The International Scientific Conference INPROFORUM 2017, *November 9, 2017, České Budějovice*, 361-365, ISBN 978-80-7394-667-8.

Various Approaches to Szroeter's Test for Regression Quantiles

Jan Kalina, Barbora Peštová

Abstract: *Regression quantiles represent an important tool for regression analysis popular in econometric applications, for example for the task of detecting heteroscedasticity in the data. Nevertheless, they need to be accompanied by diagnostic tools for verifying their assumptions. The paper is devoted to heteroscedasticity testing for regression quantiles, while their most important special case is commonly denoted as the regression median. Szroeter's test, which is one of available heteroscedasticity tests for the least squares, is modified here for the regression median in three different ways: (1) asymptotic test based on the asymptotic representation for regression quantiles, (2) permutation test based on residuals, and (3) exact approximate test, which has a permutation character and represents an approximation to an exact test. All three approaches can be computed in a straightforward way and their principles can be extended also to other heteroscedasticity tests. The theoretical results are expected to be extended to other regression quantiles and mainly to multivariate quantiles.*

Key words: Heteroscedasticity · Regression Median · Diagnostic Tools · Asymptotics

JEL Classification: C14 · C12 · C13

1 Introduction

Throughout this paper, the standard linear regression model

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

is considered, where $Y_1, \dots, Y_n$ are values of a continuous response variable and $e = (e_1, \dots, e_n)^T$ is the vector of random errors (disturbances). The task is to estimate the regression parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$. The least squares estimator assumes homoscedasticity, which is defined as a constant variance across observations, formally $var\ e_i = \sigma^2 > 0$ for each i. While the least squares estimate of β is sensitive to outliers as well as heteroscedasticity, there are numerous other estimates more suitable for non-standard situations (Matloff, 2017).

Regression quantiles represent a natural generalization of sample quantiles to the linear regression model. Their theory is studied by Koenker (2005). The estimator depends on a parameter α in the interval $(0,1)$, which corresponds to dividing the disturbances to $\alpha \cdot 100\ \%$ values below the regression quantile and the remaining $(1-\alpha) \cdot 100\ \%$ values above the regression quantile. In general, regression quantiles represent an important tool of regression methodology, which is popular in economic applications (Harrell, 2015; Vašaničová et al., 2017).

While the regression median (L1-estimator or least absolute deviation estimator) is much older than the methodology of regression quantiles, it can be defined as the regression quantile with $\alpha = 1/2$. The regression median has been at the same time investigated in the context of robust statistics because of its resistance against outlying values present in the response, which is an increasingly important property in econometrics (Kreinovich et al., 2017; Kalina, 2017).

The regression median, just like the least squares, requires to be accompanied by diagnostic tools for verifying its statistical assumptions. If homoscedasticity is violated, the confidence intervals and hypothesis tests for the regression median are namely misleading. In this work, we are interested in testing the null hypothesis (H_0) of homosceasticity against the alternative hypothesis (H_1) that the null hypothesis does not hold. In our previous work (Kalina, 2012), we derived an asymptotic test of heteroscedasticity for the least weighted squares (LWS) estimator of Víšek (2011).

This paper has the following structure. Section 2 recalls the Szroeter's test, which is one of available heteroscedasticity tests for the least squares. Further, three novel versions of the Szroeter's test are proposed for the regression median. These include an asymptotic test in Section 3, a permutation test in Section 4, and an (approximate) attempt for an exact test assuming normal errors in Section 5. An example with economic data is presented in Section 6 together with a proposal of an alternative model for the situation with a significant Szroeter's test. Finally, Section 7 concludes the paper and discusses future research topics including extension to other regression quantiles.

