EXACT INFERENCE IN ROBUST ECONOMETRICS UNDER HETEROSCEDASTICITY

Jan Kalina − Barbora Peštová

Abstract

The paper is devoted to the least weighted squares estimator, which is one of highly robust estimators for the linear regression model. Novel tests of heteroscedasticity are proposed, which have the form of a permutation Goldfeld-Quandt test and a permutation Breusch-Pagan test. Moreover, the asymptotic behavior of these permutation tests is investigated. Newly formulated theorems study the convergence of the tests to asymptotic tests based on the least weighted squares as well as to the test based on the least squares. A numerical experiment on a real economic data set on gross domestic product is presented, which also shows how to perform a robust prediction model under heteroscedasticity. The results are very different from those obtained with a standard estimation procedure. Various tests yield however rather similar results. Thus, taking the heteroscedasticity into account is very desirable, while the choice of a particular testing or estimation approach is not so crucial. Theoretical results may be simply extended to the context of multivariate quantiles.

Key words: heteroscedasticity, robust statistics, regression, diagnostic tools, economic data

JEL Code: C14, C12, C13

Introduction

Robust statistical estimators in the linear regression model are well known to require their own diagnostic tools (Víšek, 2010). One of promising highly robust estimator in the linear regression context is the least weighted squares (LWS) estimator of Víšek (2002), which does not trim away (i.e. ignore) outliers, but rather only downweights potential candidates for outliers. The estimator possesses a high breakdown point, which can be interpreted as a high resistance (insensitivity) against outlying measurements in the data and one of crucial measures of robustness of statistical estimators (Jurečková et al., 2012).

In our previous work, we proposed and investigated asymptotic diagnostic tests for the LWS including tests of heteroscedasticity (Kalina, 2009), which is one of the assumptions on random regression errors in the linear regression model. Other results include hypothesis tests

for the LWS estimator or a corresponding robust correlation coefficient (Kalina & Schlenker, 2015). However, it remains open to propose permutation tests for the same tasks and to study their asymptotic behavior. Permutation tests can be described as simple and comprehensible Monte Carlo procedures, which however require intensive computations. They can also be interpreted as an important class of resampling methodology (without replacement), which contains a wide range of different flexible tools (Efron & Tibshirani, 1994). In addition, permutation tests can be considered a nonparametric technique as investigated theoretically by Pesarin & Salmaso (2010). The only (but crucial) assumption is the exchangeability of individual observations. Sometimes, permutation tests are also called invariance tests or conditional tests, where the latter concept stresses conditioning of the procedure by the observed data.

This paper is devoted to the question of verifying the assumption of homoscedasticity for the LWS estimator. A possible violation of homoscedasticity influences the least squares as well as robust estimators. A permutation approach to testing is considered to propose new heteroscedasticity tests for the LWS estimator in Sections 2 and 3. There, also the asymptotic behavior of the permutation test statistics is investigated. Both tests are illustrated on a real data set in Section 4. Finally, Section 5 concludes the paper.

1 Least weighted squares

Throughout this paper, the standard linear regression model

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

is considered, where $Y_1, \ldots, Y_n$ are values of a continuous response variable and $e_1, \ldots, e_n$ are random errors (disturbances). The task is to estimate the regression parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$. While the classical least squares estimator denoted as b_{LS} is very well known to be too vulnerable to the presence of outlying measurements (outliers) in the data, robust statistical methods are available as alternative estimation procedures for this task.

The definition of the LWS estimator first requires the user to specify a sequence of magnitudes of weights

$$w_1 \geq w_2 \geq \cdots \geq w_n, \tag{2}$$

which are assigned to individual observations only after some permutation. While the selection of the weights influences the result, it may a good choice to use linearly decreasing weights or to allow some percentage of observations to have a zero weight to ensure high robustness. The

LWS estimator remains consistent for weights generated by any weight function which is continuous and nonincreasing (Víšek, 2010).

