Kalina, Jan

Article

Regression quantiles under heteroscedasticity and multicollinearity

Regression Quantiles under Heteroscedasticity and Multicollinearity: Analysis of Travel and Tourism Competitiveness[1]

Jan KALINA – Petra VAŠANIČOVÁ** – Eva LITAVCOVÁ***

Abstract

In the linear regression, heteroscedasticity and multicollinearity can be characterized as intertwined problems, which often simultaneously appear in econometric models. The aim of this paper is to discuss various approaches to regression modelling for heteroscedastic multicollinear data. A real economic dataset from the World Economic Forum serves as an illustration of various individual methods and the paper provides a practical motivation for quantile regression and particularly for regularized regression quantiles. In the dataset, tourist service infrastructure across 141 countries is modeled as a response of 12 characteristics of the Travel and Tourism Competitiveness Index (TTCI). Regression quantiles and their lasso estimates turn out to be more suitable for the dataset compared to more traditional econometric tools.

Keywords: *linear regression, model selection, robustness, regression quantiles, lasso, tourism*

JEL Classification: C21, C13, C14, Z32

* Jan KALINA, Institute of Computer Science of the CAS, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic and Institute of Information Theory and Automation of the CAS, Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic; e-mail: kalina@cs.cas.cz

** Petra VAŠANIČOVÁ – Eva LITAVCOVÁ, University of Prešov, Faculty of Management, Department of Mathematical Methods and Managerial Informatics, Konštantínova 16, 080 01 Prešov, Slovak Republic; e-mail: petra.vasanicova@gmail.com; eva.litavcova@unipo.sk

[1] The research was supported by the Czech Science Foundation project *Nonparametric (Statistical) Methods in Modern Econometric*s No. 17-07384S and by the Slovak Scientific Grant Agency VEGA project *Economic Activity of Tourism in European Space* No. 1/0470/18.

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-17-0166.

The authors would like to thank two anonymous referees for their constructive advice.

Introduction and Motivation

Heteroscedasticity is one of the crucial assumptions of the standard linear regression model. Although it does not have a negative influence on the least squares (LS) estimate of β and prediction of the response, it may strongly affect confidence intervals and significance tests. Common heteroscedasticity tests implicitly require a proper set of relevant variables; otherwise, their power decreases if redundant independent variables are present in the model. Nevertheless, practitioners often perform heteroscedasticity testing over all available data, without reducing the dataset to relevant variables (Jurczyk, 2012).

On the other hand, standard variable selection tools by *t*-tests assume homoscedasticity. Although not admitted in standard textbooks on regression (Matloff, 2017; Young, 2017), it is intuitively clear that alternative approaches (diagnostic tests or estimation procedures) are desirable for economic data under heteroscedasticity and (at the same time) presence of redundant variables. A rare example of such practical tools based on the least squares is the biased estimator of Alheety and Kibria (2009), which combines the benefit of ridge regression with a shrinkage estimator of Stein (1956). Regression quantiles (RQ) represent an important class of modelling tools suitable under both heteroscedasticity and multicollinearity with a potential to be successful in econometric applications (Fitzenberger, Koenker and Machado, 2002), however commonly used mainly for estimation, while the whole spectrum of corresponding hypothesis tests or tools for variable selection (lasso regression quantiles) remains underexploited (Koenker et al., 2017).

The aim of this paper is to discuss various approaches to regression modelling for heteroscedastic multicollinear data. As an illustration, we present here a complex analysis of a real economic dataset connected with tourism destination competitiveness, which deals with serious modelling issues, such as multicollinearity and heteroscedasticity (Dlouhý and Flusserová, 2007).

Taking into account the current economic processes, tourism is currently regarded as a global phenomenon and as the fastest growing world industry. According to World Tourism Organization (UNWTO), the growth of this industry has been showing its resilience to global geopolitical and economic instability for the sixth consecutive year, so its importance in the economy is evident. From the data available for 2017 (WTTC, 2018), the tourism industry contributed 8.27 trillion U.S. dollars to the global economy, representing 10.4% of the world GDP. In addition, it has created 313 million jobs, representing 1 out of every 10 jobs on the planet. Tourism turns out to be a major driving force behind economic growth and employment. Because globalization increases the level of competition, national governments have to approach the development of tourism

with a stronger emphasis (Štefko, Kiráľová and Mudrík, 2015). Unleashing new growth potential of industry within a given country also requires to enhance its competitiveness (Jenčová, 2018).

Various standard or more recent regression estimates and corresponding tests are applied in this paper to a real dataset from tourism management area, where the task is to model tourist service infrastructure in 141 economies of the world as a response of 12 characteristics of the Travel and Tourism Competitiveness Index (TTCI). Over the last two decades, the competition among destinations keeps increasing and there has been a growing need to acquire knowledge about a destination's competitive ability (Pulido-Fernández and Rodríguez-Díaz, 2016). Since 2007, the World Economic Forum has studied national competitiveness in the travel and tourism industry and has published reports allowing for cross--country comparisons of travel and tourism competitiveness.

The importance of studying the competitiveness of tourism destinations was stressed by Bucher (2015), who evaluated it by means of a comprehensive competitiveness index.

This paper has the following structure. Section 1 presents a brief overview of various approaches to modelling heteroscedasticity in linear regression, especially if there are redundant variables present in the model. In Section 2, individual methods are applied to analyze the TTCI dataset, bringing arguments in favor of regression quantiles and mainly their lasso estimators. As discussed in the Conclusions, the results may be useful for tourism policy decision making, or for management of other industries and business entities.

1. Regression Methodology: Heteroscedasticity and Redundant Regressors

In this section, various approaches to heteroscedasticity estimation as well as subsequent testing and variable selection are discussed, especially if the regression model contains also redundant variables not contributing to the variability of the response. We consider the standard linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + e_i, i = 1, \ldots, n \tag{1}$$

where $Y_1, \ldots, Y_n$ are values of a continuous response variable and $e = (e_1, \ldots, e_n)^T$ is the vector of random errors (disturbances). The least squares estimator of the parameters β of interest will be denoted by $b_{LS} = (b_0^{LS}, b_1^{LS}, \ldots, b_p^{LS})^T$.

