APPLICATION OF IMPLICITLY WEIGHTED REGRESSION QUANTILES: ANALYSIS OF THE 2018 CZECH PRESIDENTIAL ELECTION

Jan Kalina – Petra Vidnerová

Abstract

Regression quantiles can be characterized as popular tools for a complex modeling of a continuous response variable conditioning on one or more given independent variables. Because they are however vulnerable to leverage points in the regression model, an alternative approach denoted as implicitly weighted regression quantiles have been proposed. The aim of current work is to apply them to the results of the second round of the 2018 presidential election in the Czech Republic. The election results are modeled as a response of 4 demographic or economic predictors over the 77 Czech counties. The analysis represents the first application of the implicitly weighted regression quantiles to data with more than one regressor. The results reveal the implicitly weighted regression quantiles to be indeed more robust with respect to leverage points compared to standard regression quantiles. If however the model does not contain leverage points, both versions of the regression quantiles yield very similar results. Thus, the election dataset serves here as an illustration of the usefulness of the implicitly weighted regression quantiles.

Key words: linear regression, quantile regression, robustness, outliers, elections results

JEL Code: D72, C21, Y91

Introduction

Regression quantiles (quantile regression) have been commonly used in various economic applications for a complex modeling of a continuous response variable conditioning on one or more given independent variables (regressors, predictors). Implicitly weighted version based on the idea of robust regression was proposed by Kalina & Vidnerová (2019), where it was however studied only over simplistic data with one independent variable.

Let us mention at least some interesting recent demographic applications of regression quantiles. Davino et al. (2018) applied regression quantiles to path modeling, i.e. to describing dependencies among a set of variables, in a study of quality of life in Italian provinces. Wu &

Guo (2017) used Bayesian quantile regression to find a set of factors, which affect the satisfaction level of the Taiwanese population. Similarly, Ngoo et al. (2020) applied quantile regression to find factors which determine the life satisfaction in Asian countries. Maškarinec (2017) exploited spatial statistical methods to analyze Czech parliamentary elections and was able to find (at least to some extent) that right-wing parties had higher support in economically stronger regions. In all these applications, a whole set of regression quantiles allows a more complex understanding of the relationship of the response on the predictors, compared to the information provided by the standard least squares estimator. We are however not aware of applications of regression quantiles to results of elections.

The aim of this work is to apply the recently proposed implicitly weighted regression quantiles of Kalina & Vidnerová (2019) to the data from Czech presidential election of 2018, while each of the 77 Czech counties is considered as one measurement; such setup is unusual, as detailed analyses of elections are typically performed over survey samples of individual voters. Section 1 recalls standard as well as implicitly weighted regression quantiles. The data description follows in Section 2. We start with analyzing the data with beta regression in Section 3 and least squares in linear regression in Section 4, and results of the standard as well as the implicitly weighted regression quantiles are presented in Section 5.

1 Regression quantiles

We consider the standard linear regression model

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

where $Y_1, \dots, Y_n$ are values of a continuous response variable and $e_1, \dots, e_n$ are random errors (disturbances). The usual task in regression modeling is to estimate the regression parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ by means of estimating the conditional expectation of the response, given fixed values of the regressors (predictors). Especially under heteroscedasticity, it may be however more suitable to consider a whole set of several regression quantiles.

Recently, promising new forms of regression quantiles have been proposed and investigated. For example, Bleik (2019) proposed a simultaneous estimation of several Bayesian regression quantiles assuming the random errors to follow a Laplace distribution; the method was applied to simulated data. Hlubinka & Šiman (2020) investigated generalized elliptical regression quantiles in both linear and nonlinear models for a multivariate response;

their method was applied to analyze anthropometric measurements over a set of 260 physically active women.

1.1 Implicitly weighted regression quantiles

Standard regression quantiles are vulnerable to the presence of leverage points in the data and actually are not robust in terms of the breakdown point. With the motivation to improve the robustness, implicitly weighted regression quantiles denoted here as IWRQ were proposed by Kalina & Vidnerová (2019). The IWRQ procedure exploits the concept of weights assigned to individual observations, inspired by the least weighted squares (LWS) regression estimator; see Víšek (2011) or Kalina & Schlenker (2015) for a discussion of the LWS estimator, which is able to combine a high robustness (i.e. a high breakdown point) with a high efficiency.