Let us denote the squared residuals corresponding to a given estimator b of β as

$$u^2_{(1)}(b) \leq u^2_{(2)}(b) \leq \cdots \leq u^2_{(n)}(b). \tag{3}$$

The LWS estimator b_{LWS} is defined as

$$\arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} w_i u^2_{(i)}(b). \tag{4}$$

The computation of the LWS estimator is intensive and an approximate algorithm can be obtained as a weighted version of the FAST-LTS algorithm proposed for the least trimmed squares (LTS) regression (Kalina, 2009).

Čížek (2011) proposed a two-stage LWS estimator denoted as 2S-LWS and proved it to possess a high breakdown point and at the same time a 100 % asymptotic efficiency of the least squares under Gaussian errors. Further, he evaluated its relative efficiency to be high (over 85 %) compared to maximum likelihood estimators in a numerical study under various distributional models for samples of several tens of observations. The computation of the estimator starts with an initial highly robust estimator and proceeds to proposing values of the weights based on comparing the empirical distribution function of squared residuals with its theoretical counterpart assuming normality.

2 A permutation Goldfeld-Quandt test for the LWS estimator

Standard tests of heteroscedasticity for the least squares include e.g. those of Goldfeld-Quandt or Breusch-Pagan (Greene, 2002). An approximation to the permutation Goldfeld-Quandt test is proposed in this section, particularly a convergence of the p-value will be investigated.

Asymptotic tests for the LWS estimator were presented by Kalina (2009), who derived asymptotic test statistics for the LWS residuals. Analogous results were presented for regression quantiles (Kalina, 2011). In both these contexts, there were however two different versions of the approximation described and the level of convergence for both of them seems rather slow as indicated by numerical simulations. So far, we are not aware of a permutation (exact) test of heteroscedasticity for the LWS. There seems neither any combination of permutation tests with the asymptotic approximation in this context.

The Goldfeld-Quandt test considers the null hypothesis

$$H_0: var\ e_i = \sigma^2, \quad i = 1, \dots n, \tag{5}$$

and divides the observations to three groups. We consider a (one-sided) alternative hypothesis H_1 that the third part of the data has a larger variability than the first part. Assuming (1), let r_1 denote the number of observations in the first group and r_3 in the third group. Residuals corresponding to the LWS estimator b_{LWS} of β will be denoted as

$$u_{LWS} = (u_1^{LWS}, \dots, u_n^{LWS})^T . \tag{6}$$

The residual sum of squares in the first group of the data will be denoted by SSE_1, while the residual sum of squares computed in the third group by SSE_3 These quantities

$$SSE_1 = \sum_{i=1}^{r_1} \left(u_i^{LWS}\right)^2 \quad \text{and} \quad SSE_3 = \sum_{i=r_1+r_2+1}^{r_3} \left(u_i^{LWS}\right)^2 \tag{7}$$

representing a natural extension from the least squares case allow to form the test statistic

$$F_{LWS} = \frac{SSE_3}{SSE_1} \frac{r_1 - p}{r_3 - p}. \tag{8}$$

The permutation test will be based on a repated random generation of i.i.d. random variables $E_1, \dots, E_n$ following the normal distribution $N(0,1)$. These will be used to replace the errors $e_1, \dots, e_n$ within (7). For the j-th simulation ($j = 1, \dots, m$ for some m), residuals of the LWS fit will be computed and used to construct the statistic (8) denoted by F_j^*, where the star corresponds to the common way for denoting a resampling context already since Efron & Tibshirani (1994). The averaged value of these test statistics converges as follows.

Theorem 1. Let us assume (1) with i.i.d. errors $e_1, \dots, e_n$ following $N(0, \sigma^2)$ distribution with a specific $\sigma^2 > 0$. Let F_{LWS} denote the test statistic (8) computed with the LWS residuals and let $F_1^*, F_2^*, \dots$ denote values of (8) for independent realizations of independent random variables $E_1, \dots, E_n$ following a $N(0,1)$ distribution. Then, it holds for $m \to \infty$ that

$$P\left(\frac{1}{m}\sum_{j=1}^{m} F_j^* \le x\right) \to P(F_{LWS} \le x) \quad \forall x \ge 0. \tag{9}$$