Variable selection for the least squares is commonly performed by stepwise procedures, with the backward selection being the most prominent method in

both statistics and data mining. Nevertheless, common stepwise variable selection approaches, based on a repeatedly used standard *t*-test, may inappropriately focus on too simple models. Their size also has a tendency to exceed 5% due to multiple testing, and the tests may suffer from bias or from violations of homoscedasticity (Whittingham et al., 2006). There are a plethora of tests of the null hypothesis H_0 of homoscedasticity against various (either generally or more specifically formulated) forms of an alternative hypothesis. Prominent examples include tests of Breusch-Pagan, Goldfeld-Quandt, or White's general (non-constructive) test. However, their power depends heavily on the sample size and becomes smaller under multicollinearity in (1).

Also, a small deviation from normality of disturbances, which is usual in real economic data, may negatively affect analyzing the model (1). Without the normality of disturbances, the least squares estimator b_{LS} has the optimum property only within the family of linear unbiased estimators; typically, a better estimator can be found in the much wider class of biased estimators, if the normality does not hold.

Another complication is a lack of reliable guidelines for estimation if heteroscedasticity turns out to be significant. One common possibility for removing heteroscedasticity is to use Aitken estimator (also generalized least squares, heteroscedastic regression), i.e., to transform the model by introducing weights to individual observations and to apply weighted least squares (Greene, 2011). The result of the auxiliary model, which must be specified by the user, depends heavily on its ability to explain heteroscedasticity. Even if homoscedasticity is not rejected in the transformed model, it is not guaranteed that heteroscedasticity is removed completely. The transform also loses the interpretability of the original model.

White's heteroscedasticity-robust estimator of $var\, b_{LS}$ (White, 1980) is consistent under heteroscedasticity without any transform of the original model (1). Although it is commonly used routinely also under homoscedasticity, it has a lower efficiency in such a case. A significance test for individual regressors, based on White's standard errors, is rather conservative, i.e., rejects less often. This is especially true if the sample size is not large (Wooldridge, 2013; Stock and Watson, 2015). It is due to the fact that corresponding *t*-values are not distributed according to the Student's *t*-distribution, because we cannot prove the residual sum of squares corresponding to b_{LS} to have a χ^2-distribution. In other words, it is not possible to adapt the elegant trick previously used by Ronald A. Fisher, who considered a sum of squared normal random variables to obtain a χ^2-distribution with corresponding degrees of freedom, valid only under the important assumption of the same variances for the random variables (Fisher,

1920; Fisher, 1922). In addition, White's estimator does not give the answer to the basic question if there is heteroscedasticity in the data or not.

Under multicollinearity, it is common to replace the least squares by ridge regression or lasso (least absolute shrinkage and selection operator) estimator, where the latter

$$\widehat{\beta}_j = sgn\left(b_j^{LS}\right)\left(\left|b_j^{LS}\right| - \lambda\right), j = 1,\ldots, p \tag{2}$$

with some value of the regularization parameter $\lambda > 0$ will be denoted as LS--lasso to stress that it is based on least squares. The regularization by shrinking coefficients in (2) towards 0 stems from the original idea of Stein (1956). LS--lasso allows for an intrinsic variable selection, is highly stable (robust) to small changes in data, and is especially suitable for correlated regressors (Kalina, 2014). However, LS-lasso is suitable for homoscedastic variables, and its variable selection suffers under heteroscedasticity (Jia, Rohe and Yu, 2013). Ridge regression does not contain variable selection, and its interpretability is thus much less accessible (Hastie, Tibshirani and Friedman, 2009).

If data in (1) are contaminated by outliers, it is advisable to consider a (highly) robust estimator of β. The least weighted squares (LWS) estimator b_{LWS} of Víšek (2011) possesses appealing properties such as regression-equivariance or high breakdown point. We consider variable selection for the LWS estimator in the form of a backward stepwise procedure, which is analogous to the least squares; testing is performed by means of nonparametric bootstrap confidence intervals for $var\, b_{LWS}$ here, rather than by *t*-tests. There are also recently proposed tests of heteroscedasticity for the LWS (Kalina, 2012), which contributed to over-optimistic expectations that heteroscedasticity does not represent a potential danger to linear regression anymore. Nevertheless, both asymptotic and permutation-based tests following the ideas of Nyblom (2015) remain vulnerable to multicollinearity just like for the least squares, and it also remains difficult to remove heteroscedasticity if these tests are significant.

White's estimator for the LWS was derived by Víšek (2010, p. 43), whose formula (23) for the heteroscedasticity-consistent estimator, formulated for a more general situation with instrumental variables, should be corrected by the proper form of $var\, b_{LWS}$

$$\begin{aligned} &\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} u_i^2\left(b_{LWS}\right) X_i X_i^T\right]\left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right]^{-1} = \\ &=\left[\sum_{i=1}^{n} X_i X_i^T\right]^{-1}\left[\sum_{i=1}^{n} u_i^2\left(b_{LWS}\right) X_i X_i^T\right]\left[\sum_{i=1}^{n} X_i X_i^T\right]^{-1} \end{aligned} \tag{3}$$

where $X_i = (X_{i1},\ldots,X_{ip})^T$ and $u_i\left(b_{LWS}\right)$ are residuals corresponding to the LWS estimator. The steps of the proof of Víšek (2010) are valid showing that (3) converges in probability to the true covariance matrix of b_{LWS}, if the number of observations goes to infinity and if (technical) assumptions are fulfilled. It is evident, although not sufficiently investigated in the literature, that White's estimator may be mislead by the presence of redundant variables in the model. This follows from the fact that the dimensionality of $\sum_{i=1}^{n} X_i X_i^T / n$ gets larger with an increasing number of redundant variables and thus this matrix, which must be inverted, becomes likely to be ill-conditioned.