We will recall the definition of IWRQ by means of a weight function denoted as ψ. Let us use the notation $u_1(b), \dots, u_n(b)$ for residuals corresponding to a fixed $b \in \mathbb{R}^{p+1}$. Their ranks will be denoted by $R_1(b), \dots, R_n(b)$ to stress the dependence on b. Keeping in mind that both these vectors depend on b, the LWS estimator may be expressed as

$$\arg\min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \psi_1\left(\frac{R_i(b)}{n}\right)(u_i(b))^2. \quad (2)$$

Assuming two given weight functions ψ_1 and ψ_2, IWRQ is defined as

$$\arg\min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left[\psi_1\left(\frac{R_i(b)}{n}\right) I[u_i(b) > 0] + \psi_2\left(\frac{R_i(b)}{n}\right) I[u_i(b) < 0]\right](u_i(b))^2. \quad (3)$$

Particularly, if it is chosen $\psi_2 = c\psi_1$for a given $c > 0$, then (3) reduces to

$$\arg\min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \psi_1\left(\frac{R_i(b)}{n}\right)(I[u_i(b) > 0] + cI[u_i(b) < 0])(u_i(b))^2. \quad (4)$$

In the computations, we use the definition (4). The choice $c = 1$ corresponds to the LWS estimator itself, while $c > 1$ estimates a bottom quantile and $c < 1$ an upper quantile. Thus, the constant c represents a parameter analogous to τ (but with a different interpretation) of standard regression quantiles. It is recommendable that the user chooses several different values of c to examine the results for upper as well as lower quantiles. We use here trimmed linearly decreasing weights; taking $\tau = \lfloor 3n/4 \rfloor$, they are generated by the weight function

$$\psi(t) = \left(1 - \frac{t}{\tau}\right) I[t < \tau], \quad t \in [0,1]. \quad (5)$$

The computation of IWRQ may directly exploit the FAST-LTS algorithm of Rousseeuw & van Driessen (2006).

2 Data description

We analyze the second round of the presidential election in 2018 in the Czech Republic with two candidates Miloš Zeman and Jiří Drahoš. The results of Miloš Zeman as percentages are used over the individual $n = 77$ Czech counties (including Prague) as a response in (1) modeled by 4 predictors, i.e. with $p = 4$, evaluated for each of the counties:

- $X_1 =$ average wage in the fourth quarter of 2018;
- $X_2 =$ logarithm of the population density (people per km^2), according to the 2011 census;
- $X_3 =$ percentage of believers, according to the result of the 2011 census;
- $X_4 =$ percentage of people in execution in 2019.

The sources of each variable are specified in Figures 1 to 5. The variables are continuous except for X_3, for which we were only able to find the data dividing the counties to 5 groups. All computations in this paper were performed in R software (R Core Team, 2017).

The response is shown in Figure 1 with darker shades corresponding to a higher result; such map was obtained using the (very convenient) library RCzechia (version 1.6.2) of R software. Maps of the 4 predictors are shown as left images of Figures 2 to 5, where darker shades correspond to higher values of X_1, X_2, X_3, and X_4. For the sake of brevity, we skip the usual exploratory data analysis. At least, we would like to report the correlation matrix of the 4 variables, which equals

$$R = \begin{pmatrix} 1 & 0.34 & -0.15 & -0.18 \\ 0.34 & 1 & 0.05 & 0.15 \\ -0.15 & 0.05 & 1 & -0.58 \\ -0.18 & 0.15 & -0.58 & 1 \end{pmatrix}. \tag{6}$$

3 Beta regression

Beta regression is a generalized linear model suitable for data with a response in the form of percentages. We consider a beta regression model, where Y is used as the response of 4 predictors. All predictors turn out to be statistically significant using the Wald test on the level $\alpha = 0.05$. Particularly, the p-values of the 4 predictors are $0.04, 5 \cdot 10^{-5}, 8 \cdot 10^{-5}$, and $3 \cdot 10^{-10}$. The pseudo-R^2 in the model is 0.52. The mean square error, i.e. the common measure of prediction ability of (not only) the beta regression is presented in Table 1; there, values of the response were taken as percentages (with e.g. 50 per cent considered as to 0.50).