Theorem 2. Let us assume (1) with i.i.d. errors $e_1, \dots, e_n$ following $N(0, \sigma^2)$ distribution with a specific $\sigma^2 > 0$. Let F denote the test statistic (8) computed with the least squares residuals and let $F_1^*, F_2^*, \dots$ denote values of (8) for independent realizations of independent random variables $E_1, \dots, E_n$ following the $N(0,1)$ distribution. Then, it holds for $m \to \infty$ that

$$P\left(\frac{1}{m}\sum_{j=1}^{m} F_j^* \leq x\right) \to P(F \leq x) \quad \forall x \geq 0. \tag{10}$$

While the proof of Theorem 1 follows directly from elementary principles of permutation tests, Theorem 2 is a consequence of asymptotic results of Kalina (2009) and the asymptotic representation of the LWS estimator given by Víšek (2002).

We can understand the approach of Theorem 2, which stands on stronger asymptotic results than Theorem 1, as an exact version of the asymptotic test, exploiting the convergence of the test to the *p*-value of the exact test for the least squares residuals. The test based on Theorem 2 does not require an additional simulation for assessing the null distribution of F_{LWS}, but can use a simple approach exploiting the known null distribution of F to be Fisher's F distribution with $r_3 - p$ and $r_1 - p$ degrees of freedom.

Let us give a remark to the normal distribution of $E_1, \dots, E_n$. The result of the permutation test is invariant to the choice of variance of this normal distribution and the unit variance may be chosen without loss of generality.

3 A permutation Breusch-Pagan test for the LWS estimator

In the same spirit as the Goldfeld-Quandt test, also the permutation (exact) version of the Breusch-Pagan test for the LWS estimator can be constructed.

Breusch-Pagan test requires to specify the alternative hypothesis of heteroscedasticity as

$$var\ e_i = \alpha_0 + \alpha_1 Z_{i1} + \cdots + \alpha_K Z_{iK}, \quad i = 1, \dots, n, \tag{11}$$

for some variables

$$Z_1 = (Z_{11}, \dots, Z_{n1})^T, \dots, Z_K = (Z_{1K}, \dots, Z_{nK})^T. \tag{12}$$

The null hypothesis corresponds to

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_K = 0, \tag{13}$$

which is tested against a general alternative hypothesis that the null hypothesis is not true. Often, one or more regressors from (1) are selected as the auxiliary variables (12).

The test statistic is obtained as the statistic χ^2 of the score test (i.e. Lagrange multiplier test) in the model

$$u_i^2/s^2 = \alpha_0 + \alpha_1 Z_{i1} + \dots + \alpha_K Z_{iK} + v_i, \quad i = 1, \dots, n, \tag{14}$$

where s^2 is the estimator of σ^2. The score test is one of general asymptotic tests based on the likelihood function, in our case under the presence of nuisance parameters. Just like the standard Breusch-Pagan for least squares, the new test assumes normal distribution of e. The following theorems represents an analogy of Theorems 1 and 2 (with analogous proofs), exploits the score test statistics of the Breusch-Pagan test.

Theorem 3. Let us assume (1) with i.i.d. errors $e_1, \dots, e_n$ following $N(0, \sigma^2)$ distribution with a specific $\sigma^2 > 0$. Let χ^2_{LWS} denote the test statistic of the Breusch-Pagan test computed with the LWS residuals and let $\chi_1^{2*}, \chi_2^{2*}, \dots$ denote test statistics of the Breusch-Pagan test for independent realizations of independent random variables $E_1, \dots, E_n$ following the $N(0,1)$ distribution. Then, it holds for $m \to \infty$ that

$$P\left(\frac{1}{m}\sum_{j=1}^{m}\chi_j^{2*} \le x\right) \to P(\chi^2_{LWS} \le x) \quad \forall x \ge 0. \tag{15}$$

Theorem 4. Let us assume (1) with i.i.d. errors $e_1, \dots, e_n$ following $N(0, \sigma^2)$ distribution with a specific $\sigma^2 > 0$. Let χ^2 denote the test statistic of the Breusch-Pagan test computed with the least squares residuals and let $\chi_1^{2*}, \chi_2^{2*}, \dots$ denote test statistics of the Breusch-Pagan test for independent realizations of independent random variables $E_1, \dots, E_n$ following the $N(0,1)$ distribution. Then, it holds for $m \to \infty$ that

$$P\left(\frac{1}{m}\sum_{j=1}^{m}\chi_j^{2*} \le x\right) \to P(\chi^2 \le x) \quad \forall x \ge 0. \tag{16}$$

Like in Section 3, the test based on Theorem 4 exploiting the asymptotic behavior of the LWS estimator can use directly that χ^2 follows Pearson's χ^2 distribution with K degrees of freedom (in our notation) under the null hypothesis.

