeping B's utility constant" or "max B's utility keeping A's utility constant" strategies. In the former case, you are finding the highest indifference curve of A that is tangent to B's indifference curve that goes through the initial endowment. You are doing the reverse when you maximize B's utility subject to A's indifference curve that goes through the initial endowment.

STEP Click the New Endowment button a few times to move the initial endowment point around the box. When you find one you like (it does not matter), find and record a point on the contract curve. Do this several times.

You are sampling points on the contract curve and this helps you learn how Pareto optimality works. Can you discover the shape of the contract curve?

STEP Change A's preferences by setting cA to 0.5. Sample points on the contract curve (using the same method as in the previous step). What effect does this have on the contract curve?

To see the answers to these two questions (but first try to answer them on your own), click the Show CC with cA = 0.5 button.

The First Fundamental Theorem of Welfare Economics

It is no exaggeration to say that we have reached the summit of this book. We are about to see the crowning achievement of economic theory—a demonstration of the welfare effects of the market system in a general equilibrium framework.

With the Pareto criteria in hand, we are ready to judge the market allocation. Recall that the market uses prices to establish an equilibrium solution. Surpluses and shortages push the price vector to and fro until it settles down to its equilibrium solution. What can we say about the market's solution?

We can say that it is Pareto Optimal! In fact, we can say that starting from any initial endowment, a market allocation mechanism yields a Pareto Optimal solution. This is the *First Fundamental Theorem of Welfare Economics*:

> If preferences are well-behaved, a properly functioning market's equilibrium solution is Pareto Optimal.

Figure 18.9 reproduces Figure 18.5 for your convenience. It is the canonical graph of general equilibrium analysis and shows the equilibrium solution from the Edgeworth-BoxGE.xls workbook. We know we have the equilibrium solution because there is a single, common tangency point. Consumer A maximizes by choosing that combination where he reaches the highest indifference curve subject to the constraint. Consumer B does the same.

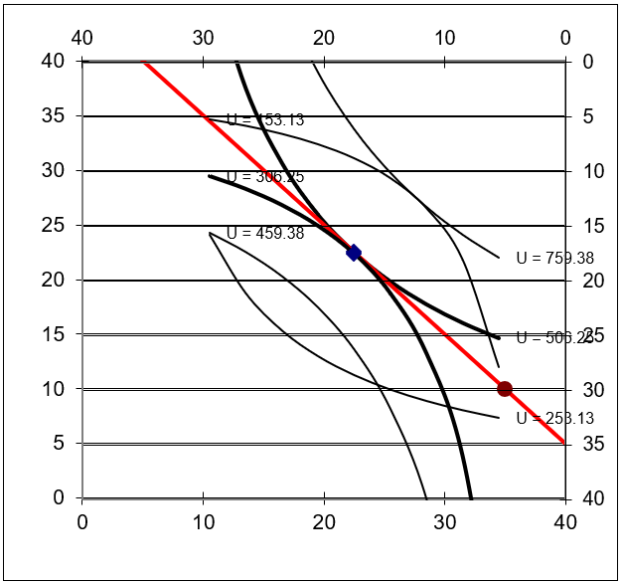


Figure 18.9: Evaluating the market allocation.
Source: EdgeworthBoxGE.xls!EdgeworthBox1 with $\frac{p_1}{p_2} = 1$.

But it is immediately obvious, given our work in this section, that the market allocation is Pareto Optimal. There are no Pareto Superior points to it.

We can use the equimarginal principle to help explain this result. Each consumer is finding a point of tangency that obeys the mathematical condition, $MRS = \frac{p_1}{p_2}$. From A's perspective, we have $MRS_A = \frac{p_1}{p_2}$. Similarly, B chooses that combination where $MRS_B = \frac{p_1}{p_2}$. Unbeknownst to them, they are ending up at a point where $MRS_A = MRS_B$.

In other words, by paying attention to prices and optimizing, the equilibrium generated by exchanging consumers is at the same time generating a Pareto Optimal solution. There is an invisible hand aspect to this in the sense that the consumers do not know and do not care about Pareto Optimality.

Geese fly in a V pattern over thousands of miles by drafting—wind resistance is minimized by aligning one-self at angle to the goose ahead, instead of flying directly behind or next to a fellow goose. The geese are completely unaware that they are generating a V-shaped pattern. Consumers in a market are just like geese—they are completely unaware that they are solving a much bigger optimization problem.

Geese also synchronize their wing beats because they take advantage of updraft. If you watch a flock, it looks like they are coordinating their flapping. This was discovered recently (see Portugal, et al., 2014) and provides an excellent example of how economists see the market system.

With each agent following a simple rule, the system produces a pattern. In the case of the market, it is an incredible result that the market allocation is Pareto Optimal.

What can't we say about the market allocation?

We certainly can't say that it is *fair*. The market will grind to a Pareto Optimal point from any initial endowment. The Pareto logic takes the initial endowment as given. What if A starts out with much more than B? What if the market does not value B's resources? The Pareto criteria have nothing to say about this. Economists have tried to include fairness in welfare analysis, but there is little consensus.

If there's a First Theorem, there must be a Second Theorem, right?

If preferences are well behaved, a properly functioning market can reach any Pareto Optimal point if the appropriate initial endowment is provided.

The Second Fundamental Theorem says that you can use the market to reach any Pareto Optimal allocation—that is, any point on the contract curve. All you have to do is set the initial endowment appropriately, then let the market work its magic.

The last two problems in the Q&A sheet ask you to show that the Second Fundamental Theorem works.

That Markets Generate Pareto Optimal Solutions Is a Truly Fundamental Idea

This section marks the end of a long trek. We began with the Theory of Consumer Behavior and learned that consumers maximize satisfaction subject to a budget constraint. An important extension of this basic model utilizes an initial endowment instead of cash income.

In a Pure Exchange Model, we combine two optimizing consumers in an Edgeworth Box. Their interaction results in an equilibrium solution.

