

Serbian Journal of Management 15 (2) (2020) 193 - 203

Serbian Journal of Management

# AN APPLICATION OF DIRECTIONAL QUANTILES TO ECONOMIC DATA WITH A MULTIVARIATE RESPONSE

### Jan Kalina

The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic, & Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic

(Received 29 July 2019; accepted 20 March 2020)

#### Abstract

Quantile regression represents a popular and useful methodology for modeling quantiles of a response variable based on one or more independent variables. Directional quantiles represent an available extension to the linear regression model with a multivariate response. However, we are not aware of any application of directional quantiles to real data in the literature. An illustration of directional quantiles to an economic dataset is presented in this paper, particularly a modeling of a two-dimensional response in the classical Engel's dataset on household consumption from the 19th century. The results reveal the directional quantiles to yield meaningful results. They order individual observations according to their depth, i.e. from the most central to the most outlying. We compare their result with those of a (more standard) outlier detection. On the whole, we perceive directional quantiles as a potentially useful tool for the analysis of data, if accompanied by a thorough analysis by standard tools.

*Keywords*: regression quantiles, multivariate response, household expenditures, heteroscedasticity, outliers

## **1. INTRODUCTION**

In a variety of management applications, linear regression represents a useful methodology of exploratory statistics for modeling and predicting a continuous response based on one or several independent variables (regressors). Thevaraja and Rahman (2019) stressed the importance of regression modeling within management tasks, allowing to understand individual activities within very complex behavioral, socio-economic and ecological systsems. Examples of recent management tasks solved by means of linear regression

\* Corresponding author: kalina@cs.cas.cz

DOI: doi.org/10.5937/sjm15-22671

include the work of Mumtaz et al. (2018), who analyzed industrial data to investigate the effect of the green supply chain management, which is considered as one of the most important environmental management approaches. Rahimi (2017) studied the effect of organizational culture factors on implementing customer relationship management in hotels.

Practical management applications however sometimes require to estimate quantiles of the response in a linear regression, i.e. depending on one or more independent variables (regressors), rather than to perform a plain modeling of the response based on the regressors. Quantile regression represents an established methodology for estimating a set of quantiles of the response variable explained by regressors in the standard linear regression model (Koenker, 2005). The list of (rather diverse) applications of quantile regression in management includes the work of Matallín-Sáez et al. (2019), who investigated the effect of active management and effect of various characteristics and abilities of managers on the performance of mutual funds. De Andrés et al. (2018) estimated effort costs within management of software development projects by means of regression quantiles. Li and Hwang (2019) used regression quantiles to investigate the effect of discretionary accruals (earnings management or earnings manipulation) on changes of stock prices of a company.

The quantile regression has become an established statistical methodology for estimating a set of quantiles of the response variable explained by regressors in the standard linear regression model (Koenker, 2005). Their ability to model the regression trend for asymmetric distribution of errors or to detect heteroscedasticity, especially if the whole set of regression  $\tau$ -quantiles is evaluated for various values  $\tau \in (0,1)$ , makes regression quantiles popular tools for analyzing real economic data (Koenker et al., 2017). They are able to capture the whole distribution of the response instead of simply considering the mean trend and allow a clear interpretation. Various generalizations of regression quantiles beyond the standard linear regression are currently available. The first such adaption was the proposal of locally polynomial quantile regression already in Koenker (2005). Other more recent modifications motivated bv econometric applications include quantile regression for clustered data (Parente & Silva, 2016) or testing for Granger causality between economic time series by means of regression quantiles (Troster, 2018).

In this paper, we are interested in regression quantiles for a multivariate response (Koenker et al., 2017). Directional  $\tau$ -quantiles based on the concept of data depth (Chernozhukov et al., 2017) represent an important direction of research aimed at modeling a multivariate response. Very central data have a large value of the depth, while the value of the depth decreases together with an increasing (but only vaguely defined) outlyingness. We are not aware of any application of directional quantiles on real economic data. This is true in spite of the facts that directional quantiles were originally intended exactly for economic applications (Paindaveine & Šiman, 2011; Boček & Šiman, 2017a) and that their implementation in R software is publicly available (Šiman & Boček, 2016).

The aim of this paper is to illustrate directional quantiles on a real economic dataset with a two-dimensional respose and to show its benefits. We claim that a proper analysis of real data by directional quantiles requires to consider a whole series of values of  $\tau$  as well as to perform a thorough analysis of the data by standard tools, i.e. which consider each of the components of the multivariate response separately. Directional quantiles are recalled in Section 2. The publicly available Engel's dataset on household expenditures is presented in Section 3, while the results are concluded in the final Section 4.

