ON THE EFFECT OF HUMAN RESOURCES ON TOURIST INFRASTRUCTURE: NEW IDEAS ON HETEROSCEDASTIC MODELING USING REGRESSION QUANTILES

Jan Kalina – Patrik Janáček

Abstract

Tourism represents an important sector of the economy in many countries around the world. In this work, we are interested in the effect of the Human Resources and Labor Market pillar of the Travel and Tourism Competitiveness Index on tourist service infrastructure across 141 countries of the world. A regression analysis requires to handle heteroscedasticity in these data, which is not an uncommon situation in various available human capital studies. Our first task is focused on testing significance of individual variables in the model. It is illustrated here that significance tests are influenced by heteroscedasticity, which remains true also for tests for regression quantiles or robust regression estimators, resistant to a possible contamination of data by outliers. Only if a suitable model is considered, which takes heteroscedasticity into account, the effect of the Human Resources and Labor Market pillar turns out to be significant. Further, we propose and present a new diagnostic tool denoted as a quintile plot, allowing to interpret immediately the heteroscedastic structure of the linear regression model for possibly contaminated data.

Key words: tourism infrastructure, human resources, regression, robustness, regression quantiles

JEL Code: C45, Z32, C21

Introduction

Sociologists, demographers, political scientitsts or managers are often interested in evaluating and/or investigating human capital. As there is a clear evidence that human capital is correlated with economic growth (Silva et al., 2018), economists and politicians in various countries around the world have been searching for ways for improving human capital. Regression modeling was many times used in various quantitative studies of human capital. To give only a few examples, Qin (2017) considered a regression model of the influence of China's one-child policy on the long-term acculumulation of human capital. Blatná (2019)

analyzed the risk of poverty in the Czech Republic in a regression model containing various macro-economic indicators as regressors.

Linear regression represents a very popular model for human capital studies, although nonlinear regression applications appear in the field as well. For example, Almeida & Azkune (2018) used neural networks to monitor and predict human behavior (actions and activities) in time. Nonlinear regression modeling by means of logistic regression was used by Rombaut and Guerry (2018) to predict voluntary turnover, i.e. the attempt of employees to leave their work, in companies using data in the internal human resources database. From the statistical point of view, assumptions of the standard linear regression model are often violated in human capital studies, as there naturally appears a larger variability in more developed countries (Kodila-Tedika & Asongu, 2016). Therefore, the regression models in such context require to handle heteroscedastic data, i.e. data violating the condition of the same variance for each of the random regression errors, for example in the situation when the variances of the random errors depend on one or more independent variables. Other important issues include robustness of the regression methods to outlying values (outliers), the need for diagnostic tools (hypothesis tests of heteroscedasticity or autocorrelation), robust methods for estimating and predicting time series or testing their stationarity, and many others.

Regression quantiles represent a popular tool in human capital investigations, as they are suitable for heteroscedastic modeling and at the same time robust to small changes of the data, although they do not possess a global robustness to the presence of severely outlying values in the data. Regression quantiles were used to model study the effect of social benefits on youth employment (Bargain and Doorley (2017), or to compare human capital variables before and after economic reforms in Brazil (Justo et al., 2018). Regression quantiles were also the main tool in the above-mentioned paper by Silva et al. (2018), who explained the relationship (correlation) between human capital and economic growth across 92 countries of the world. Kalina et al. (2019) used regression quantiles to model the tourism and travel competitiveness by means of various factors including human resources, while the main aim of the paper was however to present a review of methods suitable for heteroscedastic regression modeling (including model choice).

Section 1 of this paper recalls such regression methods, which will be used in the subsequent computations. Section 2 describes the dataset. Section 3 presents a study of testing the significance of human resources on the tourist service infrastructure, including tests for robust estimators. Section 4 allows to obtain additional economic information by means of a newly proposed quintile plot suitable for heteroscedastic data.