Using the Pareto criteria, we can compare allocations and determine which ones are Pareto Optimal. These are allocations that have no Pareto Superior points. The set of all Pareto Optimal points forms the contract curve.

Students struggle with the term Pareto optimality. Its definition, that there is no way to make someone better off without hurting someone else, can become a jumble of words with little real meaning. Here is the crucial idea: Pareto Optimality means no waste. The allocation at a Pareto optimal point cannot be improved upon (without harming someone). Thus, Pareto optimality means we have an unbeatable allocation.

The First Fundamental Theorem of Welfare Economics makes a powerful statement because it says that a properly functioning market yields a Pareto Optimal allocation. This is a highly desirable result.

It is also shocking because individual consumers have no idea they are participating in solving a resource allocation problem. Each consumer is simply maximizing utility subject to a budget constraint. Like geese that fly in a V, each consumer is responding to a signal (in the consumer's case, prices) and then the interaction is producing the coordination.

Notice that the work here has said nothing about innovation or technological change. In fact, the analysis assumes constant technology and no new products. The analysis is completely static and based solely on the market's ability to reach a Pareto Optimal solution in terms of allocating already produced goods in a pure exchange economy.

You might be wondering if all equilibria in an Edgeworth Box are Pareto Optimal? Absolutely not. The next section shows how the market can fail.

Exercises

1. Why do the Pareto criteria fail to provide a single point that is the best allocation?

2. What must be true about the exponents in the Cobb-Douglas utility functions for consumers A and B to generate a linear contract curve? Describe your procedure and explain your answer.

3. Use Word's Drawing Tools to draw an Edgeworth Box with well-behaved preferences and a point Z, where the $MRS_A > MRS_B$. Explain why point Z is not Pareto Optimal.

4. The contract curve (with $cA = 0.5$) can be transformed into a utility possibilities frontier, as shown in Figure 18.10. Where would point Z (from the previous question) be on this graph? Explain why.

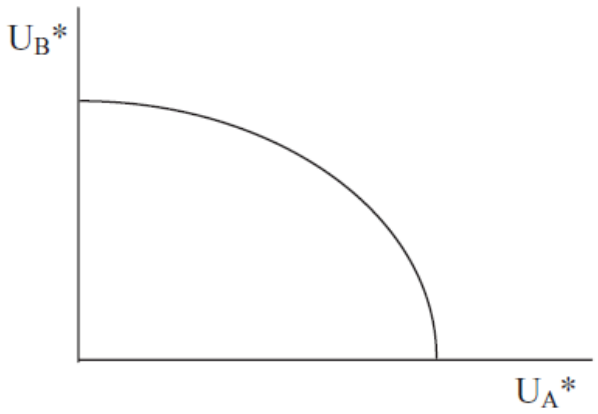


Figure 18.10: A utility possibilities frontier.

References

The epigraph is from page 36 of Vilfredo Pareto's *The Rise and Fall of Elites: An Application of Theoretical Sociology* (originally published in Italian in 1901, translated to English in 1968, and published as a paperback in 1991 and 2000). The back jacket says,

> Here in brief and incisive outline are the major ideas for which Pareto was later to become famous ... This slim volume is more readable and disciplined than most of the later elaborations, and serves well as an introduction to Pareto's political sociology ... Pareto's irony shows in his attack on elites that become humanitarian and tender-hearted rather than tough-minded.

Most economists know Pareto through his work on utility, General Equilibrium Theory, and the idea of Pareto Optimality, but Pareto grew disenchanted with "pure economics" (what we would call today economic theory) and turned to sociology. His most famous sociological work is *Mind and Society* (originally published in 1916 and first translated into English in 1935), in which he explains how the circulation of elites drives history.

See Vincent J. Tarascio, *Pareto's Methodological Approach to Economics* (published in 1968) for a comparison of Pareto's views on the scope and method of economics, especially as contrasted with Alfred Marshall. Whereas Marshall saw mathematics as a language, capable of being translated so nonmathematicians could understand, Pareto believed that "mathematics makes it possible to express relations between facts which are not possible with other facilities or ordinary language" (Tarascio, p. 106, footnote omitted). Pareto saw no need to translate heavily mathematical papers for the "literary economists." Many of Pareto's ideas on optimization and equilibrium were presented in prose form by Philip H. Wicksteed, *Common Sense of Political Economy* (first published in 1910) and available online at www.econlib.org/library/Wicksteed/wkCS.html.

On geese flying in a V and coordinating flapping, see Steven J. Portugal, Tatjana Y. Hubel, Johannes Fritz, Stefanie Heese, Daniela Trobe, Bernhard Voelkl, Stephen Hailes, Alan M. Wilson and James R. Usherwood (2014), "Upwash exploitation and downwash avoidance by flap phasing in ibis formation flight," *Nature*, 505, pp. 399–402, www.nature.com/articles/nature12939. My video, *The Invisible Hand and the Market System*, freely available at vimeo.com/econexcel/invisiblehand, has a clip of the authors explaining how the geese do it.

1) Let E be an economy such that, for every i,
(a) X_i is convex,
(b) if x_i^1 and x_i^2 are two points of X_i and if t
is a real number in $]0,1[$, then
$x_i^2 \succ x_i^1$ implies $tx_i^2 + (1-t)x_i^1 \succ x_i^1$.
An equilibrium $((x_i^*), (x_j^*))$ relative to a price system p, where no x_i^* is a satiation consumption, is an optimum.

Gerard Debreu

18.4 General Equilibrium Monopoly

Partial equilibrium analysis tells us that monopoly causes an inefficient allocation of resources—too little output (compared with the socially optimal level) is produced.

This section explores the welfare implications of monopoly in a general equilibrium setting. The procedure is the same as the one used for judging competitive markets: We determine the monopoly allocation and then test it by comparing it to the set of Pareto Optimal points (i.e., the contract curve).

To reiterate, monopoly results in an inefficient allocation of resources. There is no dispute about that. However, General Equilibrium Theory is the best way to demonstrate this inefficiency.