RNDr. Jan Kalina, Ph.D., Institute of Information Theory and Automation of the Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic & Institute of of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic, e-mail: kalina@utia.cas.cz

RNDr. Barbora Peštová, Ph.D., Institute of of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic, e-mail: pestova@utia.cas.cz

2 Szroeter's test for the least squares

Szroeter (1978) proposed a class of suitable tests of heteroscedasticity for the least squares estimator. The null hypothesis of homoscedasticity is tested against the alternative hypothesis

$$H_1: var\ e_i \leq var\ e_{i-1} \text{ with at least one sharp inequality, where } i = 2, \dots, n. \tag{2}$$

In this setup, the test is commonly used as a one-sided test. The user is required to select constants $h_1, \dots, h_n$ satisfying $h_i < h_j$ for $i < j$. The test statistic S is defined as the ratio of two quadratic forms

$$S = \frac{\sum_{i=1}^{n} h_i u_i^2}{\sum_{i=1}^{n} u_i^2} = u^T B u / u^T u, \tag{3}$$

where $(u_1, \dots, u_n)^T$ are residuals of the least squares fit and $B = diag\{h_1, \dots, h_n\}$. The test statistic S is scale-invariant, which is the crucial idea allowing the considerations in the next sections. The one-sided test rejects the null hypothesis of homoscedasticity in case $S > c$, where the critical value c depends on the particular choice of the constants $h_1, \dots, h_n$.

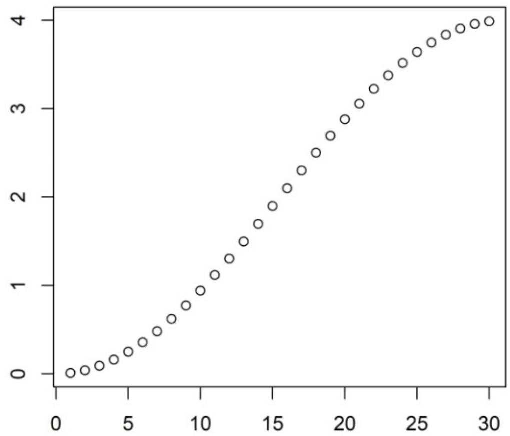
A popular choice for $h_1, \dots, h_n$ is to take

$$h_i = 2\left[1 - \cos\left(\frac{i\pi}{n+1}\right)\right], \quad i = 1, \dots, n, \tag{4}$$

leads to such form of the test, which has the same critical values as the Durbin-Watson test of independence of the disturbances e, as explained already in the original paper by Szroeter (1978). Values (4) are always strictly increasing, lie in $(0,4)$ and are illustrated in Figure 1 for $n = 30$. A less frequent possibility is to choose $h_1, \dots, h_n$ as indicators assigning data to groups similarly with the Goldfeld-Quandt test (Greene, 2011).

The Szroeter's test can be computed in a straightforward way and it is actually not needed to use the upper and lower bounds tabulated for the Durbin-Watson test. A critical value or a p-value for the least squares can be namely obtained directly, which is convenient because the test is not implemented e.g. in the popular R software.

Figure 1 Values of constants (4) for the Szroeter's test evaluated for $n = 30$. Horizontal axis: index 1,…,30. Vertical axis: values of h_i evaluated for $i = 1, \dots, 30$.



Source: Own processing

3 Asymptotic Szroeter's test for the regression median

As a first novel result, we describe an asymptotic Szroeter's test for the regression median. We formulate a theorem on the asymptotic behavior of the test statistic under H_0 and normally distributed errors e. Technical assumptions of Knight (1998), needed for the asymptotic representation of the regression median, will be denoted as Assumptions A. The asymptotic representation of Knight (1998) can be described as a special case of a more general representation for M-estimators derived by Jurečková et al. (2012). We will need the notation $M = I - X(X^T X)^{-1} X^T$, where I stands for the unit matrix with dimension nxn.

Theorem 1. Let us assume the errors e in (1) to follow a $N(0, \sigma^2)$ distribution with a constant variance $\sigma^2 > 0$. Let Assumptions A be fulfilled. Then the test statistic S evaluated for residuals of the regression median is asymptotically equivalent in probability with

$$e^T MBMe / e^T Me. \tag{5}$$

The proof follows from Kalina & Vlčková (2014), who proved the asymptotic equivalence of the Durbin-Watson test statistic computed with residuals of the LWS regression with the Durbin-Watson test statistic computed with residuals of the least squares regression. We note that (5) is the Szroeter's test statistic computed for residuals of the least squares estimator. In other words, the test is computed for the regression median in the same way as for the least squares, but such test holds exactly for the least squares and only asymptotically for the regression median.