1 Regression model and methods

This section presents regression methods used in our computations. Specific methods tailor-made for heteroscedastic data will be presented in Section 1.1. Let us first consider the standard linear regression model

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

where $Y_1, \ldots, Y_n$ are values of a continuous response variable and $e_1, \ldots, e_n$ are random errors with a common value of $var\ e_i = \sigma^2$, where $\sigma > 0$. There are p regressors in the model, and the vector of observed values of the i-th observation will be denoted as x_i for $i = 1, \ldots, n$. Possible estimators of the parameters of β include the least squares (LS), which is sensitive to the presence of outliers in the data, or various robust alternatives (Jurečková et al., 2019).

The least weighted squares (LWS) with linear weights of Víšek (2011), which is a promising (possibly highly) robust estimator based on implicit weights assigned to individual observations. The LWS estimator is robust with respect to the presence of outliers in the data, while the estimator is not fully free from any assumptions (Kalina, 2014). Let us also recall that Víšek (2011) recommended the LWS estimator for heteroscedastic data, but this result considered a modified covariance matrix of the estimator (analogous to the White heteroscedasticity-consistent covariance matrix), i.e. the result should not be interpreted as robustness of the LWS estimator itself to (any form of) heteroscedasticity. In fact, diagnostic tools for the LWS have also been proposed (Kalina, 2015).

The lasso estimator, denoted here as LS-lasso, represents an L_1-regularized version of the least squares. The L_1 estimator, also known as the regression median, is the regression quantile computed with the parameter $\tau = 0.5$.

1.1 Heteroscedasticity model

If the model (1) is heteroscedastic, it is advisable to estimate parameters in an alternative model; such approach (model and/or estimator) is often called heteroscedastic regression, Aitken estimator, generalized econometric model, generalized least squares, or weighted least squares. In the computations, we work with the particular model

$$\frac{Y_i}{\sqrt{k_i}} = \frac{\beta_0}{\sqrt{k_i}} + \frac{\beta_1 X_{i1}}{\sqrt{k_i}} + \cdots + \frac{\beta_p X_{ip} Y_i}{\sqrt{k_i}} + \frac{e_i}{\sqrt{k_i}}, \quad i = 1, \ldots., n, \tag{2}$$

while we consider a quite usual choice

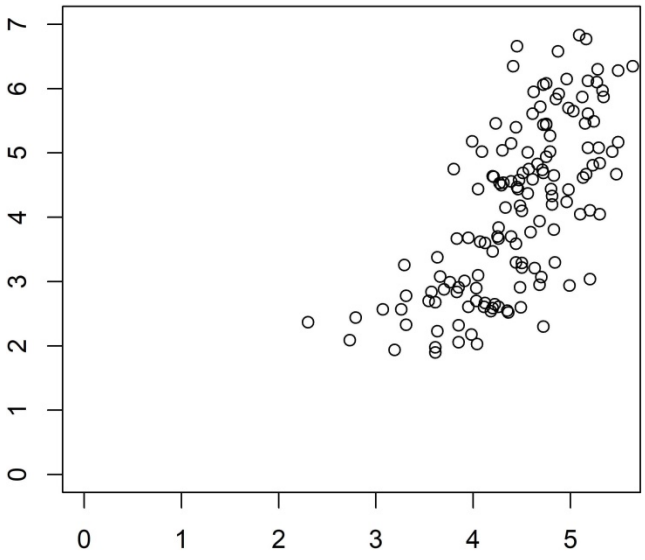
$$\sqrt{k_i} = \hat{Y}_i = b_0 + b_1 X_{i1} + \cdots + b_1 X_{i1}, \quad i = 1, \ldots n, \tag{3}$$

where $\left(b_0, b_1, \ldots, b_p\right)^T$ is the vector of estimates of β, obtained by the least squares, LTS or L_1 estimator in (3); if using a selected estimator in (1) to get (3), we use always this same selected estimator in the model (2). As we are interested only in testing (rather than predicting the response), we perform the significance tests also directly in (2).