Tab. 1: Mean square error evaluated for the regression model $Y \sim X_1 + X_2 + X_3 + X_4$.

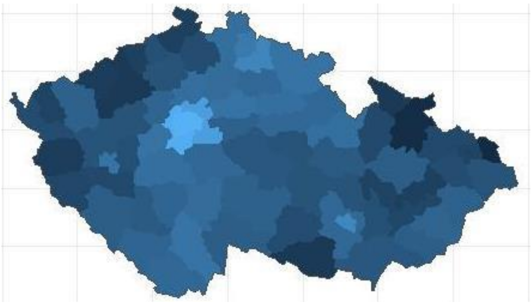
Regression estimator	MSE	Package in R software
Beta regression	25.4	betareg
Least squares in (1)	25.6	base
L1-estimator in (1)	27.7	quantreg
MM-estimator in (1)	26.0	rrcov

Source: own computations

4 Linear regression

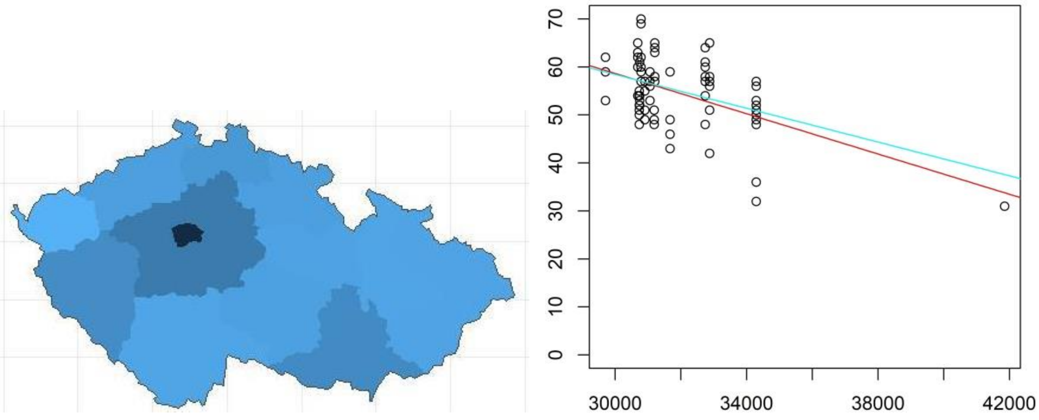
Right images of Figures 2 to 5 show the least squares estimates in the linear model of the response against each of the individual predictors separately. Some of the models turn out to be heteroscedastic, which advocates considering regression quantiles for the modeling. In Figure 2, we also show a LWS fit (cyan) to reveal that the outlying county (Prague) does not have a leverage effect on the regression fit. The coefficient of determination in the linear regression model with all 4 predictors jointly is equal to $R^2 = 0.52$. As Table 1 reveals, beta regression is slightly better than the least squares estimator in (1) in terms of the mean square error. It performs also better than the very robust MM-estimate.

Fig. 1: Map of the Czech Republic shaded according to Y, i.e. percentage of popular vote in each county for Miloš Zeman.



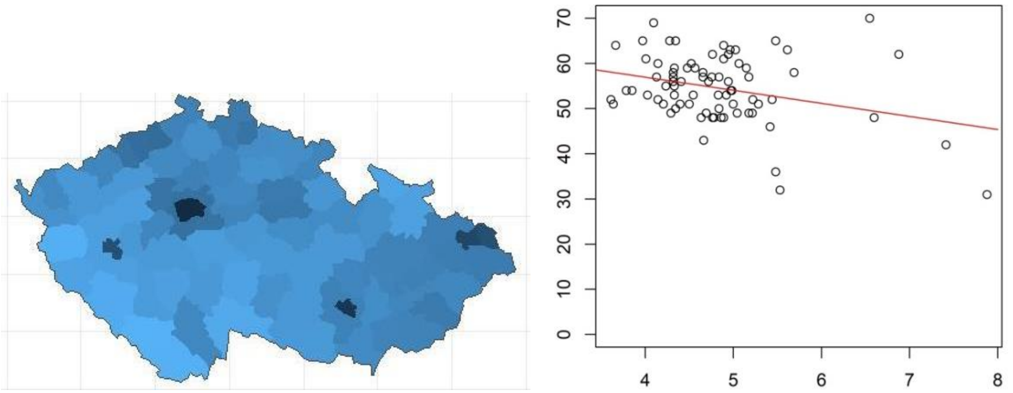
Source of the data: https://www.czso.cz/csu/czso/volba-prezidenta-republiky-2018

Fig. 2. Left: Map shaded according to X_1. Right: Y against X_1 (with a linear trend estimated by least squares).