Regression quantiles are reliable modelling tools under heteroscedasticity with a clear interpretation, if the whole set of regression τ-quantiles is evaluated for various values $\tau \in (0, 1)$. They are able to capture the whole distribution of the response instead of simply considering the mean trend. In the literature (including the seminal book by Koenker (2005)), regression quantiles are typically performed on simplistic data with a very small number of regressors, which do not actually require regression quantiles, because simple graphical illustrations give insight to such data as well.

We will use three different tests for regression quantiles in Section 2:

a) Tests of significance of individual regressors (variable selection) based on regression rank scores (RRS).

b) The test of equality of regression parameters for various (two or more) values of τ based again on RRS.

c) The Khmaladze test, originally proposed by Koenker and Xiao (2002), of equality of regression parameters for the whole range of τ (i.e., across all τ). It involves estimating nuisance parameters and is only asymptotically distribution-free, based on asymptotics of empirical processes.

The first two tests and corresponding confidence intervals exploit general results of Gutenbrunner et al. (1993) and are distribution-free without a need for a prior estimating of nuisance parameters. As investigated by Jurečková and Navrátil (2014), both are also resistant to local heteroscedasticity. Nevertheless, studies of quality of fit of regression quantiles (Wang, Zhou and Li, 2013; Ranganai, 2016) reveal that these tests are not reliable for variable selection. Backward stepwise variable selection can be performed exploiting these tests but with the same limitations as for the least squares.

Regression quantiles regularized by the lasso penalty (further denoted as RQ--lasso) proposed by Koenker (2005) can be described as an extension of LS-lasso to regression quantiles with an ability to perform variable selection. While theoretical results are available for RQ-lasso, particularly concerning the variable

selection consistency, the method has been rarely applied to economic data (cf. Jiang and Qian, 2016). This motivates us to apply them to real data from the tourism industry in this section.

2. Analysis of Travel and Tourism Competitiveness Data

For comparing the competitiveness of individual countries in the tourism area, the World Economic Forum (WEF) publishes a yearly Travel and Tourism Competitiveness Report (TTCR). Here, we work with the TTCR dataset containing TTCI characteristics in 141 countries of the whole world. Particularly, we model the Tourist Service Infrastructure (TSI) as a response of 12 pillars (indicators) of Table 1. The analysis illustrates the performance of various methods of Section 1 under heteroscedasticity and multicollinearity.

T a b l e 1

List of 12 Regressors of the Travel and Tourism Dataset

Index	Abbreviation	Name of the regressor
1	BE	Business Environment
2	SS	Safety and Security
3	HH	Health and Hygiene
4	HRLM	Human Resources and Labour Market
5	ICT	Information and Communication Technologies Readiness
6	TT	Prioritization of Travel and Tourism
7	IO	International Openness
8	PC	Price Competitiveness
9	ES	Environmental Sustainability
10	ATI	Air Transport Infrastructure
11	GPI	Ground and Port Infrastructure
12	NR	Natural Resources

Source: Crotti and Misrahi (2015).

The dataset obtained from Crotti and Misrahi (2015) contains 14 pillars (variables), which characterize areas that impact travel and tourism competitiveness. Raw data, acquired by the WEF in an opinion survey, are normalized to a common 1-to-7 scale, while the overall TTCI is calculated by gradually aggregating the results of the individual pillars using a simple average. From all available variables, we omitted the last pillar. Let us interpret the 12 pillars under consideration, using the notation of Table 1.

- Pillars 1 – 5, directly linked to economic growth and important for business development, create the first subindex of TTCI denoted as Enabling Environment.
- Pillars 6 – 9, which are more sector-specific, create a subindex Travel and Tourism Policy and Enabling Conditions.

- Pillars 10 – 11, together with TSI, are directly related to travel and tourism infrastructure and create the Infrastructure subindex.
- Pillar 12, together with the omitted pillar (Cultural Resources and Business Travel), create the fourth subindex denoted as Natural and Cultural Resources.

TSI corresponds to the level of tourism service infrastructure evaluated by means of the number of "upper-level" hotel rooms complemented by the extent of access to services such as car rentals and automated teller machines. The idea is that the availability of sufficient quality accommodation, resorts, and entertainment facilities may represent a significant competitive advantage (Crotti and Misrahi, 2015). Our analysis represents a unique application within the tourism management area, where regression quantiles seem to have been applied only to tourist spending so far (Lew and Ng, 2011; Rudkin and Sharma, 2017). A preliminary analysis of this data was presented by Vašaničová et al. (2017), however with different aims to investigate various relationships among individual variables.

The regression task in the example is to find a suitable model with regressors relevant for explaining TSI. This also required to decide whether the relationship is heteroscedastic or not. The response is a continuous variable, just like all the regressors. The data seem to contain neither missing values nor severe outliers. The regressors suffer from heavy multicollinearity with a condition number (i.e., ratio of the largest to the smallest singular value of $X^T X$) equal to 3443, where X denotes the design matrix with elements X_{ij} with $i = 1,\ldots,n$ and $j = 1,\ldots p$. The principal component analysis does not reveal any interesting structure in X, while the contribution of all regressors to the first principal components is basically equal. Our following computations are performed in R software, using several additional packages (glmnet, het.test, Qtools, quantreg, sandwich).

We performed a backward variable selection by means of various methods, which include *t*-tests or tests based on White's estimates of standard errors. Table 2 shows the resulting significant variables arranged according to the significance, starting with variables with the smallest *p*-value. Tests for the LWS estimator yield similar results; however, it is needed to take resort to a nonparametric bootstrap test (Kalina and Peštová, 2017) to replace *t*-tests. Within the LS-lasso estimator, the optimal regularization found by a 10-fold cross-validation in package glmnet is almost negligible ($\lambda = 0.004$), and thus all variables remain to contribute to the model, and no variable selection is performed.