2 Data description

Tourism, as an important sector of the economy (definitely in the pre-COVID era), is the subject of 14 characteristics of the World Economic Forum, yearly published within the Travel and Tourism Competitiveness Index (TTCI). The fourth pillar of TTCI is the Human Resources and Labor Market pillar, denoted here as X_4. The dataset from 2015, containing pillars $1, \ldots, 11$ and 13 (i.e. 12 pillars on the whole) as continuous regressors, was used in the modeling of the tourist service infrastructure (TSI) presented by Kalina et al. (2019). There, the dataset was shown to contain heteroscedastic errors (by means of asymptotic hetero-scedasticity tets) as well as outliers. Here, we use the same dataset and we consider the model (1), denoting the 12 regressors as $X_1, \ldots, X_{12}$ and using TSI as the continuous response variable. All the computations were performed in R software, version 4.0.0.

Fig. 1: Plot of the raw data. Horizontal axis: X_4 (the Human Resources and Labor Market). Vertical axis: the response Y.



Source: own computation

We are especially interested in the fourth pillar X_4 (Human Resources and Labor Market), which is obtained by aggregating 9 certain (although not specified in official materials of the World Economic Forum) macroecomic indicators. These should state how well human resources (in general terms of education and training) of a given country allow to use the skills in the labor market. Figure 1 shows the values of Y against values of X_4.

3 Study of significance of the Human Resources and Labor Market pillar

In this section, we investigate whether the effect of X_4 of TTCI on the tourist infrastructure across the world is statistically significant. As a novelty compared to the study of Kalina et al. (2019), here we perform the testing also for robust estimators in the transformed model (3). Table 1 presents the results for various methods described above in Section 1.

Particularly, testing for least squares is performed by *t*-tests. We use tests based on regression rank scores for the L_1 estimator, and nonparametric bootstrap tests for the LWS estimator. All these tests are performed within the standard backward elimination. Model choice for LS-lasso is performed by retaining only the regressors with non-zero estimates of the regression parameters, while the regularization parameter was found by a 10-fold cross validation. Significant results are denoted here with a star, if the p-value is below 0.05 but above 0.01; highly significant results with p-value below 10^{-3} are denoted by three stars. The p-value is not available in R software for some of the estimators, although the software states if the result is highly significant, significant, or non-significant. Table 1 also shows which package of R software was used for the computation, or it states that we used our own implementation.

As Table 1 reveals, the least squares as well as the robust LWS estimator are not able to determine X_4 as significant. LS-lasso and the L_1 estimator do find X_4 to be significant, as indicated by a star in Table 1. Especially the result of the L_1 estimator is important, as the method is popular for heteroscedastic data. The heteroscedastic models in the last three rows of Table 1 give very significant results for X_4, as indicated by three stars there. We can conclude the presented results by stating that heteroscedasticity has a strong (harmful) influence on standard significance tests in (1). This remains to be true even if robust estimation in (1) is used. On the whole, we may recommend to use heteroscedastic model (2) instead of (1), also if robust estimation is intended to be used. In other words, the robustness (to outliers) does not ensure robustness to heteroscedasticity. Such property is revealed here (to the best of our knowledge) as an original contribution not previously reported in the

literature, presumably because diagnostic tests (and subsequent alternative models) for robust regression have not acquired sufficient attention in the literature so far.

Tab. 1: Results of significance tests of Section 4.1, i.e. tests of significance of the effect of the Human Resources and Labor Market pillar on tourist infrastructure

Estimator (model)	Method	Significance	P-value (if available)	Implementation in R software
LS (1)	*t*-test	-	0.178	base
LWS (1)	Boostrap	-	0.093	own
L_1 (1) (regression median)	Regression rank scores	*	-	rq
LS-lasso (1)	Non-zero estimates of β	*	-	rq
LS in the heter. model (2)	*t*-test	***	$1.9 \cdot 10^{-10}$	own
LWS in the heter. model (2)	Boostrap	***	$6.4 \cdot 10^{-7}$	own
L_1 in the heter. model (2)	Regression rank scores	***		rq

Source: own computation

4 Quintile plot

As regression quantiles have been many times successfully applied to heteroscedastic data, we propose here to compute a novel type of a plot denoted as quintile plot with the aim to model and explain the dependence of the response (tourist infrastructure) on a single regressor (the Human Resources and Labor Market pillar) under the presence of heteroscedasticity. The computation of the quintile plot proceeds as follows:

1. The 4 regression quintiles, i.e. regression quantiles corresponding to the parameter (usually denoted by τ) equal to 0.2, 0.4, 0.6 and 0.8, are computed in the given regression model. Denote by $Q_1(x_i)$ the value of the first regression quintile evaluated for a given x_i, where $i = 1, \dots, n$. Use the notation $Q_2(x_i), Q_3(x_i)$ and $Q_4(x_i)$ for the second, third and fourth quintile.
2. The observations are divided to 5 parts according to the 4 quintiles (below the first quintile; between the first and the second; between the second and the third; between the third and the fourth; and above the fourth).