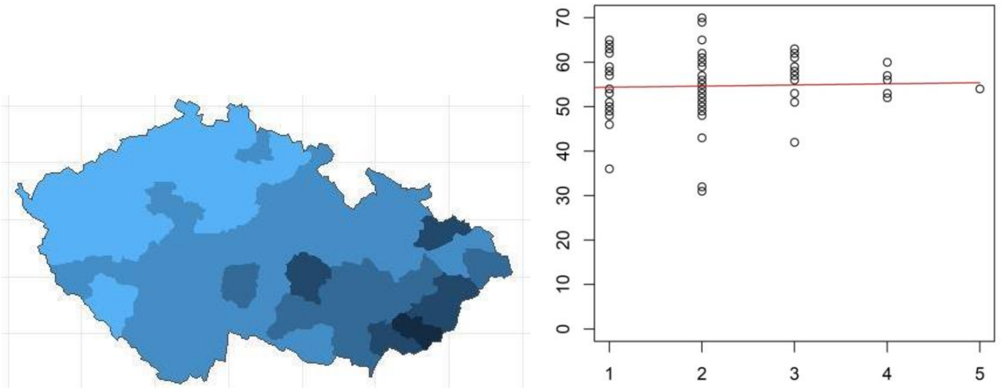
Source of the data: https://www.czso.cz/csu/xb/prumerna-mzda-ve-4-ctvrtleti-2018-a-v-roce-2018

Fig. 3. Left: Map shaded according to X_2. Right: Y against X_2 (with a linear trend estimated by least squares).



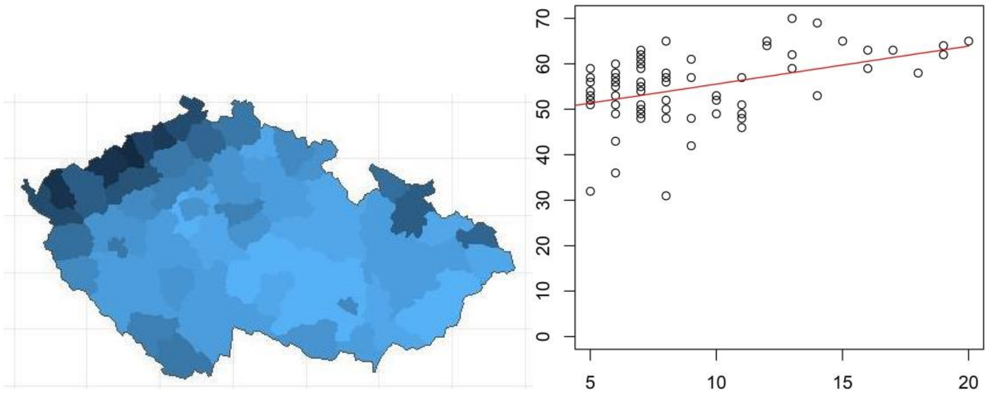
Source of the data: https://cs.wikipedia.org/wiki/Seznam_okresů_v_Česku

Fig. 4. Left: Map shaded according to X_3. Right: Y against X_3 (with a linear trend estimated by least squares).



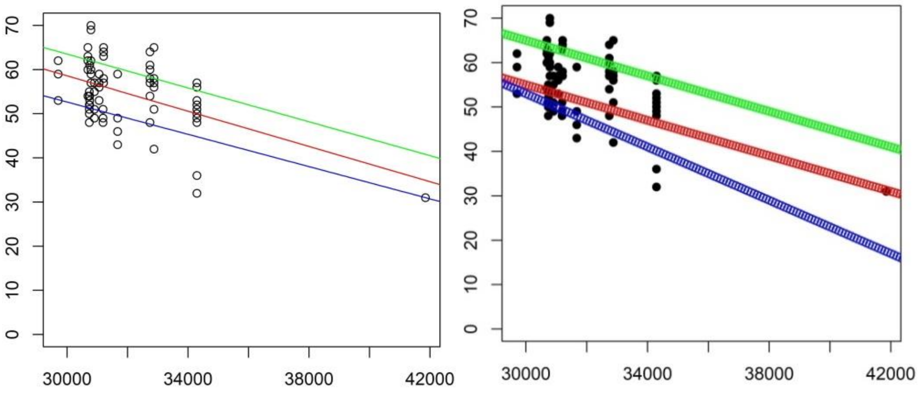
Source of the data: Růžičková (2014)

Fig. 5. Left: Map shaded according to X_4. Right: Y against X_4 (with a linear trend estimated by least squares).