4 Permutation Szroeter's test for the regression median

Next, we describe a test based on random permutations of residuals. The methodology of permutation tests for the least squares was summarized by Nyblom (2015). Here, we are interested in performing the total number of K permutations of residuals of the regression median.

For the k-th permutation ($k = 1, \dots, K$), the regression median is computed and the vector of its residuals is denoted as $\tilde{u}_k$. In addition, the test statistic

$$\tilde{u}_k^T B \tilde{u}_k / \tilde{u}_k^T \tilde{u}_k \tag{6}$$

is computed for each k. The empirical distribution over all K permutations is now considered to find the quantile c so that

$$\frac{1}{K} \sum_{k=1}^{K} I\left(\frac{\tilde{u}_k^T B \tilde{u}_k}{\tilde{u}_k^T \tilde{u}_k} > c\right) = 0.05, \tag{7}$$

where I denotes an indicator function. Then, the test rejects H_0 if the statistic S evaluated from residuals of the regression median is larger than the quantile c.

Permutation tests were proposed already by R.A. Fisher in 1930s. In general, not many theoretical results can be derived concerning properties of permutation tests (Pesarin & Salmaso, 2010). The test of this section is ensured to keep the level on the 5 % value and can be described as a nonparametric method, because it does not need any distributional assumption on the errors in the model (1). In other words, normality of residuals is not required.

5 Exact approximate approach to Szroeter's test for the regression median

Finally, we present an approximation to an exact Szroeter's test. Here, in contrary to a permutation test, random variables with the same distribution as the errors are repeatedly randomly generated. The idea here is to approximate directly the exact p-value of the test with an arbitrary precision assuming that the vector e comes from the normal distribution with zero expectation. By repeated random generating of mutually independent random variables $E_i \sim N(0,1)$ for $i = 1, \dots, n$, the exact p-value of the test against the one-sided alternative can be approximated by an empirical probability as evaluated in the following theorem.

Theorem 2. Let us assume (1) with independent identically distributed (i.i.d.) errors $e_1, \dots, e_n$ following $N(0, \sigma^2)$ distrbution with a specific $\sigma^2 > 0$. Let S denote the test statistic computed with residuals of the regression median and let $S_1^*, S_2^*, \dots$ denote values of (3) for independent realizations of independent random variables $E_1, \dots, E_n$ following $N(0,1)$ distribution. Then, it holds for $m \to \infty$ that

$$P\left(\frac{1}{m} \sum_{j=1}^{m} S_j^* \le x\right) \to P(S \le x) \quad \forall x \ge 0. \tag{8}$$

This aproximation, which states that the p-value can be approximated with an arbitrary precision, does not rely on the asymptotic behavior of the regression median. In other words, the empirical distribution obtained by the simulation allows to approximate the distribution of the Szroeter's test statistic S and theoretical probabilities can be estimated by their empirical counterparts. Here, the unit variance matrix of each E_i may be used valid without loss of generality thanks to the independence of the Szroeter's test statistic on β.

Finally, the next approximation between the p-value of the exact approximate test and the p-value of the least squares relies on the asymptotic results of Section 3 of this paper and represents a bridge connecting the exact testing with the asymptotic approach.

Theorem 3. The approximate p-value of the Szroeter's test for the regression median assuming errors $e_1, \dots, e_n$ following a $N(0, \sigma^2)$ distribution with a certain $\sigma^2 > 0$ converges with probability one to the p-value of the exact test for the regression median for $n \to \infty$ under Assumptions A.

6 Example

The performance of the novel tests will be now illustrated on a real economic data set. In this section, we also discuss a suitable estimation in the situation with a significantly heteroscedastic result of the regression median.

A gross domestic product (GDP) data set is analyzed which contains quarterly data from the first quarter of 1995 to the third quarter of 2007 measured in the USA in 10^9 USD, i.e. with $n = 50$. The data set was downloaded from website of the Federal Reserve Bank of St. Louis. The linear regression model has the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + e_i, \quad i = 1, \dots, n, \tag{9}$$

where Y is the GDP considered as a response of four regressors. Particularly, X_1 represents consumption, X_2 government expenditures, X_3 investments, and X_4 represents the difference between import and export.