Monopoly in an Edgeworth Box

Suppose we start with the usual Edgeworth Box. It has an initial endowment that is the point of departure for trade between the two consumers.

Competitive markets are modeled in an Edgeworth Box by supposing that prices are determined by the interaction of many buyers and sellers. To implement price-taking behavior in a two-person Edgeworth Box, we use an auctioneer who calls out prices. Each consumer determines optimal amounts to buy and sell based on the given prices. The Edgeworth Box is used to check whether the amounts that each consumer wants to buy and sell are compatible. If not, prices adjust based on the shortages and surpluses generated by the plans of each consumer.

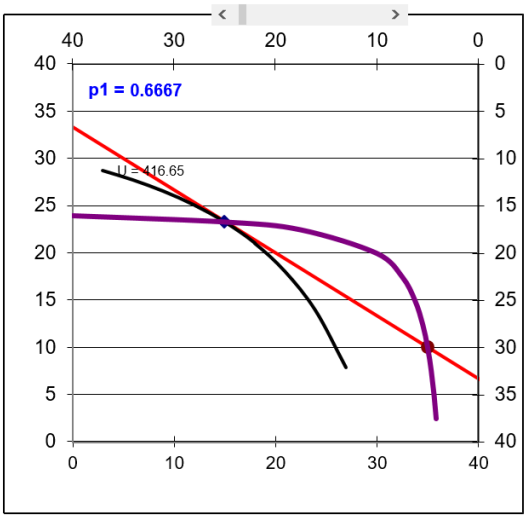
We model monopoly in a pure exchange Edgeworth Box by eliminating the auctioneer. We give one of the consumers monopoly power. They can set the price vector to have any slope.

Suppose that A is a monopolist. What does this mean in the context of the Edgeworth Box? A will quote prices to B and let B decide how much to buy and sell. A will choose a price ratio and this determines the final allocation.

We can think of A as an auctioneer who first shouts out prices to see how B will respond, then picks the best prices—from A's point of view.

STEP Open the Excel workbook *EdgeworthBoxMonopoly.xls*, read the *Intro* sheet, then go to the *PriceOfferCurveB* sheet.

Figure 18.11 (and your screen) shows B's *price offer curve*, which tells A how much x_1 and x_2 B wishes to hold given the price ratio, $\frac{p_1}{p_2}$.



B's Price Offer Curve

p1/p2	x1B	x2B
0.4	40.00	16.00
0.6667	25.00	16.67
1	17.50	17.50
2	10.00	20.00
3	7.50	22.50
4	6.25	25.00
5	5.50	27.50
6	5.00	30.00
7	4.64	32.50
8	4.38	35.00
9	4.17	37.50

Figure 18.11: B's offer curve.
Source: EdgeworthBoxMonopoly.xls!PriceOfferCurveB.

Initially, A has set $p_1 = 0.6667$ (p_2 is the numeraire). B maximizes utility, given that price ratio, by choosing the combination 25,16.67. This is shown by the black indifference curve that is tangent to the red price vector. B will want to buy 20 units of good 1 and offer (hence the name offer curve) 13.33 units of good 2 for sale to A.

A can set any price for good 1 she wishes, but B gets to decide how much to buy and sell at A's chosen price. Also, we assume A will honor the deal and buy the amount B wants to sell.

STEP Click the scroll bar above the graph a few times to change the price of good 1.

With each click, the red budget constraint line rotates about the initial endowment and B chooses a new optimal bundle.

The locus of points that B chooses as p_1 is varied, ceteris paribus, is the *price offer curve*. For any given price, B finds the place at which the highest indifference curve is tangent to the budget constraint—and this point is on the price offer curve.

Having explained B's price offer curve, we bring A into the picture. A knows B's price offer curve and has the monopoly power to set any price for x_1. Given $p_2 = 1$, A has the power to set the slope of the price vector. The key question is: Which price will A choose?

In one sense, the answer is obvious: Choose p_1 that maximizes satisfaction for A. But how can this problem be solved so we find the best price from A's point of view?

STEP Proceed to the *EdgeworthBox* sheet.

The display is the same as on the *PriceOfferCurveB* sheet, except that now we have added A's indifference curves. We also can easily see A's utility in cell C28.

Is the initial price of 0.6667 the best solution for A? No, because by increasing p_1, A gets greater satisfaction.

STEP Confirm that this is true by clicking on the scroll bar to increase p_1 and keeping your eye on A's utility in cell C28.

You can also control the price with the scroll bar over cells A9 and B9. Notice how the price has been moved under the heading of *Endogenous Variables*. Because A chooses the price—this is what monopoly power means—price is endogenous to the monopolist.

In the *Wealth of Nations*, Adam Smith says, "The price of monopoly is upon every occasion the highest which can be got" (Book I, Chapter VII, www.econlib.org/library/Smith/smWN.html?chapter_num=10#book-reader).

But is this true? Would the monopolist literally charge the highest price possible?

STEP Drag the scroll box in the scroll bar all the way to the right.

The chart is hard to read, but we can see from the table next to the chart that with $p_1 = 9$ (the highest price we can set with the scroll bar), B wants to end up with 4.17 units of x_1 and 37.5 units of x_2. This means p_1 is so high that B does not want to buy any of it and, in fact, wants to sell 0.83 units to A!

More importantly, a quick glance at cell C28 reveals that A's utility is under 90. This means that, taken literally, a monopolist will not charge the highest price possible.

Just like a monopoly in a partial equilibrium setting, A is operating under a constraint. A monopolist takes the demand curve as given. Consumer A takes B's offer curve as given and B's offer curve acts as constraint for A.

With this knowledge, can you solve A's problem? What is A's optimal p_1?

STEP Use the scroll bar to manipulate p_1. Keep an eye on A's utility. Can you find the value of p_1 that maximizes A's utility?