4 Example: A heteroscedastic model for GDP

The performance of the novel permutation tests will be now illustrated on a real economic data set. If the LWS regression fit in the original model is significantly heteroscedastic, a specific heteroscedastic regression model will be condidered.

A 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,

which comes originally from the Federal Reserve Bank of St. Louis, was analyzed by Špaková (2011), but only by standard (non-robust) methods.

The linear regression model has the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + e_i, \quad i = 1, \dots, n, \tag{17}$$

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. A graphical analysis reveals a relationship close to linear between the response and each of the regressors.

We estimate parameters of the model (17) by means of the least squares and the LWS. Residuals of the least squares do not contain severe outliers but their distribution is far from unimodal. Also the Shapiro-Wilk test of normality is rejected. Tests of significance for both estimators reveal β_4 not to be significantly different from zero. Therefore, we reduce the model (17) to a more suitable submodel.

Tab. 1: Results of the example of Section 7.

	β_0	β_1	β_2	β_3
Least squares estimator				
Linear regression model (15)	-3123	1.67	1.31	-8.04
Heteroscedastic model (16)	-52	0.32	0.27	-2.16
LWS estimator				
Linear regression model (15)	-2402	1.98	0.61	-10.88
Heteroscedastic model (16)	-57	0.39	0.14	-2.33

Source: own computation

The model under consideration has the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i, \quad i = 1, \dots, n, \tag{18}$$

while it remains important to check the assumption of homoscedasticity of the random errors. The results of the least squares and the LWS with linearly decreasing weights in (17) are shown in Table 2. Let us now perform both the asymptotic and exact version of the Breusch-Pagan test. The test statistic χ^2 for the LWS is equal to 11.94 and is significant with a *p*-value of 0.0010. The permutation test yields a *p*-value of 0.0009.

Because the heteroscedasticity turns out to be significant, we consider the following model as a replacement of (18). The model

$$\frac{Y_i}{\sqrt{k_i}} = \frac{\beta_0}{\sqrt{k_i}} + \frac{\beta_1 X_{i1}}{\sqrt{k_i}} \dots + \frac{\beta_p X_{ip}}{\sqrt{k_i}} + \frac{e_i}{\sqrt{k_i}}, \quad i = 1, \dots, n, \tag{19}$$

is used with the choice $k_i = u_i^2$ exploiting values $u_1^2, \dots, u_n^2$ of (14). Such heteroscedastic model for the LWS is an analogy of a model described by Greene (2002). In our example, the general approach (19) uses $p = 3$ and the three regressors were chosen to play the role of Z_1, Z_2 and Z_3 for the model (14). Further, the parameters in this heteroscedastic model were estimated and the results are shown again in Table 2. Further, asymptotic and exact tests of heteroscedasticity are performed again, which both yield unsignificant results. The p-value of the permutation Goldfeld-Quandt test equals 0.09 and of the Breusch-Pagan 0.07.

To summarize the computations, the standard linear regression model is not adequate due to a severe heteroscedasticity of the random regression errors. Not even a robust regression estimator is able to improve the quality of the model. Only in a specific model tailor-made for heteroscedastic errors, the assumption of homoscedasticity of the errors is fulfilled by means of both versions of the Breusch-Pagan test. The final model (19) 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 (17) or (18).

5 Conclusions

This paper is devoted to diagnostic tools for the LWS estimator in the linear regression model. While robust regression diagnostics has been declared as an important problem for in linear regression (Salini et al., 2016) as well as in related models (Kalina, 2012), asymptotic approximation for heteroscedasticity tests have been available so far (Kalina, 2009). This paper fills the gap of permutation versions of the Goldfeld-Quandt and Breusch-Pagan test for the LWS estimator.

