ISSN 2045-3345

ISSN 2045-3345

TESTING HETEROSCEDASTICITY IN ROBUST REGRESSION

Jan KALINA
Institute of Computer Science of the Academy of Sciences of the Czech Republic

ABSTRACT

This work studies the phenomenon of heteroscedasticity and its consequences for various robust estimation methods for the linear regression, including the least weighted squares, regression quantiles and trimmed least squares estimators. We investigate hypothesis tests for these regression methods and removing heteroscedasticity from the linear regression model. The new asymptotic heteroscedasticity tests for robust regression are asymptotically equivalent to standard tests computed for the least squares regression. Also we describe an asymptotic approximation to the exact null distribution of the test statistics. We describe a robust estimation procedure for the linear regression with heteroscedastic errors.

JEL CLASSIFICATION & KEYWORDS

■ C14 ■ C12 ■ C21 ■ HETEROSCEDASTICITY ■ ROBUST REGRESSION

INTRODUCTION

Homoscedasticity is known to be one of essential assumptions of linear regression. It is important not only for the classical least squares estimator, but also for any other (for example robust) estimator of regression parameters. The paper starts by defining the phenomenon of heteroscedasticity, which is the violation of homoscedasticity, and presents its negative consequences. Tests of heteroscedasticity are presented in for the least squares estimator, namely the tests of the Goldfeld-Quandt and Breusch-Pagan test. The new result is the asymptotic version of these tests derived for the least weighted squares, regression quantiles and trimmed least squares estimators. The solution of estimating parameters in the heteroscedastic model is called heteroscedastic regression, which is described again for various regression estimators.

Linear regression

In the whole paper we consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i\,,\ i = 1, 2, \ldots, n. \qquad (1)$$

The variance of the disturbances σ^2 is known to be a nuisance parameter. The homoscedasticity assumption

$$var\ e_i = \sigma^2\,,\ i = 1, \ldots, n \qquad (2)$$

is called homoscedasticity, while its violation is denoted as heteroscedasticity.

There can be severe negative consequences of heteroscedasticity, especially if the equality of variances of the disturbances is violated heavily. Regression parameters β cannot be estimated efficiently. Denoting the least squares estimator of β by b, the classical estimator of var b is biased. This disqualifies using classical hypothesis tests and confidence intervals for β as well as the value of the coefficient of determination R^2. Diagnostic tools checking the assumption of equality of variances of the disturbances can be based on residuals $u = (u_1, \ldots, u_n)^T$, where

$$u_i = Y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip},\quad i=1, \ldots, n, \qquad (3)$$

and ^{T} denotes a vector transposition.

This paper however has also the aim to show that σ^2 is a key parameter also in estimating β. It is crucial to estimate σ^2 reliably in order to obtain reliable tests of hypotheses about β and also its reliable confidence intervals. Also the robust estimators of linear regression parameters are sensitive to the assumption of homoscedasticity. In the linear regression it is known that σ^2 is a nuisance parameter in estimating the regression parameters β. This does not mean that σ^2 is not important or that its estimation stands aside during the inference of β. We bring arguments that σ^2 play a very important role in the statistical inference and influences the estimation procedures, which aim only at the regression parameters β. While the regression is based on the (very non-robust) sum of squares of residuals, the estimation of σ^2 is based exactly on the same sum of squares. This connects the problem of non-robustness of estimating β and σ^2.

Heteroscedasticity for least squares

We describe the classical Goldfeld-Quandt test and Breusch-Pagan test for the least squares regression. Each of these tests is designed for a different alternative hypothesis. More details on standard heteroscedasticity tests can be found in econometric references (Greene, 2002) or (Judge et al., 1985). Although originally proposed in econometric journals, they serve as basic diagnostic tools for a general statistical (not only econometric) context.

Goldfeld-Quandt test (Goldfeld and Quandt, 1965) is easy to be computed and interpreted. It tests the null hypothesis

$$H_0 : var\ e_i = \sigma^2\,,\ i = 1, \ldots, n, \qquad (4)$$

against the alternative hypothesis

$$H_1 : var\ e = \sigma^2\ diag\{k_1, \ldots, k_n\},\ i = 1, \ldots, n, \qquad (5)$$

which models heteroscedasticity in a particular way. The constants $k_1, \ldots, k_n$ must be selected by the statistician already before the computation. In fact the test does not depend on these values, but its power depends on them. The alternative hypothesis expresses that the variance of the disturbances $e_1, \ldots, e_n$ depends on some variable (or a combination of variables) in a monotone way. Typically one of the regressors in the linear regression model or fitted values of the response are selected to explain the variability of the disturbances in this way. The test is based of dividing the data to three groups according the values of the constants $k_1, \ldots, k_n$. Let SSE_1 denote the residual sum of squares in the first group of the data and let SSE_3 denote the residual sum of squares computed in the third group. Let r_1 denote the number of observations in the first group, r_3 in the third group and p is the number of regression

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

parameters in the linear regression model. Under homoscedasticity the test statistic

$$F = \frac{SSE_3}{SSE_1}\frac{r_1 - p}{r_3 - p} \quad (6)$$

follows Fisher's *F*-distribution with r_3 - p and r_1 - p degrees of freedom.