You cannot beat $p_1 = 2$. This is the optimal solution. This is what A will charge B for x_1. At this price for good 1, B wants to have 10 and 20 units of goods 1 and 2. B will buy 5 units of x_1 (adding this to the initial endowment of 5 units) financed from the sale 10 units of x_2. A ends up with 30 and 20 units of goods 1 and 2. A sells 5 of her initial endowment of 35 units of x_1 for \$2/unit and buys 10 units of good 2. The plans match and we are at a stable position.

You can also find this answer with Solver.

STEP Click the scroll bar so p_1 is not equal to 2 and run Solver.

Notice that the changing cell is B9, which is the cell connected to the scroll bar. Solver does not need a constraint because the sheet is set up so that B optimizes based on p_1 and then A's x_1 and x_2 are the total units available for each good minus B's optimal decision. Thus, B's offer curve has been included in A's optimization problem.

In addition, you could use analytical methods, using A's utility as the objective function and B's offer curve as the constraint. All of these methods give the same answer—A's utility maximizing p_1 is 2.

The monopoly solution is displayed in Figure 18.12. Notice that A's indifference curve is tangent to B's offer curve. This is how a monopolist maximizes utility.

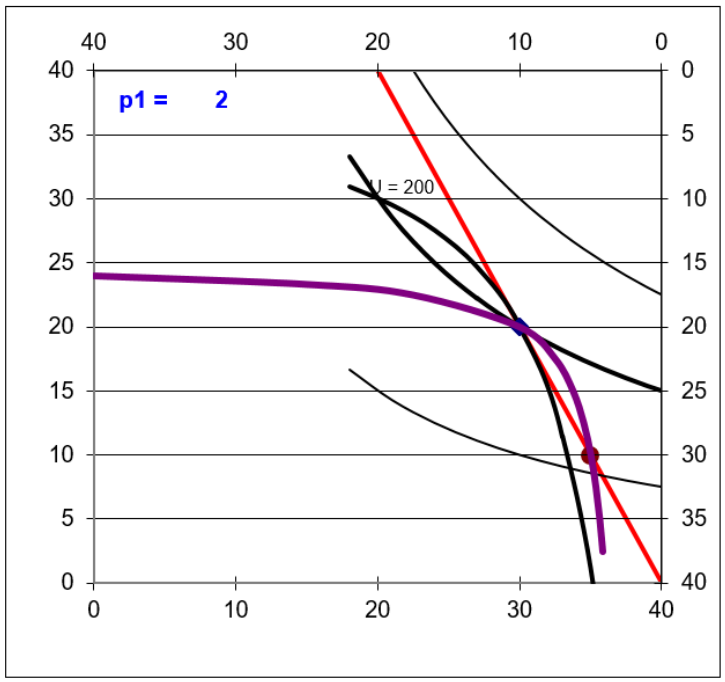


Figure 18.12: Monopoly's optimal solution.
Source: EdgeworthBoxMonopoly.xls!EdgeworthBox with $p_1 = 2$.

Judging Monopoly

What can we say about the monopoly allocation? With Pareto's criteria we can instantly proclaim: Monopoly is not Pareto Optimal.

Figure 18.12 shows that the monopoly allocation is at a point (from A's view it is coordinate 30,20) where the $MRS_A \neq MRS_B$ because the indifference curves intersect. This means that there are Pareto Superior points to the monopoly allocation. It also means that the monopoly allocation is not on the contract curve.

By moving northwest, into the lens created by the two indifference curves at the monopoly solution, an omniscient, omnipotent social planner could make both A and B better off.

Why doesn't A do this? Because all A can do is set the price of good 1 and with this monopoly power, she must charge the same price for all the units sold. This leads to the allocation in Figure 18.12.

If A could perfectly price discriminate, charging different prices for different units, we would get a different result. A could sell the first unit of x_1 at a high price and decrease the price as B purchased more units. As explained in the chapter on monopoly in a partial equilibrium setting, this is called perfect price discrimination. The Q&A sheet asks you to work out the welfare implications of this type of monopoly in a general equilibrium analysis. The welfare results for perfect price discrimination in partial and general equilibrium are the same.

Unlike partial equilibrium, we report no deadweight loss measure in this pure exchange, general equilibrium analysis. We simply note that the monopoly allocation is not Pareto Optimal and this is enough to doom monopoly because we know there are Pareto Superior allocations to the monopoly result.

We cannot say how much damage the inefficiency of monopoly causes because utility can only be measured ordinally. We cannot express, in utils or dollars, the wasted value from monopoly, but we know it is there. Once we say that there are Pareto Superior points, we stamp monopoly as a poor allocation mechanism.

Monopoly is not Pareto Optimal

We found, as we did with partial equilibrium analysis, that monopoly is inefficient. This time, however, we used a general equilibrium analysis that adhered to the strict limitations imposed by ordinal utility. Thus, this analysis is theoretically sound.

In a pure exchange Edgeworth Box, if one agent is granted monopoly power, he or she will choose a price to maximize his or her utility. This does not generate a Pareto Optimal allocation. The monopolist is not interested in Pareto optimality—she simply wants to maximize her own utility.

Recall, however, that this is simply a pure exchange economy. A true general equilibrium model must include production of goods and services and then combine production and exchange. This is beyond the scope of this book. The monopoly result stays the same; however, it still fails to yield a Pareto Optimal allocation.

Exercises

1. Is the monopoly solution better than the initial endowment? Explain.

   Hint: Use Figure 18.12 as a reference.

2. Suppose A really liked x_1, so that cA (cell B21) was 2. How would this change A's utility maximizing price of x_1? What is the monopoly solution? Describe your procedure.

3. In the previous chapter, we used a supply and demand (partial equilibrium) analysis to show that price ceilings in a competitive market cause an inefficient allocation of resources. Use Word's Drawing Tools to create an Edgeworth Box with a price ceiling on x_1. Explain why price ceilings are undesirable in this general equilibrium setting.

References

The epigraph is from page 94 of Gerard Debreu's *Theory of Value: An Axiomatic Analysis of Economic Equilibrium* (originally published in 1959). Debreu won the Nobel Prize in Economics in 1983 "for having incorporated new analytical methods into economic theory and for his rigorous reformulation of the theory of general equilibrium" (www.nobelprize.org/prizes/economic-sciences/1983/summary).