# 2. DIRECTIONAL QUANTILES

called Directional quantiles (also quantiles) projection proposed in Paindaveine and Šiman (2011) represent an important class of regression quantiles for models with a multivariate response. Directional quantiles were proposed with the intention to be applied for modeling heteroscedastic data, particularly for detecting heteroscedasticity (Boček & Šiman, 2017a), e.g. for portfolio optimization. However, there are only population versions of directional quantiles illustrated in the paper by Paindaveine and Šiman (2011), without providing any example on real data.

The idea of directional quantiles is based on the concept of data depth, particularly on the most traditional and probably also the most popular data depth, namely the halfspace depth (also known as Tukey depth), see Liu et al. (1999). While their model rigorously described in Paindaveine and Šiman (2011) is rather complicated, the user only selects a parameter  $\tau$  known as the level.

The directional quantile for a given  $\tau$  has the form of (multivariate) quantile regions with some boundary, obtained as intersections of (several) corresponding halfspaces (Boček & Šiman, 2017a). The directional quantiles fulfil a set of necessary and sufficient Karush-Kuhn-Tucker conditions formulated by Paindaveine and Šiman (2011). The directional quantile itself (for a given  $\tau$ ) is obtained as intersection of certain halfspaces. We can say that the directional quantile (again for a given  $\tau$ ) divides the data to three sets:

• The boundary (i.e. contour),

• The space within the boundary, i.e. central observations, or observations with a higher depth,

• The space outside the boundary, i.e. observations with a lower depth.

While the computation of directional quantiles requires to solve a complicated optimization efficient task. their implementation is available in the package modQR (Šiman & Boček, 2016) of R software. The implemented deterministic algorithm documented in Boček and Šiman (2017a) exploits parametric programming and is granted to yield a unique solution. The package modQR offers two different methods (denoted as Method 1 and 2) for computing directional quantiles, corresponding to two different formulations of constraints; thus, the constraints correspond to two different versions of the Lagrange multipliers (Boček & Šiman, 2017a).

# 3. AN ILLUSTRATIVE ANALYSIS OF THE ENGEL'S DATASET

The aim of the example is to illustrate directional quantiles on real economic data, namely their ability to model a multivariate response depending on multiple regressors. We work with a famous data set of Ernst Engel (1821–1896) on household expenditures (Engel, 1857). Engel was a German (Saxon) economist, statistician and also propagator of statistics. The interest of Engel was to investigate the economic disparity and particularly the dependence of household expenditures on the household income, i.e. construction of so-called Engel curves.The dataset contains n=235 measurements, obtained as a mixture of 199 Belgian measurements with 36 obtained from other European countries. It was exactly on these data that Engel studied and formulated the famous Engel's law, stating that families with a lower income tend to spend a greater part of their budget on food. In fact, he was able to understand heteroscedasticity in linear regression as probably the first researcher aware of the phenomenon of heteroscedasticity in linear regression (Haughton & Haughton, 2011).

The dataset of Engel was used by various researchers, while most of them considered only various analyses with a onedimensional response. One-dimensional analysis of Engel's dataset (with a single response) was presented e.g. in the fundamental book (Koenker, 2005). Carlier et al. (2016) considered a two-dimensional response depending on a single regressor and used an original vector quantile regressor and based on Monge-Kantorovich depth (Chernozhukov et al., 2017). This dataset (i.e. the particular response and regressor) will be considered in this section as well, while we obtained the dataset from the website of Galichon (2017).

We consider here three continuous variables, while the multivariate response  $(Y_1, Y_2)^T$  is explained by the regressor X (and of course on the intercept as well):

•  $X = (X_1, ..., X_n)^T =$  total expenditures,

•  $Y_1 = (Y_{11}, ..., Y_{1n})^T =$ food expenditures,

•  $Y_2 = (Y_{21}, ..., Y_{2n})^T$  = expenditures on housing, fuel and light.

We perform all computations in R software (R core team, 2019). The relationships of the two components of the response on X are shown in Figure 1.