Results of heteroscedasticity tests for the least squares as well as for the LWS estimators for (1) with all 12 regressors are shown in Table 3, where however none of the tests yields a significant result. This is revealed as very misleading in Figure 1 showing the relationship between the response and variables 5, 6 and 10,

which will be later explained to be the crucial regressors in the model. A linear trend with a major heteroscedasticity can be observed in each of the figures, in spite of the results of tests.

T a b l e 2

Variable Selection Applied to Various Estimators with the List of Significant Variables, Starting from the Most Significant to the Least Significant

Estimator	Method	Significant variables
LS	*t*-tests	6, 5, 9, 1, 10
LS	White's estimator	6, 5, 9, 1, 3, 10
LWS	Nonparametric bootstrap	6, 5, 9
LWS	White's estimator	6, 5, 1, 9, 3
LS-lasso (with optimal λ)	Non-zero estimate of β_j	6, 5, 9, 1, 3, 10, 4, 7, 2, 11, 12, 8
Regr. quantile ($\tau = 0.1$)	Test based on RRS (A)	5, 6
Regr. quantile ($\tau = 0.3$)	Test based on RRS (A)	5, 6
Regr. quantile ($\tau = 0.5$)	Test based on RRS (A)	6, 5, 1, 10, 4
Regr. quantile ($\tau = 0.7$)	Test based on RRS (A)	6, 5, 10
Regr. quantile ($\tau = 0.9$)	Test based on RRS (A)	6, 10

Source: Own computation.

T a b l e 3

Results of Heteroscedasticity Tests

Heteroscedasticity test	Degrees of freedom	*p*-value for LS	Asymptotic *p*-value for LWS
Breusch-Pagan on 12 variables	12	0.362	0.401
Breusch-Pagan on the set of regressors {1, 5, 6, 9, 10}	5	0.053	0.063
White's test	20	0.089	0.094

Source: Own computation.

Let us now consider generalized least squares (Aitken estimator) for the model (1) with all 12 variables. This popular heteroscedastic regression approach (see Greene (2011), Sec. 9.3) proceeds in two stages. First, least squares are used to estimate β in (1) and thus to obtain residuals $u_1,\ldots,u_n$. In the second stage, the model

$$\frac{Y_i}{|u_i|} = \frac{\gamma_0}{|u_i|} + \frac{\gamma_1 X_{i1}}{|u_i|} \ldots + \frac{\gamma_p X_{ip}}{|u_i|} + v_i, i = 1,\ldots,n \qquad (4)$$

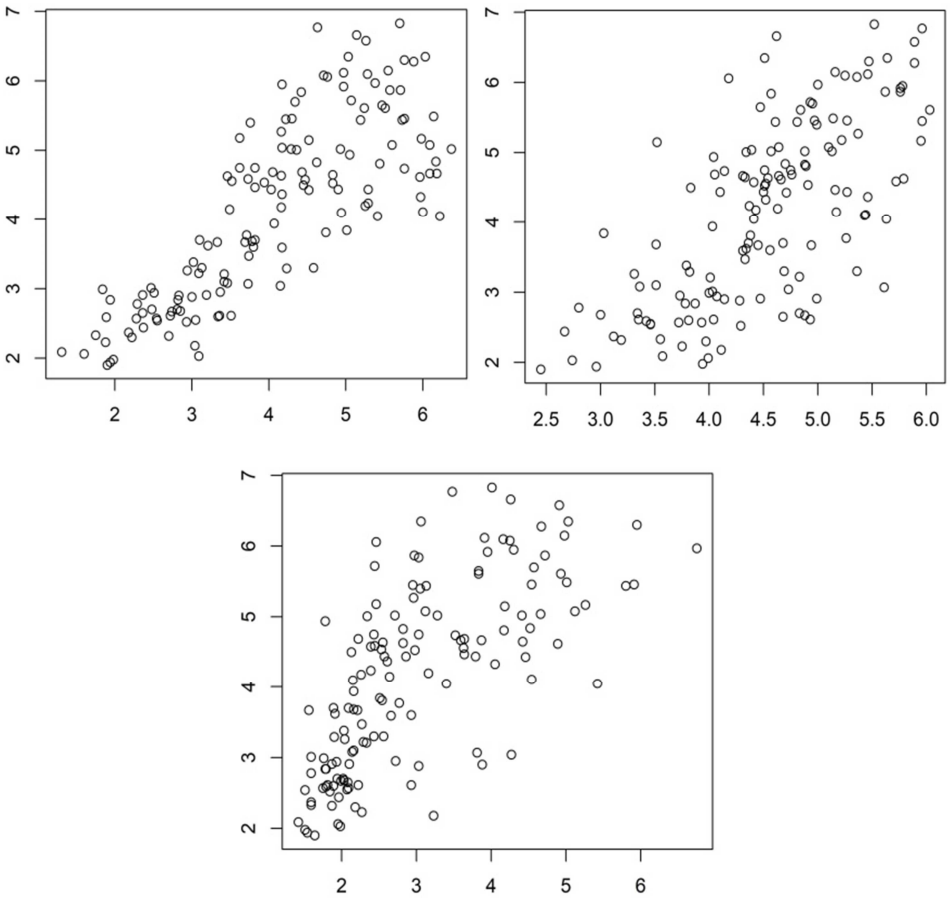
is considered and least squares are used to estimate parameters $\gamma_0,\ldots,\gamma_p$ (instead of the original parameters $\beta_0,\ldots,\beta_p$). This two-stage approach is motivated by the idea (although not directly confirmable) that $var\, e_i = \sigma^2 u_i^2$, which implies the errors $v_1,\ldots,v_n$ in (4) to be homoscedastic with variance σ^2. The results of all regressors in (4) except for variable 12 turn out to be significant and the Breusch-Pagan test

in the transformed model (4) yields a *p*-value 0.063. This is, however, very misleading due to the presence of redundant variables as we will now see.