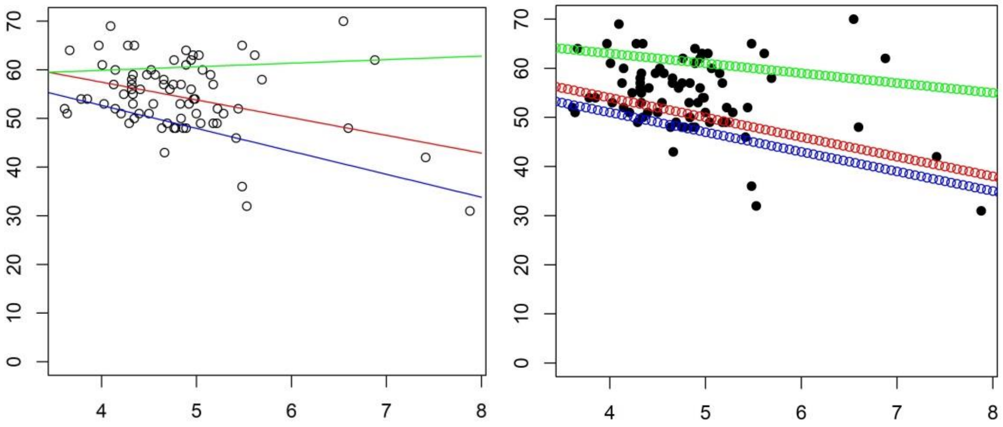
Source of the data: http://www.mapaexekuci.cz

Fig. 6. Regression quantiles (left) and IWRQ (right) in the model $Y \sim X_1$.



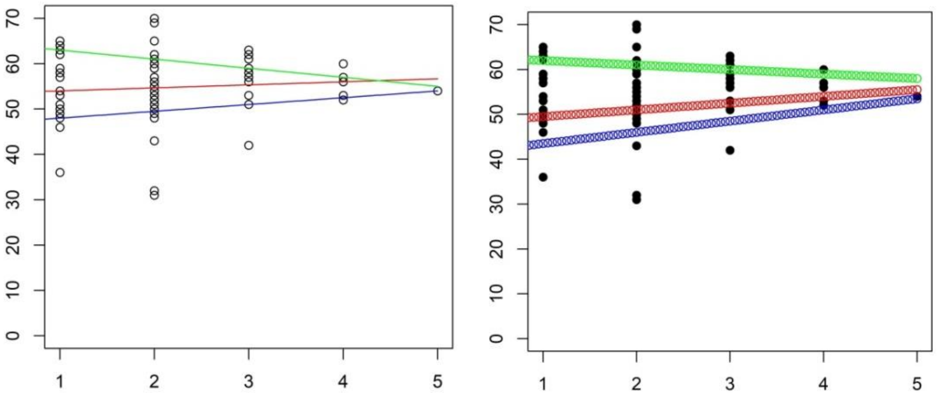
Source: own computations

Fig. 7. Regression quantiles (left) and IWRQ (right) in the model $Y \sim X_2$.