Breusch-Pagan test (Breusch and Pagan, 1979) requires to specify the alternative hypothesis of heteroscedasticity in the form

$$\text{var } e_i = \alpha_0 + \alpha_0 Z_{1i} + ... + \alpha_K Z_{Ki}, \; i = 1,...,n \quad (7)$$

for some variables

$$Z_1 = (Z_{11},...Z_{1n})^T, ..., \; Z_k = (Z_{K1},...,Z_{kn})^T. \quad (8)$$

Often one or more regressors in the original linear regression model are selected as these auxiliary variables. The null hypothesis corresponds to

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_K = 0, \quad (9)$$

which is tested against a general alternative hypothesis that the null hypothesis is not true. Breusch and Pagan (1979) derived the test statistic in the form of the Rao score test, which is one of general asymptotic tests based on the likelihood function, in our case under the presence of nuisance parameters. This tests assumes a normal distribution of the disturbances *e*.

White (1980) proposed a general test which is known as White test. The test exploits White's proposal of an estimator of the variance matrix var e, which is consistent also under heteroscedasticity. The test is based on comparing two estimators of the variance matrix, where the classical estimator is consistent only under homoscedasticity, while the White's estimator is consistent also under the alternative hypothesis. Therefore large values of the test statistic speak in favour of the alternative hypothesis. However the White test is a special case of Breusch-Pagan test. Here the particular choice of auxiliary variables $Z_1,...,Z_K$ is performed to contain squares of all regressors in the original model and also products of pairs of regressors in the form X_iX_j for $i \neq j$.

The least squares estimator is known to be too vulnerable with respect to violation of the assumption of the normal distribution of the disturbances e. Therefore robust statistical methods are studied intensively in the literature (see Jurečková and Sen, 1996), which represent a diagnostic tool for the least squares estimator or they can be used as an independent tool for the statistical modeling. One of efficient estimator is the least weighted squares proposed by Víšek (2001), which will now be presented.

Heteroscedasticity for least weighted squares

We recall the definition of the least weighted squares (LWS) regression estimator and describe asymptotic heteroscedasticity tests, which can be used as diagnostic tools for the LWS regression. The tests are based on the test statistics of the Goldfeld-Quandt and Breusch-Pagan test computed for residuals of the least weighted squares.

The least weighted squares (LWS) regression is a robust regression method with a high breakdown point proposed by Víšek (2001). There must be nonnegative weights w_1, w_2, ..., w_n specified before the computation of the estimator. While the classical weighted regression assigns a fixed and known weight to each observation, in the context of least weighted squares only the magnitudes of the weights are known a priori. These are assigned to the data after a permutation, which is determined automatically only during the computation based on the residuals. It is reasonable to choose such weights so that the sequence w_1, w_2, ..., w_n is decreasing (non-increasing), so that the most reliable observations obtain the largest weights, while outliers with large values of the residuals get small (or zero) weights.

Let us denote the i-th order value among the squared residuals for a particular value of the estimate b of the parameter β by $u_i^2(b)$. The least weighted squares estimator b_{LWS} for the model (1) is defined as

$$b_{LWS} = \text{argmin} \sum_{i=1}^{h} w_i u_{(i)}^2(b). \quad (10)$$

Kalina (2007) proposed an approximative algorithm for the intensive computation of the LWS estimator and described diagnostic tests for the estimator, which are equivalent with those computed for the least squares regression. A special case with weights equal to either 1 or 0 is the popular least trimmed squares (LTS) estimator, which has excellent properties in outlier detection (see Hekimoglu et al., 2009).

The least weighted squares estimator has interesting applications, which follow from its robustness and at the same time efficiency for normal data. Theoretical properties including the breakdown point of the estimator are studied by Víšek (2001). It is especially suitable to use the LWS estimator rather than other robust regression estimators, because diagnostic tools (such as tests of heteroscedasticity and autocorrelation of the errors e) can be computed directly using the weighted residuals and again are not affected by outliers. Another advantage of the estimator is that no detection of outliers is actually needed to compute it, because outlying data are downweighted automatically. Víšek (2010) conjectures that the LWS estimator is a reasonable compromise between the least squares and least trimmed squares, namely the estimator combines the efficiency of the least squares with the robustness of the least trimmed squares.