The Nobel Prize web site explains Debreu's contribution in more detail, of course, but for the real scoop, consider this excerpt from E. Roy Weintraub's *How Economics Became a Mathematical Science* (published in 2002):

> While it was the case that most economists would have been unfamiliar at that time with the novel tools of set theory, fixed point theorems, and partial preorderings, there was something else that would have taken them by surprise: a certain take-no-prisoners attitude when it came to specifying the "economic" content of the exercise. Although there had been quantum leaps

> of mathematical sophistication before in the history of economics, there had never been anything like this (p. 114).

Weintraub reports that he had better luck interviewing Debreu than did George Feiwel, who prefaced many of his questions with, "For the benefit of the uneducated." When Feiwel asked why existence of an equilibrium solution is so important, "Debreu shot back, 'Since I have not seen your question discussed in the terms I would like to use, I will not give you a concise answer'" (Weintraub, p. 113). In addition to providing an entire transcript of the interview, Weintraub explains how Debreu led a wave of mathematical formalism into economics in the 1950s.

The general equilibrium ideas you have encountered in this book are a mathematical step below the more formal, axiomatic exposition of General Equilibrium Theory developed in the 1950s and used in graduate economics courses. Pick up Debreu's *Theory of Value* or a modern, PhD-level Micro Theory text (such as David M. Kreps, *A Course in Microeconomic Theory*, 1990) to see exactly what a formal, axiomatic exposition of general equilibrium entails.

Part IV

Conclusion

> But when time and the means for achieving ends are limited and capable of alternative application, and the ends are capable of being distinguished in order of importance, then behaviour necessarily assumes the form of choice. Every act which involves time and scarce means for the achievement of one end involves the relinquishment of their use for the achievement of another. It has an economic aspect . . . Here, then, is the unity of the subject of Economic Science, the forms assumed by human behaviour in disposing of scarce means.
>
> Lionel Robbins

Conclusion

Throughout this book, Excel has been used to solve optimization problems and equilibrium models. Repeated emphasis has been placed on comparative statics and elasticity.

This conclusion has three parts:

1. Excel's Solver: There is a review of basic Solver skills with emphasis on the lesson that Solver is not perfect.
2. Overall view: A quick tour of the topics covered enables a clear statement of the economic way of thinking.
3. An open problem: Markets in a static framework are well understood, but the economic growth generated over time by capitalism is not.

1. Excel's Solver

Consider a perfectly competitive (PC) firm with a total cost function given by $TC = 100q^{\frac{1}{2}}$. Dividing both sides by q gives us the average cost function, $ATC = 100q^{-\frac{1}{2}}$. Taking the derivative of TC with respect to q yields $MC = 50q^{-\frac{1}{2}}$.

If this PC firm faced a market price of \$5/unit, what is the profit-maximizing level of output?

This book has solved optimization problems via numerical and analytical methods. We will apply both methods to this problem. First, we will use Solver.

But we will not use a prepared Excel workbook. Instead, you will create your own implementation of this problem. There are, of course, helpful steps to guide you.

STEP Open a blank Excel workbook. In cell A1, type the word *quantity*. Cell B1 will hold a number that represents the quantity. In cell A2, type the word *profits*. In cell B2, enter the formula for profit.

The price is \$5/unit and $TC = 100q$ so the formula in cell B2 is: = 5*B1 − 100*SQRT(B1).

STEP Run Solver. The target cell is B2, the goal is obviously to maximize profits, and the changing cell is B1. There are no constraints because the PC firm is free to produce as much output as it wants at the given price.

Excel gives a miserable result. Depending on your Solver defaults, it might go negative and, since Excel cannot take the square root of a negative number, it gives up and announces its failure.

If so, make A1 zero and run Solver again, but this time, check the *Make Unconstrained Variables Non-negative* option. Your Solver may be set up so the *Make Unconstrained Variables Non-negative* option was already checked so you might not see the first miserable result.

Starting from zero (or a blank cell) in A1, with the non-negativity constraint, Solver says the answer is zero. This is worrisome. Could the optimal quantity really be zero?

Maybe the issue is that we are starting from blank cell, which is zero. This is poor practice. Excel interprets blanks as a zero and the formula in B1 evaluates to zero. Treating blanks as zero is one of the most dangerous things a spreadsheet does (Google sheets behaves the same way). You should always avoid this.

We can change where Solver starts from to see if that helps.

STEP Change cell B1 to 25. Cell B2 should display −375. Run Solver.

Solver appears convinced that the optimal solution is zero. We turn to analytical methods to see if we can confirm Solver's result.

We know $MC = 50q^{-\frac{1}{2}}$ and since it is a PC firm $MR = P$ so $MR = 5$. We can set $MR = MC$ and solve for optimal q.

$$5 = 50q^{-\frac{1}{2}} \rightarrow q^{\frac{1}{2}} = 10 \rightarrow q^* = 100$$

This is confusing. We now have two answers: $q = 0$ and $q = 100$. Which one is correct?

Maybe a graph will help. We can draw the canonical graph of the firm's output profit maximization problem. Figure IV.1 shows the cost curves and we can clearly see that $MR = MC$ yields a negative profit rectangle.

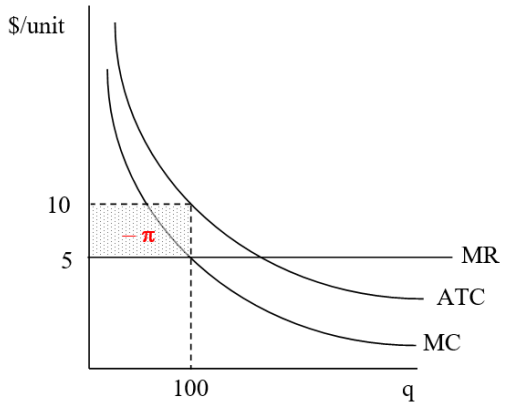


Figure IV.1: The firm at $q = 100$ where $MR = MC$.