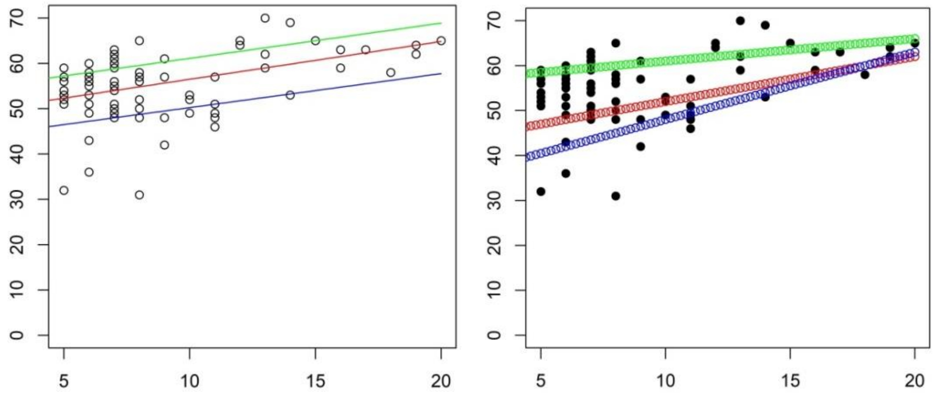
Source: own computations

Fig. 8. Regression quantiles (left) and IWRQ (right) in the model $Y \sim X_3$.



Source: own computations

Fig. 9. Regression quantiles (left) and IWRQ (right) in the model $Y \sim X_4$.



Source: own computations

5 Standard and implicitly weighted regression quantiles

Assuming the linear model with all 4 predictors jointly, we compute standard regression quantiles using the library quantreg of R software with $\tau = 0.2, \tau = 0.5$, and $\tau = 0.8$; these are shown in the left images of Figures 6 to 9. Further, we used $c = 0.005, c = 1$, and $c = 5$ to compute IWRQ; the obtained quantiles are shown in the right images of Figures 6 to 9.

Conclusions

Political scientists may use quintile regression to find interesting connections between the election results and regional economic characteristics on the level of individual counties. In comparison to modeling by the least squares estimator, regression quantiles may bring additional knowledge compared to that following from spatial statistical methods exploited e.g. by Maškarinec (2017). This paper compares the performance of standard regression quantiles with the performance of the recently proposed IWRQ procedure. The results of

IWRQ are meaningful for heteroscedastic regression models. They seem robust with respect to leverage points, while yielding similar results to standard regression quantiles over data without leverage points. Further, the IWRQ computation does not suffer from convergence problems, although we use in fact also categorized predictors and only $n = 77$ counties. For comparison, it would be also interesting to analyze the election data also by robust regression methods (e.g. of those overviewed in Kalina (2013)) or geographically weighted regression.

Acknowledgment

The work was supported by the project GA21-05325S of the Czech Science Foundation.

References

Bleik, J.M. (2019): Fully Bayesian estimation of simultaneous regression quantiles under asymmetric Laplace distribution specification. *Journal of Probability and Statistics*, 2019, Article 8610723.

Davino, C., Dolce, P., Taralli, S., Vinzi, V.E. (2018): A quantile composite-indicator approach for the measurement of equitable and sustainable well-being: A case study of the Italian provinces. *Social Indicators Research*, 136, 999-1029.

Hlubinka, D. & Šiman, M. (2020): Parametric elliptical regression quantiles. *REVSTAT—Statistical Journal*, 18, 257-280.

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

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

Kalina, J. & Vidnerová, P. (2019): Implicitly weighted robust estimation of quantiles in linear regression. In Houda, M. & Remeš, R., eds.: *37th International Conference on Mathematical Methods in Economics 2019*, České Budějovice: University of South Bohemia, 25-30.

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

Maškarinec, P. (2017): A spatial analysis of Czech parliamentary elections, 2006-2013. *Europe-Asia Studies*, 69, 426-457.

Ngoo, Y.T., Tan E.C., & Tey, N.P. (2020): Determinants of life satisfaction in Asia: Quantile regression approach. *Journal of Happiness Studies.* In press.

Rousseeuw, P.J. & Van Driessen, K. (2006): Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12, 29-45.

Růžičková, M. (2014): Regional differentiation of the population in the Czech Republic according to religious creed. Bachelor thesis, Prague: Charles University, Faculty of Science. (In Czech.)

Víšek, J.Á. (2011): Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47, 33-49.

Wu, S. & Guo, J. (2017): The Bayesian quantile regression and rough set classification. Taiwanese satisfaction level analysis. *Kybernetes*, 46, 1262-1274.

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
& The Czech Academy of Sciences, Institute of Information Theory and Automation
Pod Vodárenskou věží 4, 182 00 Prague 8, Czech Republic
kalina@cs.cas.cz

Petra Vidnerová
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
petra@cs.cas.cz