Permutation tests are proposed in Sections 2 and 3 for heteroscedasticity tests based on LWS residuals. These can be described as approximations to exact Goldfeld-Quandt and Breusch-Pagan tests. While theoretical results are obtained for the probability of type I error, numerical simulations would be necessary to investigate the performance of the tests under the alternative hypothesis.

A permutation test gives only a p-value without a direct possibility to estimate the power of the test, which would be feasible in the asymptotic setup. This limits the usage of the tests to

situations when testing is the very aim of the analysis. The example described in Section 4 has its main to predict the response rather than to perform testing, but a suitable estimation method is presented as a tool for an improved prediction compared to a standard model ignoring the heteroscedasticity structure.

As an alternative, there remains also another possibility to use the White test of heteroscedasticity, which exploits a covariance matrix estimator of regression coefficients estimates. Víšek (2010) showed this estimator within the White test to be reasonably robust to heteroscedasticity. Still another possibility to modelling heteroscedasticity is to use the methodology of regression quantiles, which may be used not only for a subjective detection but also for rigorous testing of heteroscedasticity (Gutenbrunner and Jurečková, 1992).

Nevertheless, the approach of the current paper can be interpreted as a preparation for a generalization to the context of elliptical quantiles. Such generalization of the Goldfeld-Quandt and Breusch-Pagan tests 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. To the best of our knowledge, there have been however no diagnostic tools for linear regression with a multivariate response available. 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 (2015) are used and volume of the constructed ellipses is used to build statistical decision rules for detecting heteroscedasticity and its subsequent modelling.

Acknowledgment

The research was supported by the Czech Science Foundation project No. 17-07384S.

References

1. Čížek, P. (2011): Semiparametrically weighted robust estimation of regression models. *Computational Statistics & Data Analysis*, 55 (1), 774-788.
2. Efron, B., Tibshirani, R.J. (1994): *An introduction to the bootstrap*. Chapman & Hall/ CRC, Boca Raton.
3. Greene, W.H. (2002): *Econometric Analysis*. 5th edn. Macmillan, New York.
4. Gutenbrunner, C. & Jurečková, J. (1992): Regression rank scores and regression quantiles. *Annals of Statistics*, 20 (1), 305-330.

5. Hlubinka, D. & Šiman, M. (2015): On generalized elliptical quantiles in the nonlinear quantiles regression setup. *Test*, 24 (2), 249-264.
6. Jurečková, J., Sen, P.K., & Picek, J. (2012): *Methodology in robust and nonparametric statistics.* CRC Press, Boca Raton.
7. Kalina, J. (2012): Highly robust statistical methods in medical image analysis. *Biocybernetics and Biomedical Engineering*, 32 (2), 3-16.
8. Kalina, J. (2011): On heteroscedasticity in robust regression. *Conference Proceedings International Days of Statistics and Economics MSED* 2011, Melandrium, Slaný, 228-237.
9. Kalina, J. (2009): Least weighted squares in econometric applications. *Journal of Applied Mathematics, Statistics and Informatics*, 5 (2), 115-125.
10. Kalina, J. & Schlenker, A. (2015): A robust supervised variable selection for noisy high-dimensional data. *BioMed Research International*, 2015, Article 320385.
11. Pesarin, F. & Salmaso, L. (2010): *Permutation tests for complex data: Theory, applications and software*. Wiley, New York.
12. Salini, S., Cerioli, A., Laurini, F., & Riani, M. (2016): Reliable robust regression diagnostics. *International Statistical Review*, 84 (1), 99-127.
13. Špaková, M. (2011): *Testing heteroscedasticity*. Bachelor thesis, MFF UK, Prague.
14. Víšek, J.Á. (2010): Heteroscedasticity resistant robust covariance matrix estimator. *Bulletin of the Czech Econometric Society*, 17, 33-49.
15. Víšek, J.Á. (2002): The least weighted squares II. Consistency and asymptotic normality. *Bulletin of the Czech Econometric Society*, 9, 1-28.

Contact

Jan Kalina
Institute of Information Theory and Automation of the Czech Academy of Sciences
Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic
& Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
kalina@cs.cas.cz

Barbora Peštová
Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
pestova@cs.cas.cz