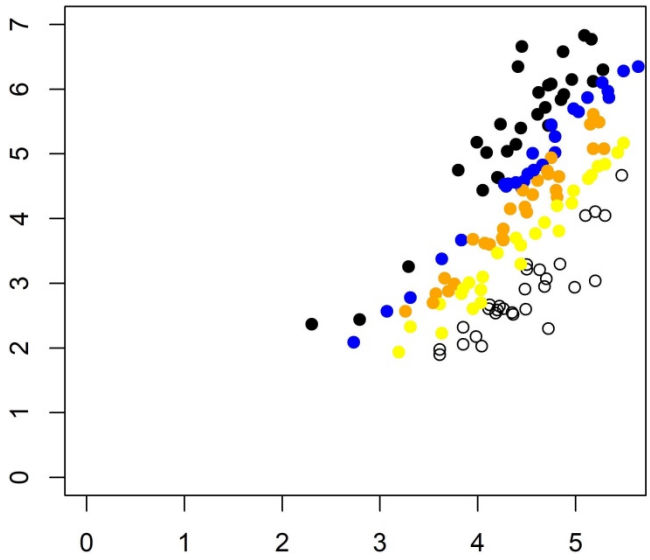
3. For each observation (denoting its regressors as x_i and the response as Y_i for $i = 1, \dots, n$), s is obtained as

$$s = \sum_{j=1}^{4} \mathbb{1}\left[Y_i \leq Q_j(x_i)\right], \tag{4}$$

where $\mathbb{1}$ denotes an indicator function.

4. Observations with $s = 1$ are shown in white color, $s = 2$ in yellow, $s = 3$ in orange, $s = 4$ in blue, and $s = 5$ in black.

Fig. 2: Quintile plot in the model $Y \sim X_4$. Horizontal axis: X_4 (the Human Resources and Labor Market). Vertical axis: the response Y.



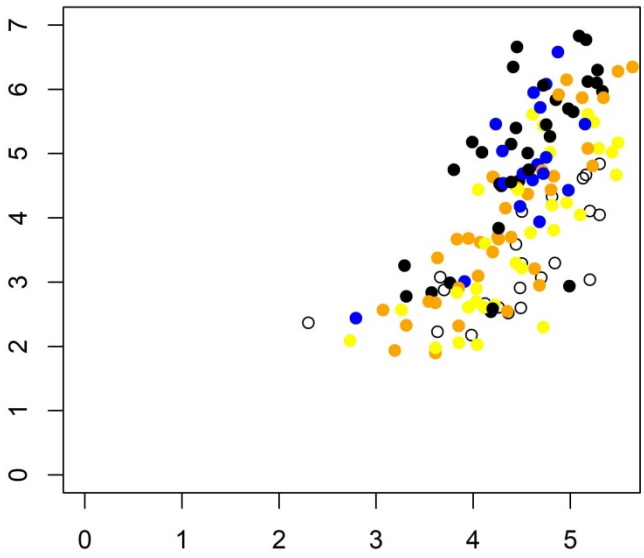
Source: own computation

The idea of dividing the observation is inspired by the regression median, which divides the data to 2 categories (below the median, and above). Step 4 is motivated by the lacking monotonicity of regression quantiles, i.e. we may sometimes encounter such phenomenon in real data that a given observation is e.g. above the first quintile, but below the second one, above the third one and finally below the fourth one.