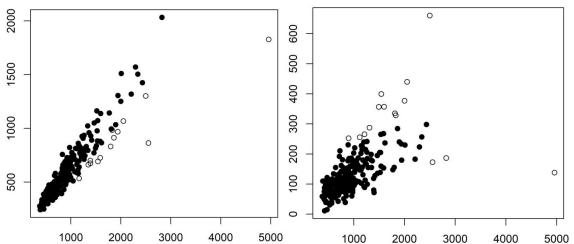


Figure 1. Left:  $Y_1$  (food expenditures, vertical axis) against X (total expenditures, horizontal axis). Right:  $Y_2$  (expenditures on housing, fuel and light) against X (total expenditures). In both graphs, outliers detected by the LTS are shown as empty circles, the remaining measurements are full circles.

# **3.1.** Statistical analysis with a onedimensional response

First, let us present results of (more or less) standard analysis using always a onedimensional response. Particularly, we investigate the two models

$$Y_{1i} = \beta_{10} + \beta_{11}X_i + e_{1i}$$
  
and  
$$Y_{2i} = \beta_{20} + \beta_{21}X_i + e_{2i}$$
 (1)

for each i=1,...,n. These will be symbolically denoted as  $Y_1 \wedge X$  and  $Y_2 \wedge X$ , respectively. We use the following estimators (including robust ones) of parameters of linear regression:

• Least squares (LS);

• The least trimmed squares estimator (see e.g. Jurečková et al., 2019) denoted as  $LTS(\alpha)$  defined by

$$\arg\min_{\beta\in\mathbb{R}^2}\frac{1}{h}\sum_{i=1}^h u_{(i)}^2(b)$$
(2)

with a given  $\alpha \in [1/(2,1))$ , where  $u_{(i)}(b)$  are residuals corresponding to a fixed  $b \in \mathbb{R}^2$ ,  $h=\lfloor \alpha n \rfloor$ , and  $\lfloor x \rfloor$  denotes the integer part of  $x \in \mathbb{R}$ ;

• The least weighted squares (LWS) estimator (Víšek, 2011; Kalina, 2013) with linear weights or weighted generated by a (decreasing) logistic function (Kalina & Tichavský, 2020);

• MM-estimator with breakdown point 0.5 and efficiency 0.95 (see e.g. Jurečková et al., 2019);

• Regression  $\tau$ -quantiles (for different fixed values of  $\tau$ ).

The results are presented in Table 1. In the model  $Y_1 \lor X$ , the regression quantiles indicate a clear heteroscedasticity. This is confirmed also by the Wald's test based on the White's estimator of the covariance matrix, implemented in the package sandwich of R software; there, however, the test is based on comparing the test statistic with confidence interval and no p-value is reported. In our case, the confidence limits [0.35; 0.62] does not cover 0 and thus the test is significant. In addition, the Breusch-Pagan test gives a p-value below 10<sup>-6</sup> here and confirms heteroscedasticity as well. The LTS estimator with  $\alpha = 3/4$  detects 14 outliers here. If these are ignored, the data remain heteroscedastic; the Wald's test yields the confidence interval [0.60; 0.68] and the Breusch-Pagan test the p-value below 10<sup>-6</sup> again. Thus, the significant heteroscedasticity in the model  $Y_1 \lor X$  is not caused merely by outliers.

In the model  $Y_2 \lor X$ , there is again a clear heteroscedasticity revealed by regression quantiles. The Wald's test based on the White's covariance matrix yields a significant result based on the confidence interval [0.03; 0.16] for the test statistic, and the Breusch-Pagan test as well (with a *p*value below 10<sup>-6</sup>). The LTS estimator detects 15 outliers here. If these are ignored, the data remain to be significantly heteroscedastic; the Wald's test yields the confidence interval [0.08; 0.68] and the Breusch-Pagan test the *p*-value 0.0011 again.

The multivariate response  $(Y_1, Y_2)^T$  contains clear outliers, as revealed in Figures 2 and 3. The correlation coefficient between  $Y_1$  and  $Y_2$  equals 0.56, while the Spearman's nonparametric analogue is even larger and equal to 0.60. A sophisticated (but subjective) multivariate outlier detection, implemented in the chisq.plot procedure of

the mvoutlier package (Filzmoser & Gschwandtner, 2018) of R software, finds 14 outliers in the data  $(Y_1, Y_2)^T$ . These are shown in Figure 2 (left).