This graph helps explain what is going on here, but we need a better visual. This book claimed that looking directly at the profit function made clear the Shutdown Rule so let's try that approach.

STEP Create a column from 0 to 500 by 10. This is the quantity. Use the profit formula to create a column for profit based on the quantity. Create a graph of the two columns.

If you get stuck, this 2-minute video at vimeo.com/425873093 shows how to do it.

Figure IV.2 shows the graph made in the video. It makes clear that the point where $MR = MC$ is actually a point of minimum profit. Although the first-order condition is met (we did find a flat spot on the profit function at $q = 100$), this solution fails the second-order condition for a maximum.

Thus, the correct answer is to produce an infinity of output. Profits rise as more is produced past 100 units of output. Higher output leading to greater profit continues forever so the optimal solution is infinity.

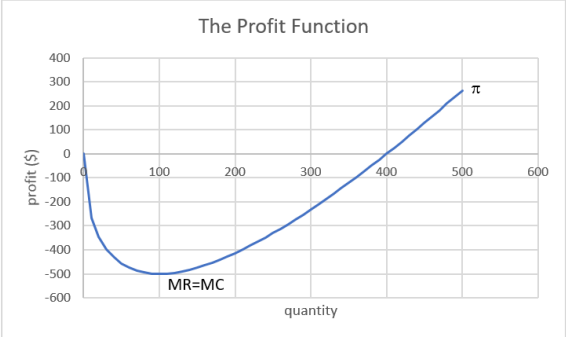


Figure IV.2: The profit function shows that optimal q is infinity.

How can we explain Solver's answer of zero? Why doesn't it give us the correct answer? When Solver starts from below 100 (we started from zero and 25), it goes to zero (or negative output if you do not have a non-negativity constraint). What happens if it starts from a number greater than 100?

STEP Enter 110 in cell A1 and run Solver.

Solver reports that "Objective cells do not converge." Is this a miserable result? No, actually, it is the correct answer! When Solver starts from more than 100, it goes right on the x axis and profits rise and it keeps going and going. As we know, this is the right answer.

It is worth remembering that Solver's algorithm is naive. It evaluates the function at the starting value, then moves left and right. The size of the move depends on the numerical values in the problem. Starting from $q = 25$, for example, Solver moves a little bit right, sees that profits fell, then goes in the opposite direction and lowers output. You can see Solver's steps by checking the *Show Iteration Results* option after clicking the *Options* button in the Solver dialog box.

You might be thinking that since we are in the long run, $ATC = AVC$ and it is clear that $P < AVC$ at $MR = MC$, which means the firm should shut down. That is not bad thinking, except the rule does not work at $MR = MC$ in this case because that is not the profit-maximizing output.

The takeaway of this final example is that you have to know what you are doing with Solver. It is not perfect and you cannot blindly rely on its results. This example shows that numerical methods are to be used with caution. Be careful out there.

2. Overall View

This book covered modern-day, orthodox microeconomic theory at the college undergraduate level. It used Excel to present difficult material and showed how mathematics can be used to solve problems in economics.

The economic approach or economic way of thinking provided the framework for analyzing observed behavior. The basic idea is to set up and solve an optimization problem or equilibrium model. Next, a single variable is changed, ceteris paribus, and the new solution is compared to the initial solution. This procedure is called comparative statics. Elasticity captures the logic of comparative statics in a single number.

When the economic approach is applied to consumers, it is called the Theory of Consumer Behavior. The key comparative statics analysis is deriving the demand curve. Figure IV.3 is a canonical graph of deriving demand.

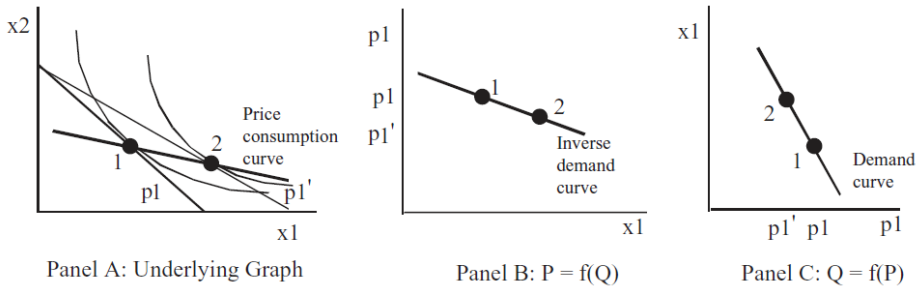


Figure IV.3: Deriving the demand curve.

When the economic approach is applied to producers, it is called the Theory of the Firm. The key comparative statics analysis is deriving the supply curve. Figure IV.4 is a canonical graph of deriving supply.

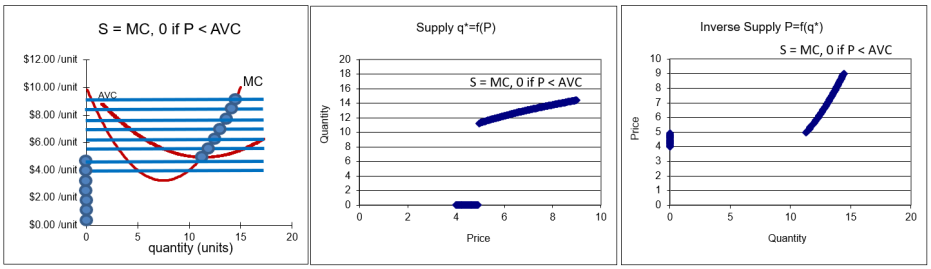


Figure IV.4: Deriving the supply curve.

The firm is more complicated than the consumer because firms hire inputs to produce output. In fact, the firm is really a set of three interrelated optimization problems: input cost minimization, output profit maximization, and input profit maximization.