If we use (4) only with the set of variables {1, 5, 6, 9, 10}, which are the 5 variables significant by *t*-tests of above, all of them turn out to be significant in (4), and the Breusch-Pagan test yields a *p*-value 0.009. Additional graphs (not presented here) do not reveal a linear trend of the transformed response against most of the transformed regressors. In addition, the transform introduces severe outliers (or propagates their effect) to the data. The results remain rather similar if the LWS model is used, which is more appropriate for data contaminated by outliers; there remains even heavier heteroscedasticity in the transformed model compared to the original model (1).

F i g u r e 1

Plot of the Response Against Variable 5 (left), 6 (right) and 10 (bottom)



Source: Own computation.

Results of three versions of tests for regression quantiles (see Section 1) are presented in Table 2. Tests (A) yield 5 significant regressors for the regression median. For τ equal to 0.1 or 0.9, they yield not more than two significant regressors. Such dependence of the test result is due to narrower confidence intervals for more extreme τ (see Koenker, 2005, p. 72). The effect of regressors on the response is more complex, and regressor 5 (and 6) turns out to be important for smaller τ, while regressor 6 (together with 5 and 10) becomes crucial for larger τ. We consider the set {5, 6, 10} to be result of variable selection by regression quantiles, while each of the three variables comes from a different subindex of TTCI: Enabling Environment (variable 5), Travel and Tourism Policy (6), and Infrastructure (10), respectively. Also variables 1 and 4, significant for the regression median, are related to macroeconomic environment and highly correlated with variable 5.

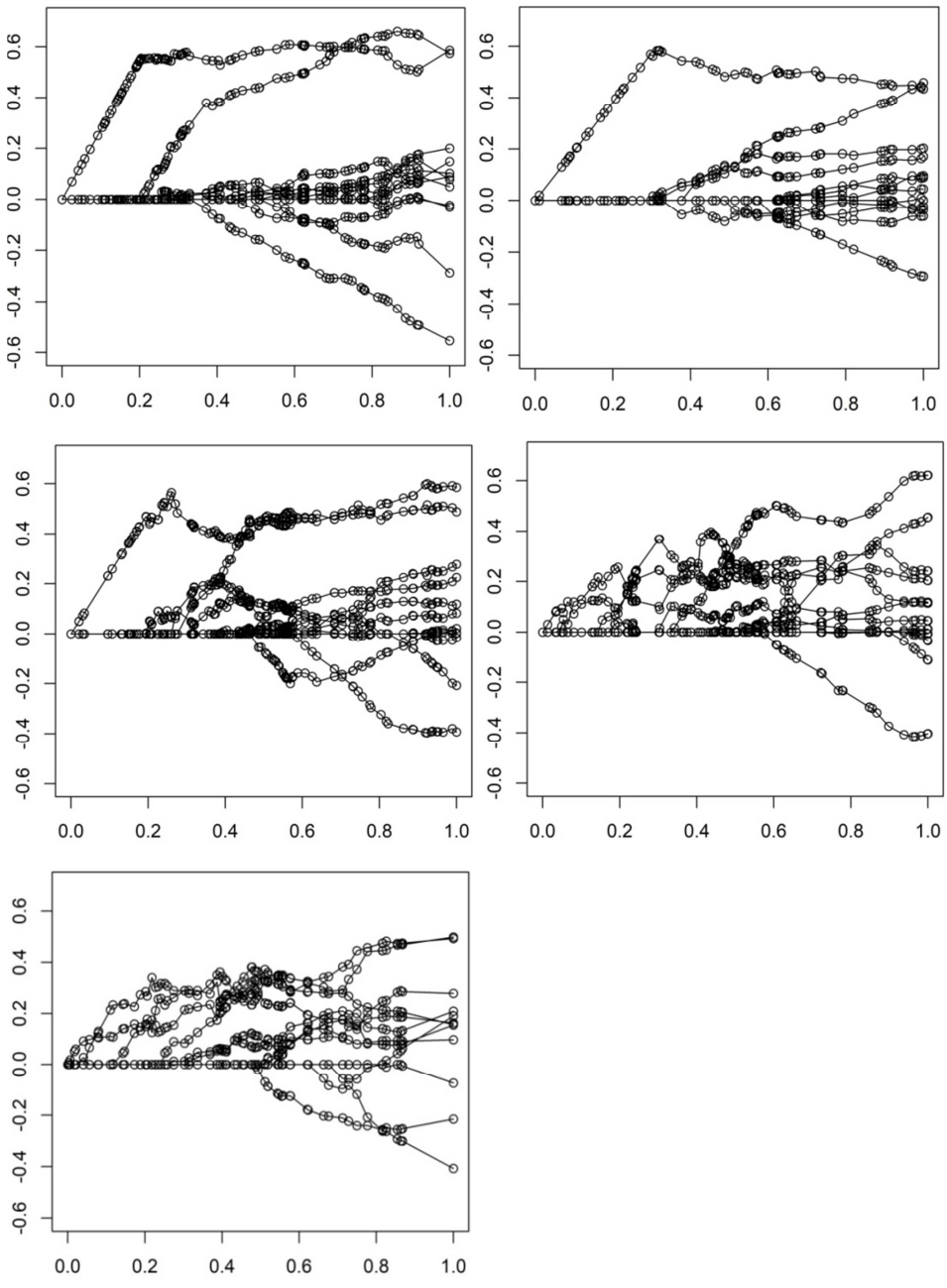
The test (B) of equality of all the 12 slopes of β for 5 different values of τ was performed across values equal to {0.1, 0.3, 0.5, 0.7, 0.9}. The result is highly significant with p-value 10^{-9}. The test remains significant also if various subsets of regressors are considered. This is true for the subset {1, 4, 5, 6, 10} with $p = 8 \cdot 10^{-6}$ or {1, 3, 5, 6, 9, 10} with $p = 2 \cdot 10^{-4}$

The Khmaladze test (C) yields a highly significant result if used for all 12 regressors as well as for these subsets. Its p-value can also be approximated, but this is not provided in R software. While the Kmaladze test allows considering the same null hypothesis for individual regressors, we observe such test to be extremely vulnerable to multicollinearity, as it is heavily dependent on omitting a single regressor from the model.