#### 3.2. Results of directional quantiles

Now we apply directional quantiles to the Engel's dataset exploiting their available implementation in the modQR package (Šiman & Boček, 2016) of R software. The results for particular choices of  $\tau$  are shown in Figures 2 (right) and 3. As explained in

Section 2, the output of the computation of directional quantiles in modQR contains for each given  $\tau$  a list of measurement within the quantile, on the boundary, and outside the quantile. At the same time, the output contains a figure with the same information. The number of observations at the boundary or outside the boundary for the directional quantile with a particular  $\tau$  is shown in Table 2. These numbers are unpredictable and depend on the data in a very complicated way. Methods 1 and 2 give different results, while an inspection of the ordering of

*Table 1. Results of one-dimensional analysis of regression models*  $Y_1 \lor X$  *and*  $Y_2 \lor X$  *in the example of Section 3.* 

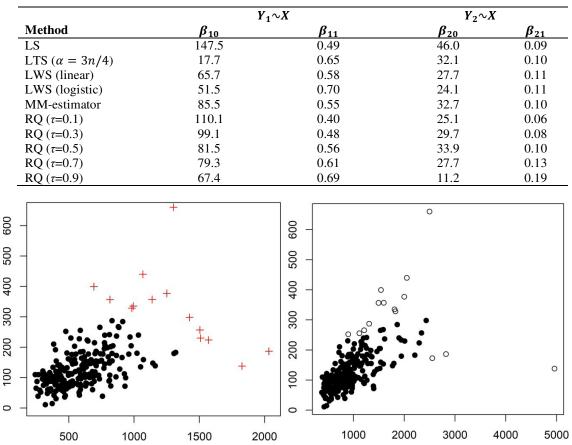


Figure 2. Plots of  $Y_1$  and  $Y_2$ . Left: outliers detected by the multivariate outlier detection of Filzmoser and Gschwandtner (2018) are denoted by a plus sign, the remaining measurements are full circles. Right: measurements outside the contour of the directional quantile for  $\tau$ =0.005 are empty circles, the remaining measurements are full circles.

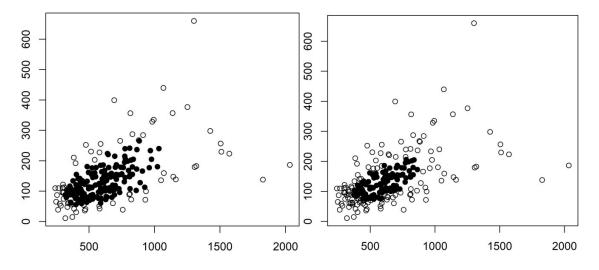


Figure 3. Plots of  $Y_1$  and  $Y_2$ . Measurements outside the contour of the directional quantile for  $\tau$ =0.050 (left) and  $\tau$ =0.120 (right) are empty circles, the remaining measurements are full circles.

Table 2. The number of observations within the boundary (at the boundary, outside the boundary) of directional quantiles in the example of Section 3.

	Method 1 Number of observations			Method 2 Number of observations		
	within	at	outside	within	at	outside
τ	the boundary			the boundary		
0.005	222	3	10	229	1	5
0.050	169	6	60	208	2	25
0.120	113	5	117	195	0	40
0.250	39	1	195	145	1	89

observations according to the halspace depth (i.e. centrality of the observations) seems very similar.

It is possible to perceive the boundaries of the quantiles (i.e. contours) and their development with an increasing (or decreasing) value of  $\tau$ . Based on Figures 2 (right) and 3, there is a clear presence of outliers and heteroscedasticity in  $(Y_1, Y_2)^T$ . While outliers and heteroscedasticity are well visible from the graph of  $Y_1$  and  $Y_2$ without any sophisticated tools, Figures 2 (right) and 3 confirm their presence in the regression model. It is not possible to compare the results with those of vector quantile regression presented by Carlier et al., (2016), who analyzed the data without the ambition to investigate or interpret the presence of outliers or heteroscedasticity. Directional quantiles order the observations, from the most central to the most outlying; we are not aware of any other ways for doing this. While heteroscedasticity may be influenced by outliers, which is the case of our dataset, directional quantiles do not clearly decide which observations are outliers. In other words, directional quantiles do not perform outlier detection as their byproduct. Heteroscedasticity is identified in the model with a two-dimensional response (while the two components of the response are significantly correlated), which would not be possible by standard econometric tools. Some households are outlying and

these are those with a large income and thus with potentially large total expenditures.