Figure 2 shows the quintile plot for the simple model $Y \sim X_4$, while all 12 pillars are used as regressors in the quintile plot of Figure 3. Figure 2 is simple, but Figure 3 allows to present a multivariate knowledge clearly, bringing new interpretation. Figure 2 must be elegant with the strips of colors above each other, as no additional regressors are used in the model. Thus, Figure 2 corresponds to a standard plot of regression quantiles (in the form of regression lines), while the information in Figure 3 has to be more complex. Still, Figure 3

reveals a nice linear trend and at the same time reveals heteroscedasticity. In our opinion, the colors of all the observations are immediately apparent at one sight. The quintile plot can be computed also for other types of regression quintiles (e.g. nonlinear). The quintile plot can be also used as a diagnostic tool able to reveal if the linear model in (1) is not adequate.

Fig. 3: Quintile plot in the model $Y \sim X_1 + \cdots + X_{12}$. Horizontal axis: X_4 (the Human Resources and Labor Market). Vertical axis: the response Y.



Source: own computation

Conclusion

Improving human capital is known to be correlated with a growth of the national economy around the world. As our literature research indicates, various human capital studies work with heteroscedastic data require specific statistical tools for their analysis. This paper presents two particular studies performed over the TTCI dataset. As a methodological novelty, we propose and present here a quintile plot revealing the particular form of hetero-scedasticity in the relationship between the Human Resources and Labor Market pillar and the tourist service infrastructure. This plot shows the linear model to be meaningful and may find applications as a tool of regression diagnostics in linear regression.

The presented quintile plots may be exploited also within other types of regression quantiles. Although nonparametric regression quantiles available in the quantreg package of R software (function rqss) require the number of regressors not to exceed 2, alternative nonlinear regression quantiles can be computed exploiting neural networks within robust data

mining (Kalina, 2013). Such approach may be suitable also for modeling big heteroscedastic data, which slowly start to appear also in demographic research (Bohon, 2018).

Acknowledgment

The research was supported by the projects GA19-05704S and GA18-01137S of the Czech Science Foundation. Eva Litavcová and Petra Vašaničová provided the dataset.

References

Almeida, A. & Azkune, G. (2018). Predicting human behaviour with recurrent neural networks. Applied Sciences, 8, Article 305

Bargain, O. & Doorley, K. (2017). The effect of social benefits on youth employment: Combining RD and a behavioral model. Journal of Human Resources, 52, 1032-1059

Blatná, D. (2019). Risk of poverty rate by education in the Czech Republic in the period 2005-2017. RELIK 2019, Reproduction of Human Capital – Mutual Links and Connections. University of Economics, Prague, 53-63

Bohon, S.A. (2018). Demography in the big data revolution: Changing the culture to forge new frontiers. Population Research and Policy Review, 37, 323-341

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

Justo, W.R., Alencar, N.d.S., de Alencar, M.O., & Alves, D.F. (2018). Return on human capital: Quantile regression evidence in Brazil 2003-2013. International Journal of Finance and Accounting, 7, 153-159

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

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

Kalina, J. (2015). Three contributions to robust regression diagnostics. Journal of Applied Mathematics, Statistics and Informatics, 11 (2), 69-78

Kalina, J., Vašaničová, P., & Litavcová, E. (2019). Regression quantiles under heteroscedasticity and multicollinearity: Analysis of travel and tourism competitivenes. Ekonomický časopis, 67 (1), 69-85

Kodila-Tedika, O. & Asongu, S.A. (2016). Genetic distance and cognitive human capital: A cross-national investigation. Journal of Bioeconomics, 18, 33-51

Qin, X., Zhuang, C.C., & Yang, R. (2017). Does the one-child policy improve children's human capital in urban China? A regression discontinuity design. Journal of Comparative Economics, 45, 287-303

Rombaut, E. & Guerry, M.A. (2018). Predicting voluntary turnover through human resources database analysis. Management Research Review, 41, 96-112

Silva, F.R., Simoes, M., & Andrade, J.S. (2018). Health investments and economic growth: A quantile regression approach. International Journal of Development Issues, 17, 220-245

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

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
& Charles University, Faculty of Mathematics and Physics
Sokolovská 83, 186 75, Praha 8, Czech Republic
kalina@cs.cas.cz

Patrik Janáček
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
& Charles University, Faculty of Mathematics and Physics
Sokolovská 83, 186 75, Praha 8, Czech Republic
janacek@cs.cas.cz