For the RQ-lasso for different values of τ, Figure 2 shows graphs of estimates for the 12 regressors, which are analogous to graphs commonly used for LS-lasso. The figures show solution paths of each b_j (estimate of β_j) against values of $s = |b_j| / max|b_j|$, depending on the regularization parameter λ. Here, the maximal values in the denominator of s are equal to values of (non-regularized) regression τ-quantiles. Figure 2 thus reveals the effect of the regularization on the estimates with $s = 0$ corresponding to maximum possible regularization and $s = 1$ to none. Cross-validation minimizing the mean square error makes no sense here for finding a suitable estimate of the regularization level, but no alternative method seems to be available. The set of relevant variables in the model seems not to be the same across τ. Still, based on a subjective evaluation of Figure 2, we consider the regressors 3, 5, 6 and 10 as the variables relevant for predicting the response. Here, variables 3 and 5 are highly correlated with correlation coefficient $r = 0.82$; both are connected to macroeconomic environment, while variable 3 is more important for larger τ and variable 5 for smaller τ.

F i g u r e 2

RQ-lasso Estimates Depending on *s*



Note: Horizontal axis: values of *s* between 0 and 1. Vertical axis: RQ-lasso estimates of regression parameters for all 12 regressors for $\tau = 0.1$ and 0.3 (top row), $\tau = 0.5$ and 0.7 (middle row), and $\tau = 0.9$ (bottom).

Source: Own computation.

For the relevant set {3, 5, 6, 10} of regressors, Table 4 presents an overview of various estimates of the regression parameters. RQ-lasso indicates the heteroscedasticity in most of the regressors clearly, as the estimates of the parameters are very different across τ. For the subset {3, 5, 6, 10}, Table 5 shows confidence intervals for the important regressors 5 and 6. The width of confidence intervals is larger for extreme quantiles, although not depending on τ in a monotone way; this is in accordance with theoretical knowledge already discussed above. We consider RQ-lasso to be a very suitable tool for the analysis of the given data. Lasso estimators are known as useful, when there is a small number of dominant variables in the whole set of variables, which is clearly the case of our dataset. For smaller τ, RQ-lasso reveals a single variable (regressor 5) to be very dominant. For larger τ, 2 or 3 variables turn out to be dominant, while all other regressors do contribute to the response as well.

Table 4

Estimates of Regression Parameters for the Subset of Regressors {3, 5, 6, 10} for the Least Squares and Regression Quantiles (RQ)

	LS	RQ ($\tau = 0.1$)	RQ ($\tau = 0.3$)	RQ ($\tau = 0.5$)	RQ ($\tau = 0.7$)	RQ ($\tau = 0.9$)
β_0	−0.18	0.99	−0.80	0.32	−0.04	−0.58
β_3	0.20	−0.01	0.08	0.09	0.26	0.25
β_5	0.38	0.54	0.49	0.52	0.25	0.44
β_6	0.52	0.26	0.52	0.49	0.53	0.40
β_{10}	0.15	−0.04	0.07	0.09	0.22	0.29

Source: Own computation.

Table 5

A Detailed Overview of Confidence Intervals for β_5 and β_6 for Regression Quantiles in the Model with Regressors {3, 5, 6, 10}

τ	Conf. interval for β_5	Width	Conf. interval for β_6	Width
0.1	(0.47, 0.85)	0.38	(0.09, 0.48)	0.39
0.3	(0.10, 0.68)	0.58	(0.20, 0.71)	0.51
0.5	(0.30, 0.57)	0.27	(0.39, 0.59)	0.20
0.7	(0.17, 0.58)	0.41	(0.32, 0.72)	0.40
0.9	(−0.10, 0.66)	0.76	(0.14, 0.86)	0.72

Source: Own computation.

Table 6

Regressors Arranged According to their Relevance for Explaining the Response, Based on the LS-lasso and RQ-lasso (RQL)

Method	Order of regressors from the most important to the most redundant											
LS-lasso	5	3	10	6	7	9	8	12	1	11	2	4
RQL ($\tau = 0.1$)	5	6	7	8	9	2	1	3	4	12	10	11
RQL ($\tau = 0.3$)	5	6	10	9	4	3	12	1	2	8	7	11
RQL ($\tau = 0.5$)	5	3	10	6	11	7	8	9	2	1	4	12
RQL ($\tau = 0.7$)	10	5	3	11	6	7	12	9	1	2	4	8
RQL ($\tau = 0.9$)	10	5	3	6	11	12	2	7	9	1	4	8

Source: Own computation.

Results of LS-lasso and RQ-lasso (for different τ) are compared in Table 6, where all 12 regressors are ordered according to their contribution to the variability of the response. With an increasing regularization parameter λ, these estimates of each β_j ($j = 1, \ldots, p$) vary until becoming exactly 0. In Table 6, regressors are arranged according to such values of λ, for which their estimates of β_j become 0. While the variable selection of RQ-lasso has theoretically appealing properties as explained in Section 1, it finds rather different variables from those obtained with a backward variable selection with (standard) regression quantiles, where the latter approach cannot be supported by theoretical arguments.