In Figures 2 (right) and 3, measurements with a large depth lie inside the contour of the quantile, and measurements with a small depth tend to lie outside. Based on these figures, we can say that the directional quantiles indeed fulfil what we intuitively expect from multivariate quantiles. A comparison of these results with those of multivariate outlier detection, i.e. of Figure 2 (left), reveals that the task of quantile estimation is conceptually very different from outlier detection.

#### 4. CONCLUSIONS

Directional quantiles represent an important type of regression quantiles for a multivariate response. This paper presents the very first application of (parametric) directional quantiles (Paindaveine & Šiman, 2011) on a real dataset. The analysis of the publicly available Engel's dataset reveals their meaningfulness and at the same time, the implementation in the modQR package of R software (which has to solve a complicated nonlinear optimization) is reliable and computationally feasibile for the number of observations equal to n=235. Independently on this good experience, let us recall the warnings of Boček and Šiman (2017a) that the computation may not be reliable under the presence of discrete regressors or too many regressors.

It is clear from the model example presented in Section 3 that directional quantiles can be properly interpreted only if their series for different values of  $\tau$  is used. Moreover, directional quantile cannot replace a thorough exploratory data analysis by means of tools for a univariate response. On the other hand, we believe that using directional quantiles should be accompanied by a series of one-dimensional (exploratory) analyses in order to perform a valuable data analysis of data with a multivariate response. While robustness of regression modeling with respect to outliers and/or measurement errors is commonly desirable (Kalina & Schlenker, 2015), the robustness of available quantiles for a multivariate continuous response remains an open question.

There are other possible approaches to quantile estimation for data with a multivariate response, which deserve to be investigated within future research applications, mainly a nonparametric version of directional quantiles (Boček & Šiman, 2017b) or elliptical quantiles, e.g. those for nonlinear regression (Hlubinka & Šiman, 2015) or elliptical multiple-output regression (Hallin & Šiman, 2016). In addition, it is planned to investigate quantile estimation for neural networks with a multivariate output layer (e.g. with several neurons in the output layer; cf. Kalina, 2014); neural networks namely represent a nonlinear regression methodology with a recently increasing number of applications within various management tasks. To give only a few examples, neural networks were used to improve the safety management within a chemical enterprise (Yuan et al., 2019), or to develop a cost-effective strategy for waste management (Azadi & Karimi-Jashni, 2016).

Acknowledgements: The research was supported by the grant 18-01137S of the Czech Science Foundation. We are thankful to A. Galichon for providing the Engel's dataset (Galichon, 2017) and to Miroslav Šiman and Pavel Boček for discussion.

# ПРИМЕНА УСМЕРЕНИХ КВАНТИЛА НА ЕКОНОМСКЕ ПОДАТКЕ СА МУЛТИВАРИЈАТНИМ ОДГОВОРОМ

#### Jan Kalina

#### Извод

Квантила регресија представља популарну и корисну методологију за моделирање квантила променљиве одговора на основу једне или више независних променљивих. Усмерени квантили представљају доступно проширење модела линеарне регресије са мултиваријантним одзивом. Међутим, није позната ниједна примена усмерених квантила на стварне податке у литератури. У овом раду је представљена илустрација усмерених квантила према економском скупу података, посебно моделирање дводимензионалног одговора у класичном Енгеловом скупу података о потрошњи домаћинстава из 19. века. Резултати откривају да усмерени квантили дају значајне резултате. Они одређују појединачна посматрања у складу са њиховом дубином, тј. од централног до најудаљенијег. У раду се упоређује њихов резултат са резултатима (стандарднијег) откривања одступања. У целини, усмерени квантили се могу посматрати као потенцијално корисно средство за анализу података, ако је праћено темељном анализом стандардним алатима.

*Кључне речи:* регресијски квантили, мултиваријантни одговор, издаци домаћинства, хетероскедастичност, одступања

#### References

Azadi, S., & Karimi-Jashni, A. (2016). Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: A case study of Fars province, Iran. Waste Management 48, 14-23.

Boček, P., & Šiman, M. (2017a). Directional quantile regression in R. Kybernetika, 53 (3), 480-492.

Boček, P., & Šiman, M. (2017b). On weighted and locally polynomial directional quantile regression. Computational Statistics, 32 (3), 929-946.

Carlier, G., Chernozhukov, V., & Galichon, A. (2016). Vector quantile regression: An optimal transport approach. Annals of Statistics, 44 (3), 1165-1192.

De Andrés J., Landajo M., & Lorca P.