Table 1 Estimated values of parameters in the example of Section 6

	β_0	β_1	β_2	β_3
Least squares estimator				
Linear regression model (11)	-3183	1.72	1.27	-8.24
Heteroscedastic model (12)	-52	0.35	0.24	-2.01
Regression median				
Linear regression model (11)	-2402	1.94	0.58	-10.68
Heteroscedastic model (12)	-56	0.39	0.18	-2.23

Source: Own processing

Table 2 Results of the Szroeter's test for the regression median

Linear regression model (11)	
Asymptotic test	$p = 0.0010$
Permutation test	$p = 0.0009$
Approx. exact test	$p = 0.0009$
Heteroscedastic model (12)	
Asymptotic test	$p = 0.07$
Permutation test	$p = 0.08$
Approx. exact test	$p = 0.08$

Source: Own processing

We estimate parameters of the model (10) by means of the least squares and the regression median. Tests of significance for both estimators reveal β_4 not to be significantly different from zero. Therefore, we reduce the model (9) to

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i, \quad i = 1, \dots, n. \tag{10}$$

Estimates of parameters obtained by the least squares and the regression median are shown in Table 1. Residuals of the least squares do not contain severe outliers but their distribution is far from unimodal. Moreover, the Shapiro-Wilk test of normality is rejected. Let us also check the assumption of homoscedasticity of the random errors. Let us now perform all three possible versions of the Szroeter‘s test with the choice (4). All these versions are significant, while approximate exact test and the permutation test yield a very similar result and the asymptotic test is slightly different from them.

Because the heteroscedasticity turns out to be significant, the question is how to find a more suitable model for explaining the response based on the regressors. While we have found no solution in references for the Szroeter' test, it remains possible (although not optimal) to consider the following model as a replacement of (10). The model

$$\frac{Y_i}{\sqrt{k_i}} = \frac{\beta_0}{\sqrt{k_i}} + \frac{\beta_1 X_{i1}}{\sqrt{k_i}} \dots + \frac{\beta_p X_{ip}}{\sqrt{k_i}} + \frac{e_i}{\sqrt{k_i}}, \quad i = 1, \dots, n, \tag{11}$$

is used with the choice $k_i = \hat{u}_i^2$ exploiting estimated values $u_1^2, \dots, u_n^2$ from an auxiliary model

$$u_i^2 = \gamma_0 + \gamma_1 X_{i1} + \dots + \gamma_p X_{ip} + v_i, \quad i = 1, \dots, n, \tag{12}$$

where $u_1, \dots, u_n$ are residuals computed in (10). Such heteroscedastic model (11) is an analogy of a so-called heteroscedastic regression (Greene, 2011) for models with a significant result of the Goldfeld-Quandt or Breusch-Pagan test.

We estimated parameters in this heteroscedastic model and the results are shown again in Table 1. Further, all versions of the Szroeter's test are performed again. The results are not significant any more. Such positive result (definitely not theoretically guaranteed) allows us to claim that parameters can be estimated better in the model (11) compared to model (10). We can say again that the approximate exact test and the permutation test yield a very similar result, while the asymptotic test is slightly different from them.

To summarize the computations, the standard linear regression model is not adequate due to a severe heteroscedasticity of the random regression errors. Only in a specific model tailor-made for heteroscedastic errors, the assumption of homoscedasticity of the errors is fulfilled by means of all versions of the Szroeter‘s test. The final model (11) considers weighted values of the response as well as regressors and therefore the interpretation of its parameters remains uncomparable to that of the original models (9) or (10).

7 Conclusions

This paper is devoted to testing heteroscedasticity for the regression median. While this estimator as a special case of regression quantiles is commonly used for analyzing heteroscedastic data, it itself requires to be accompanied by a test of heteroscedasticity.

We propose three different tests for the regression median in Sections 3, 4 and 5. Surprisingly, although the asymptotic representation for the regression median has been derived already in 1998, our asymptotic test for the regression median is novel. We are also not aware of any other approaches to the Szroeter's test for the regression median. The novel tests can be characterized as modifications of the basic Szroeter's test statistic, which has in the standard form the same critical values as the Durbin-Watson test of autocorrelated residuals, which is very popular in econometrics. The asymptotic test as well as the approximate exact test require to assume normally distributed errors, which is not the case of the (nonparametric) permutation test.