Conclusions

The objective of the paper is to discuss various approaches to regression modelling under heteroscedasticity and multicollinearity. This discussion provides a motivation for quantile regression and especially regularized regression quantiles. Limitations of standard tools for heteroscedasticity testing or variable selection are discussed together with benefits of more modern tools including regression quantiles and RQ-lasso. The real dataset from the tourism management area, which is downloadable from the website shown in the references under Crotti and Misrahi (2015), is analyzed here as an illustration of particular methods. In the task to explain and predict the Tourist Service Infrastructure (TSI) as a response of 12 pillars (as regressors) of TTCI in 141 countries of the world, the relevant variables turn out to be Health and Hygiene, Information and Communication Technologies Readiness, Prioritization of Travel and Tourism, and Air Transport Infrastructure.

Heteroscedasticity tests do not give satisfactory results in the given dataset in spite of a relatively large *n*. Standard tests do not find the (otherwise apparent) heteroscedasticity in the model. If a subset of regressors is considered, heteroscedasticity becomes revealed, but then it remains difficult to remove it. This is true for diagnostic tests for the least squares as well as for the robust LWS estimator, for which a novel proper form of White's estimator is proposed here in (3). The LWS estimator, which is theoretically proven to be suitable under heteroscedasticity (Víšek, 2011), does not bring many benefits here and its stepwise variable selection suffers from multicollinearity, just like for the least squares. The whole set of regression τ-quantiles (across various τ) is illustrated here as a useful tool for investigating the relationship of the response on regressors. In the given data, this complex relationship turns out to be influenced by heteroscedasticity, so that it cannot be simply characterized by a single mean trend.

The travel and tourism dataset shows that lasso estimation brings benefits not only for a large number of variables *p*, for which it was originally designed. The

variable selection of RQ-lasso is known to possess a variable selection consistency, which is not true for any stepwise procedure for (standard) regression quantiles. Thus, we evaluate RQ-lasso as preferable, although we admit that its subjective choice of relevant variables requires an experienced user. RQ-lasso is also more complicated than the LS-lasso and their resulting models are not equivalent; the latter performs shrinking of estimates of β_j ($j = 1,\ldots, p$) depending on an additional regularization parameter λ.

On the whole, it turns out that variable selection by RQ-lasso can be performed in an analogous spirit as by LS-lasso, although the former has been rarely used in econometric applications. Nevertheless, both RQ-lasso and LS-lasso implicitly assume $var b_1,\ldots, var b_p$ to be at least approximately equal, which has not been theoretically investigated for RQ-lasso. While LS-lasso exhibits some drawbacks such as instability, too restrictive conditions on the design matrix, or confounding variables (as discussed by Kalina, 2014), RQ-lasso may potentially inherit these properties from LS-lasso as well.

Beyond the scope of this paper, other possibilities for regression under heteroscedasticity include transforms of the response, possibly combined with a suitable dimension reduction method, such as sliced inverse regression of Li (1991). A ridge LWS estimator, i.e., a tailor-made version of the LWS for multicollinear data, was investigated by Jurczyk (2012). As future methodological research, we plan to extend and investigate lasso estimation to the LWS estimator or to multivariate quantiles of Hlubinka and Šiman (2013). As tourism represents a significant part of the total economy in many countries and provides numerous employment opportunities, exploration of development patterns in tourism should be a focal point for policymakers around the world. Using new statistical methods can help the economic development of any industry. For tourism policymakers, an in-depth analysis of conditional quantiles of the response can make the right strategic decisions possible. This may contribute to strengthening the competitiveness of tourism industry and to ensuring sustainable tourism development.

References

ALHEETY, M. I. − KIBRIA, B. G. (2009): On the Liu and almost Unbiased Liu Estimators in the Presence of Multicollinearity with Heteroscedastic or Correlated Errors. Suveys in Mathematics and its Applications, *4*, No. 1, pp. 155 – 167.

BUCHER, S. (2015): Konkurencieschopnosť európskych destinácií cestovného ruchu: Hodnotenie komplexného indexu konkurencieschopnosti. Ekonomický časopis/Journal of Economics, *63*, No. 7, pp. 634 – 655.

CROTTI, R. − MISRAHI, T. (2015): The Travel & Tourism Competitiveness Report 2015. Growth Through Shocks. Geneva: World Economic Forum. Available at: <http://www3.weforum.org/ docs/TT15/WEF_Global_Travel&Tourism_Report_2015.pdf>.

DLOUHÝ, M. – FLUSSEROVÁ, L. (2007): Three Approaches to the Analysis of Cost Function in Health Care. Ekonomický časopis/Journal of Economics, *55*, No. 1, pp. 69 – 78.

FISHER, R. A. (1920): A Mathematical Examination of the Methods of Determining the Accuracy of Observation by the Mean Error, and by the Mean Square Error. Monthly Notices of the Roayl Astronomical Society, *80*, No. 8, pp. 758 – 770.

FISHER, R. A. (1922): On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London, *222*, Series A, pp. 309 – 368.

FITZENBERGER, B. − KOENKER, R. − MACHADO, J. A. F. (2002): Economic Applications of Quantile Regression. Heidelberg: Physica-Verlag. ISBN 3-7908-1448-2.

GREENE, W. H. (2011): Econometric Analysis. 7th ed. Old Tappan: Prentice Hall. ISBN 978-0-13-139538-1.

GUTENBRUNNER, C. – JUREČKOVÁ, J. − KOENKER, R. − PORTNOY, S. (1993): Tests of Linear Hypotheses Based on Regression Rank Scores. Journal of Nonparametric Statistics, *2*, No. 4, pp. 307 – 331.

HASTIE, T. − TIBSHIRANI, T. − FRIEDMAN, J. (2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer. ISBN 978-0-387-84857-0.

HLUBINKA, M. − ŠIMAN, M. (2013): On Elliptical Quantiles in the Quantile Regression Setup. Journal of Multivariate Analysis, *116*, No. 3, pp. 163 – 171.

JENČOVÁ, S. (2018): Aplikácie pokročilých metód vo finančno-ekonomickej analýze elektrotechnického odvetvia Slovenskej republiky. Ostrava: VŠB-TU Ostrava. ISBN 978-80-248-4219-6.