(2018). Using Nonlinear Quantile Regression for the Estimation of Software Cost. In: de Cos Juez F. et al. (eds) Hybrid Artificial Intelligent Systems. HAIS 2018. Lecture Notes in Computer Science, 10870. Springer, Cham. 422-432.

Engel, E. (1857). The relations of production and consumption in the Kingdom of Saxony (in German). Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Inneren, 8, 1-54.

Filzmoser, P., & Gschwandtner, M. (2018). mvoutlier: Multivariate outlier detection based on robust methods. Retreived from https://CRAN.R-project.org/package=mvoutlier.

Galichon, A. (2017). Data file. Retrieved from http://alfredgalichon.com/wp-content/uploads/2017/03/.

Hallin, M., & Šiman, M. (2016). Elliptical

multiple-output quantile regression and convex optimization. Statistics and Probability Letters, 109, 232-237.

Haughton, D., & Haughton, J. (2011). Living standards analytics. Development through the lens of household survey data. Springer. New York, NY, USA

Hlubinka, D., & Šiman, M. (2015). On generalized elliptical quantiles in the nonlinear quantile regression setup. Test, 24, 249-264.

Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. Annals of Statistics, 45 (1), 223-256.

Jurečková, J., Picek, J., & Schindler, M. (2019). Robust statistical methods with R. 2nd edn. CRC Press. Boca Raton, FL, USA

Kalina, J. (2013). Highly robust methods in data mining. Serbian Journal of Management 8 (1), 9-24.

Kalina, J. (2014). On robust information extraction from high-dimensional data. Serbian Journal of Management, 9 (1), 131-144.

Kalina, J., & Schlenker, A. (2015). A robust supervised variable selection for noisy high-dimensional data. BioMed Research International, 2015, 320385.

Kalina, J., & Tichavský, J. (2020). On robust estimation of error variance in (highly) robust regression. Measurement Science Review, 20 (1), 6-14.

Koenker, R. (2005). Quantile regression. Cambridge, UK: Cambridge University Press.

Koenker, R., Chernozhukov, V., He, X., & Peng, L. (2017). Handbook of quantile regression. Chapman & Hall/CRC. Boca Raton, FL, USA.

Li, L., & Hwang, N.C.R. (2019). Do market participants value earnings management? An analysis using the quantile regression method. Managerial Finance, 45 (1), 103-123.

Liu, R., Parelius, J.M. & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. Annals of Statistics, 27 (3), 783-858.

Matallín-Sáez, J.C., Soler-Domínguez, A., & Tortosa-Ausina, E. (2019). Does active management add value? New evidence from a quantile regression approach. Journal of the Operational Research Society, 70 (10), 1734-1751.

Mumtaz, U., Ali, Y., & Petrillo, A. (2018). A linear regression approach to evaluate the green supply chain management impact on industrial organizational performance. Science of the Total Environment, 624, 162-169.

Paindaveine, D., & Šiman, M. (2011). On directional multiple-output quantile regression. Journal of Multivariate Analysis, 102 (2), 193-212.

Parente, P., & Silva, J.S. (2016). Quantile regression with clustered data. Journal of Econometric Methods, 5 (1), 1-15.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Version 3.5.1 Retrieved from https://www.R-project.org/.

Rahimi, R. (2017). Organizational culture and customer relationship management: A simple linear regression analysis. Journal of Hospitality Marketing & Management, 26 (4), 443-449.

Šiman, M., & Boček, P. (2016). modQR: Multiple-output directional quantile regression. R package version 0.1.1 [ h t t p s : / / C R A N . R project.org/package=modQR].

Thevaraja, M., & Rahman, A. (2020). Assessing robustness of regularized regression models with applications. In J. Xu, S.E. Ahmed, F.L. Cooke, G. Duca (Eds), Proceedings of the Thirteenth International Conference on Management Science and Engineering Management (ICMSEM 2019), Springer, Cham, 401-415.

Troster, V. (2018). Testing for Grangercausality in quantiles. Econometric Reviews, 37 (8), 850-866.

Víšek, J.Á. (2011). Consistency of the least weighted squares under heteroscedasticity. Kybernetika, 47 (2), 179-206.

Yuan, Y., Zhou, X., Man, J., Jiao, H., Jiang, Q., Xu, Q., Kong, S., & Gao, W. (2019). The safety evaluation of management in chemical enterprise with generalized regression neural network. IOP Conference Series Earth and Environmental Science, 295, 042010.