The individual demand and supply curves derived from the consumer and firm models can be added up to produce market demand and supply curves. This enables a partial equilibrium analysis of how markets solve society's resource allocation question. Figure IV.5 shows supply and demand flanked by its consumer and firm source graphs.

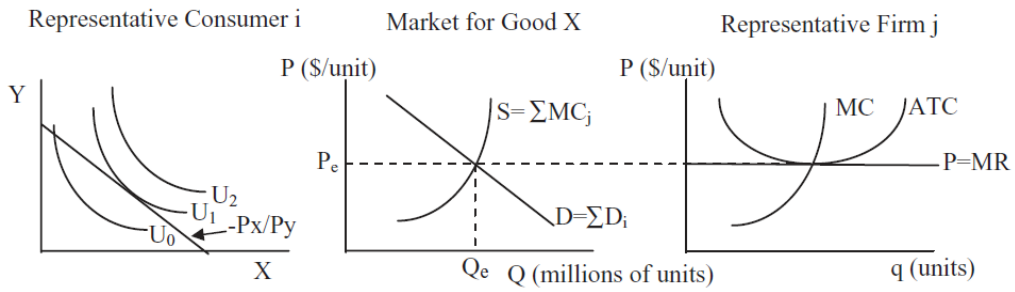


Figure IV.5: The market's resource allocation solution for one good.

Price ceilings, taxes, monopoly, import quotas, and externalities are all examples of situations where we have a misallocation of resources in a single market.

Partial equilibrium enables calculation of a measure of inefficiency called deadweight loss (also known as the Harberger triangle), but this should be interpreted as an approximation because consumers' surplus requires that an adjustment be made to the ordinary demand curve (compensated demand must be used) and the effects on other markets are ignored. Partial equilibrium analysis is commonly used in empirical work. Think of deadweight loss as a rough measure of inefficiency in the allocation of resources.

General equilibrium is a more rigorous and sophisticated analysis because it looks at all markets as a total system. Pareto's criteria show that a properly functioning market yields an optimal allocation and monopoly is not Pareto optimal. Figure IV.6 is the canonical graph of a market's general equilibrium and it makes clear that the market's allocation has no Pareto Superior points.

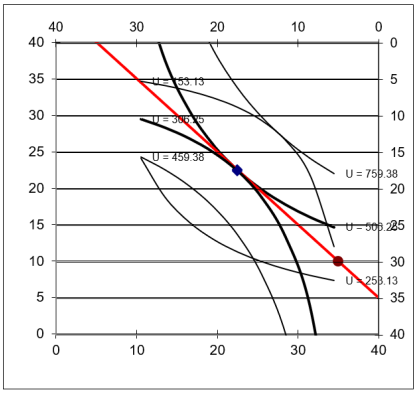


Figure IV.6: The market's allocation in an Edgeworth Box.

General equilibrium does not suffer from the same problems as partial equilibrium, but it is much harder to implement in the real world. In the epigraph to the section introducing the Edgeworth Box, mention was made of computable general equilibrium models. This shows that there is an empirical side to general equilibrium analysis, but it is a relatively modern development.

It is reasonable to view mainstream microeconomics as a theory of the price mechanism. The market system uses prices as signals to allocate resources. Optimizing agents react to price changes and their interactions as buyers and sellers drive the system toward equilibrium. The Theories of Consumer Behavior and the Firm are stepping stones that explain how the market answers society's resource allocation question. Figure IV.5 puts the Theory of Consumer Behavior, Theory of the Firm, and partial equilibrium analysis together. These three graphs and how they fit together are worth remembering.

Another organization of microeconomics splits it into two parts—individual agents (consumers and firms) that optimize and what happens when these optimizing agents interact in a market. The former is about optimization and the latter is about equilibrium. The order that is spontaneously generated by interacting, optimizing agents is a remarkable result. Economists see supply and demand not as the simple intersection of two lines, but as a pattern that is unwittingly generated by the agents themselves—just like geese that fly in a V.

This book was designed to provide you with practice in applying the economic approach. We tackled unconstrained and constrained optimization

problems, computed many different elasticities, and solved several equilibrium models at the partial and general levels.

The many applications of the economic approach demonstrate its remarkable flexibility. The Theory of Consumer Behavior, at first, seems ridiculously unrealistic—a robot consumer chooses between two goods with prices, tastes, and income given! But that is just the basic model. By changing the goods to consumption in the present and the future, it becomes an intertemporal choice model. We analyzed charitable giving, portfolio theory, and the effect of safety features in automobiles with the Theory of Consumer Behavior.

In every application, the economic way of thinking was prominent. We set up and solved an optimization problem, then changed a variable, ceteris paribus, to see how the optimal solution changed. There are countless applications of the economic approach, but they share the same framework and logic.

In fact, the economic approach is what defines economics today. It may be the only discipline that defines itself by a methodology instead of by what it studies. Most people have a content-based definition of economics: They think that the study of interest rates, unemployment, and money is economics. But this is wrong. The proper definition of economics is the application of the economic approach to explain observed behavior. Crime, marriage, and war, if analyzed with the economic approach, fall under the heading of economics.

From now on, when you hear the phrase "an economic analysis of," you will know that the economic approach is about to be applied, you will know what to expect, and you will be comfortable as the speaker talks about constraints, optimality, comparative statics, and elasticity.

3. An Open Problem

Neither this book nor modern, mainstream economics explains the dynamic process of capitalism. A few hundred years of the market system make it obvious that creativity, innovation, and technological change are endogenously generated by market-based societies. No one really knows why.

The question has been with economics since the very beginning. Many people know that Adam Smith wrote a book called the *Wealth of Nations*, but only a few know that the actual title is, *An Inquiry into the Nature of Causes of the Wealth of Nations*. But what was Smith's inquiry, simply put?

He wanted to know why England was so much richer than its neighbors. In 1776, Smith could see British wealth all around him. He could see the economy taking off and he wondered why some places develop and grow, while others cannot seem to do so? This question remains unanswered and, in the language of mathematics, it is the biggest open problem in economics.