JIA, J. − ROHE, K. − YU, B. (2013): The Lasso under Poisson-like Heteroscedasticity. Statistica Sinica, *23*, No. 1, pp. 99 **–** 118.

JIANG, R. − QIAN, W. M. (2016): Quantile Regression for Single-index-coefficient Regression Models. Statistics and Probability Letters, *110*, No. 1, pp. 305 – 317.

JURCZYK, T. (2012): Outlier Detection under Multicollinearity. Journal of Statistical Computation and Simulation, *82*, No. 2, pp. 261 – 278.

JUREČKOVÁ, J. − NAVRÁTIL, R. (2014): Rank Tests in Heteroscedastic Linear Model with Nuisance Parameters. Metrika, *77*, No. 3, pp. 433 – 450.

KALINA, J. (2012): On Multivariate Methods in Robust Econometrics. Prague Economic Papers, *21*, No. 1, pp. 69 – 82.

KALINA, J. (2014): Classification Methods for High-dimensional Genetic Data. Biocybernetics and Biomedical Engineering, *34*, No. 1, pp. 10 – 18.

KALINA, J. – PEŠTOVÁ, B. (2017): Exact Inference in Robust Econometrics under Heteroscedasticity. In: 11th International Days of Statistics and Economics MSED 2017. [Proceedings.] Slaný: Melandrium, pp. 636 − 645.

KOENKER, R. (2005): Quantile Regression. Cambridge: Cambridge University Press. ISBN 978-0-521-84573-1.

KOENKER, R. − CHERNOZHUKOV, V. − HE, X. − PENG, L. (2017): Handbook of Quantile Regression. Boca Raton: Chapman and Hall/CRC. ISBN 978-1-498-72528-6.

KOENKER, R. − XIAO, Z. (2002): Inference on the Quantile Regression Process. Econometrica, *70*, No. 4, pp. 1583 – 1612.

LEW, A. A. − NG, P. T. (2011): Using Quantiles Regression to Understand Visitor Spending. Journal of Travel Research, *51*, No. 3, pp. 278 – 288.

LI, K. C. (1991): Sliced Inverse Regression for Dimension Reduction. Journal of the American Statistical Association, *86*, No. 414, pp. 316 – 327.

MATLOFF, N. (2017): Statistical Regression and Classification. From Linear Models to Machine Learning. Boca Raton: CRC Press. ISBN 978-1-498-71091-6.

NYBLOM J. (2015): Permutation Tests in Linear Regression. In: NORDHAUSEN, K. and TASKINEN, S. (eds): Modern Nonparametric, Robust and Multivariate Methods. Cham: Springer, pp. 69 – 90.

PULIDO-FERNÁNDEZ, J. I. − RODRÍGUEZ-DÍAZ, B. (2016): Reinterpreting the World Economic Forum's Global Tourism Competitiveness Index. Tourism Management Perspectives, *20*, No. 1, pp. 131 – 140.

RANGANAI, E. (2016): Quality of Fit Measurement in Regression Quantiles: An Elemental Set Method Approach. Statistics and Probability Letters, *111*, No. 1, pp. 18 – 25.

RUDKIN, S. − SHARMA A. (2017): Enhancing Understanding of Tourist Spending Using Unconditional Quantiles Regression. Annals of Tourism Research, *66*, No. 3, pp. 188 – 191.

STEIN, C. (1956): Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, *1*, No. 1, 197 – 206.

STOCK, J. H. − WATSON, M. W. (2015): Introduction to Econometrics. 3rd ed. Harlow: Pearson. ISBN 978-0-13-348687-2.

ŠTEFKO, R. – KIRÁĽOVÁ, A. – MUDRÍK, M. (2015): Strategic Marketing Communication in Pilgrimage Tourism. Procedia – Social and Behavioral Sciences, *175*, No. 1, pp. 423 – 430.

VAŠANIČOVÁ, P. − LITAVCOVÁ, E. − JENČOVÁ, S. − KOŠÍKOVÁ, M. (2017): Dependencies between Travel and Tourism Competitiveness Sub-indexes: The Robust Quantile Regression Approach. In: 11th International Days of Statistics and Economics MSED 2017. [Proceedings.] Slaný: Melandrium, pp. 1729 − 1739.

VÍŠEK, J. Á. (2010): Heteroscedasticity Resistant Robust Covariance Matrix Estimator. Bulletin of the Czech Econometric Society, *17*, No. 27, pp. 33 – 49.

VÍŠEK, J. Á. (2011): Consistency of the Least Weighted Squares under Heteroscedasticity. Kybernetika, *47*, No. 2, pp. 179 – 206.

WANG, H. J. – ZHOU, J. − LI, Y. (2013): Variable Selection for Censored Quantile Regression. Statistica Sinica, *23*, No. 1, pp. 145 – 167.

WHITE, H. (1980): A heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. Econometrica, *48*, No. 4, pp. 817 – 838.

WHITTINGHAM, M. J. − STEPHENS, P. A. − BRADBURY, R. B. − FRECKLETON, R. P. (2006): Why Do We Still Use Stepwise Modeling in Ecology and Behaviour? Journal of Animal Ecology, *75*, No. 5, pp. 1182 – 1189.

WOOLDRIDGE, J. M. (2013): Introductory Econometrics: A Modern Approach. 5th ed. Boston: Cengage Learning. ISBN 1-111-53104-8.

WTTC (2018): Travel & Tourism Global Economic Impact & Issues 2018. Available at: <https://www.wttc.org/-/media/files/reports/economic-impact-research/documents-2018/global-economic -impact-and-issues-2018-eng.pdf>.

YOUNG, D. S. (2017): Handbook of Regression Methods. Boca Raton: Chapman and Hall/CRC. ISBN 978-1-4987-7529-8.