In an example, we illustrate the performance of various versions of the Szroeter's test for the regression median. In addition, the example shows that an attempt for an alternative estimation procedure in the form of a heteroscedastic regression allows to give a reasonable solution, preferable to a standard linear regression model.

Heteroscedasticity can be also revealed by regression quantiles. The methodology of regression quantiles, which may be used not only for a subjective detection but also for rigorous testing of heteroscedasticity by means of regression rank scores. While our attempt was to investigate diagnostic tools for regression quantiles from a more general perspective, the proposed tests cannot be directly applied to regression quantiles because the distribution of residuals would be (perhaps highly) asymmetric for a regression α-quantile with $\alpha \neq 1/2$.

As a future research, an intensive simulation study is needed to investigate the speed of asymptotics. Other simulations are needed to study the performance of the tests under H_0 (also for non-normal errors) and also under various forms of heteroscedasticity, i.e. under the alternative hypothesis. It would be also possible to extend the novel approaches to tests for other regression estimates, e.g. for linear regression with a multivariate response, for which there seem no diagnostic tools available.

On the whole, the approaches of the current paper can be interpreted as a preparation for a generalization to the context of elliptical quantiles. Such generalization of the Szroeter's test to elliptical quantiles may be performed in a rather straightforward way. Such future tests may find applications in testing heteroscedasticity for linear regression models with a multivariate response, which have recently penetrated to econometric modelling. The use of any form of multivariate quantiles seems a promising tool as the quantiles are heavily influenced by a possible heteroscedasticity. Such idea may be especially valuable if elliptical quantiles studied by Hlubinka and Šiman (2013) are used and volume of the constructed ellipses is used to build statistical decision rules for detecting heteroscedasticity and its subsequent modelling.

Acknowledgement

The research was supported by the Czech Science Foundation project „Nonparametric (statistical) methods in modern econometrics“ No. 17-07384S.

References

Greene, W.H. (2011). *Econometric Analysis*. 7th edition. Old Tappan: Prentice Hall.

Harrell, F.E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd edition. Cham: Springer.

Hlubinka, D., & Šiman, M. (2013). On elliptical quantiles in the quantile regression setup. *Journal of Multivariate Analysis*, 116, 163-171.

Jurečková, J., Sen, P.K., & Picek, J. (2012): *Methodology in robust and nonparametric statistics.* Boca Raton: CRC Press.

Kalina, J. (2017). High-dimensional data in economics and their (robust) analysis. *Serbian Journal of Management*, 12(1), 157-169.

Kalina, J., & Vlčková, K. (2014). Highly robust estimation of the autocorrelation coefficient. In *Proceedings 8th International Days of Statistics and Economics MSED 2014* (pp. 588-597), Melandrium, Slaný.

Kalina, J. (2012). Highly robust statistical methods in medical image analysis. *Biocybernetics and Biomedical Engineering*, 32(2), 3-16.

Knight, K. (1998). Limiting distributions for L1 regression estimators under general conditions. *Annals of Statistics*, 26(2), 755-770.

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

Kreinovich, V., Sriboonchitta, S. & Huynh, V.-N. (2017). *Robustness in econometrics*. Cham: Springer.

Matloff, N. (2017). *Statistical regression and classification. From linear models to machine learning.* Boca Raton: CRC Press.

Nyblom, J. (2015). Permutation tests in linear regression. In K. Norhausen & S. Taskinen (Eds.), *Modern nonparametric, robust and multivariate methods* (pp. 69-90). Cham: Springer.

Pesarin, F. & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. New York: Wiley.

Szroeter, J. (1978). A class of parametric tests of heteroscedasticity in linear econometric models. *Econometrica*, 46(6), 1311-1327.

Vašaničová, P., Litavcová, E., Jenčová, S., & Košíková, M. (2017). Dependencies between travel and tourism competitiveness sub-indexes: The robust quantile regression approach. *Proceedings 11th International Days of Statistics and Economics MSED 2017* (pp. 1729-1739). Melandrium, Slaný.

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