Explaining the dynamism of the market system is a much different question than the static optimization and equilibrium models that explain why markets allocate resources efficiently. In the static world, there are no new products, cost-saving innovations, or new firms. The static world is stable and markets are in equilibrium.

This static model clashes violently with reality. Joseph Schumpeter's portrayal of what he called plausible (i.e., real-world) capitalism, captured in the oxymoron "creative destruction," highlights the rise and fall of firms, explosive growth, and dislocation produced by markets. For Schumpeter, the driving force is the entrepreneur, a hero whose desire to dominate the business world results in economic success for society. But Schumpeter's story (best captured in *Capitalism, Socialism and Democracy*, originally published in 1942), thrilling though it may be, is not part of mainstream economics today.

It is plainly clear that markets do generate spectacular economic growth, unparalleled by any other organizational form. Even the harshest critics of capitalism concede this point:

> The bourgeoisie, during its rule of scarce one hundred years, has created more massive and more colossal productive forces than have all preceding generations together. Subjection of Nature's forces to man, machinery, application of chemistry to industry and agriculture, steam-navigation, railways, electric telegraphs, clearing of whole continents for cultivation, canalisation of rivers, whole populations con-jured out of the ground—what earlier century had even a presentiment that such productive forces slumbered in the lap of social labour?

That was written by Karl Marx and Friedrich Engels in *The Communist Manifesto* in 1848, available at www.marxists.org/archive/marx/works/download/pdf/Manifesto.pdf.

Marx and Engels argued capitalism will self-destruct, but not because it failed to make goods and services. They thought it was the most productive system ever devised. They were amazed by capitalism's ability to generate output.

Marx and Engels were not the first nor the last to be awed by the productive power of the market system. Yet, even though we can easily see that productive power, we simply do not know the answer to basic questions about how markets generate growth. Beyond superficial generalities about the institutional environment, such as needing rule of law and established property rights, we have no explanation for how the interaction of multitudes of agents drives the system over time. We cannot even answer the most basic question, posed by Adam Smith—why are some countries rich and others poor?

If we knew how and why markets caused technological change and output per person to grow exponentially, we would know how to help those societies mired in poverty. Nobel Prize winning economist Robert Lucas poses the issue this way:

> Is there some action a government of India could take that would lead India's economy to grow like Indonesia's or Egypt's? If so, what, exactly? If not, what is it about 'the nature of India' that makes it so? The consequences for human welfare involved in questions like these are simply staggering: Once one starts to think about them, it is hard to think about anything else. (Lucas, 1988, p. 5)

The point is this: Markets can be analyzed from static and dynamic perspectives. The former focuses on resource allocation at a single moment in time. It freezes the movie and asks how markets work in this motionless environment. We know how markets work as a resource allocation mechanism.

The latter perspective is about the dynamic nature of markets, we want to know how markets work over time. The movie runs—spurts of rapid growth are followed by recessions, then more growth, but output per person trends upward. Will this continue? We do not know. How do the institutions we rely on (including property rights) emerge from the interaction of optimizing agents? We do not know.

Explaining markets as a dynamic process remains the most important open problem in economics. Perhaps you can work on it.

References

The epigraph is from pages 14 and 15 of Lionel Robbins, *An Essay on the Nature and Significance of Economic Science* (originally published in 1932) and available online at mises.org/library/essay-nature-and-significance-economic-science.

We began with a famous quotation from Robbins, defining economics as "the science which studies human behavior as a relationship between given ends and scarce means which have alternative uses." This book takes this definition seriously and has stressed static optimization, but this last chapter makes clear that we have a great deal to discover and learn about dynamics and technological progress.

Robert Lucas, "On the Mechanics of Economic Development," *Journal of Monetary Economics*, Vol. 22 (1988), pp. 3–42, www.sciencedirect.com/science/article/abs/pii/0304393288901687.

The fact that perfect competition is incompatible with increasing returns (as the Solver example with $TC = 100q^{1/2}$ showed) led to a heated debate in the 1920s. Economics continues to struggle to develop a model that combines the fact that average cost falls as output rises for many products with competitive markets. See David Warsh (2006), *Knowledge and the Wealth of Nations: A Story of Economic Discovery*, for a review of how economics has grappled with the issue of increasing returns.

If you are interested in the trajectory of capitalism and markets, then modern economic theory will not be of much help. For an entertaining review of capitalism and how it has been treated in economics, no one has beaten this classic: Robert Heilbroner, *The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers* (New York: Touchstone, 1999, 7th edition, originally published 1953).

INTERMEDIATE MICROECONOMICS WITH MICROSOFT EXCEL®

This unique text uses Microsoft Excel® workbooks to instruct students. In addition to explaining fundamental concepts in microeconomic theory, readers acquire a great deal of sophisticated Excel skills and gain the practical mathematics needed to succeed in advanced courses. In addition to the innovative pedagogical approach, the book features explicitly repeated use of a single central methodology, the economic approach. Students learn how economists think and how to think like an economist. With concrete, numerical examples and novel, engaging applications, interest for readers remains high as live graphs and data respond to manipulation by the user. Finally, clear writing and active learning are features sure to appeal to modern practitioners and their students. The website accompanying the text is found at www.depauw.edu/learn/microexcel.

Humberto Barreto is Professor of Econmics and Management at DePauw University. He earned his Ph.D. from the University of North Carolina at Chapel Hill. Professor Barreto has lectured around the world on teaching economics with computer-based methods, including Cuba, Brazil, Canada, Scotland, Spain, Poland, India, Burma, Japan, and Taiwan. He was a Fulbright Scholar in the Dominican Republic and has taught National Science Foundation (NSF) Chautauqua short courses using simulation. He has received several research and teaching awards. His book, *The Entrepreneur in Microeconomic Theory*, was translated into Arabic in 1999. He has written numerous articles and books on using Excel to teach economics (including introductory level material, micro, macro and econometrics). He offers an annual workshop for faculty on teaching economics. Visit his website for more information: Teaching Economics with Excel.