Kalina (2009) proposed the asymptotic Goldfeld-Quandt test and the asymptotic Breusch-Pagan test for the least weighted squares estimator. Víšek (2010) derives the White's estimator of var e for the LWS regression, which is based on the LWS estimation and is consistent under heteroscedasticity. This allows to define directly a test statistic of White (1980), which is tailor-made for the context of the LWS regression. Now we use these existing results and the ideas of proofs to derive asymptotic heteroscedasticity tests for regression quantiles.

*The*orem 1. Let the test statistic F of the Goldfeld-Quandt test be computed using residuals of the LWS regression estimator with a parameter α. Then F has asymptotically Fisher's F-distribution with r_3 - p and r_1 - p degrees of freedom under the null hypothesis of homoscedasticity and assuming normal distribution of disturbances in the linear regression model.

Theorem 2. Let the LWS estimator be computed in the linear regression model. Let the test statistic of Breusch-Pagan test χ^2 be computed as one half of regression sum of squares in the model

$$\frac{u_i^2}{s^2} = \alpha_0 + \alpha_0 Z_{1i} + ... + \alpha_K Z_{Ki} + v_i \qquad i = 1,...,n, \quad (11)$$

where $u = (u_1,...,u_n)^T$ is the vector of residuals of the regression quantile estimator and s^2 is the estimator of σ^2. Then the test statistic χ^2 is asymptotically χ^2_K distributed assuming the null hypothesis of homoscedasticity and

normal distribution of disturbances in the linear regression model.

The proof of the theorems follows from Kalina (2009) and the asymptotic representation of the LWS estimator given by Víšek (2001). Analogous steps were used by Kalina (2007) in the study of the asymptotic behavior of the Durbin-Watson test statistic computed with the residuals of the LWS estimator.

Approximation to the exact heteroscedasticity tests for the least weighted squares

We present a computational method, which allows to obtain the approximative critical value of approximative p-value of the Goldfeld-Quandt test for the LWS residuals. We use the notation of above and we denote by m_{ij} the elements of the matrix $M = I - X (X^TX)^{-1} X^T$,where I is a unit matrix of size *nxn*. It holds for the residuals of the least squares estimator that

$$SSE_1 = \sum_{i=1}^{r1}(\sum_{j=1}^{n} m_{ij}e_j)^2 \quad \text{and} \quad SSE_3 = \sum_{i=r_1+r_2+1}^{n} (\sum_{j=1}^{n} m_{ij}e_j)^2. \qquad (12)$$

This evaluation will be now used for the test for the LWS estimator.

Thanks to the asymptotic behavior of the residuals of the LWS regression, the test statistic of the Goldfeld-Quandt test (6) computed from the residuals of the LWS regression converges in probability to the test statistic (6) computed from the least squares residuals. Therefore the *p*-value can be obtained by a numerical simulation, which generates random variables $E_1,\ldots,E_n$ following the normal distribution *N*(0,1). They are plugged into (12) replacing the uknown errors, which consequently allows to compute (6).

We consider the Goldfeld-Quandt test of homoscedasticity against the (one-sided) alternative that the third part of the data (according to (5)) has a larger variability than the first part. The exact p-value of the test can be approximated by the empirical probability

$$P[F \le \frac{SSE_3}{SSE_1}\frac{r_1 - p}{r_3 - p}], \qquad (13)$$

where *F* is the test statistic (6) computed from the residuals of the LWS estimator and SSE_1 and SSE_3 are averages of sums of squares obtained with the randomly generated samples $E_1,\ldots,E_n$ from the N(0,1) distribution.

Because of the scale-invariance of the test statistic (6) and the approach in (13), the unit variance of the variables $E_1,\ldots,E_n$ is valid without loss of generality. In the same spirit the exact computation of the Breusch-Pagan test for the LWS estimator can be approximated.

Heteroscedasticity for regression quantiles

Regression quantiles represent a natural generalization of sample quantiles to the linear regression model. Their theory is studied by Koenker (2005) and their asymptotic representation was derived by Jurečková and Sen (1996). The estimator depends on a parameter α in the interval (0,1), which corresponds to dividing the disturbances to α·100% values below the regression quantile and the remaining (1-α)·100% values above the regression quantile. Here we describe asymptotic heteroscedasticity tests for regression quantiles, which are derived based on their asymptotic representation. The proof of the theorems follows from the asymptotic considerations of Kalina (2009).

Theorem 3. Let the test statistic F of the Goldfeld-Quandt test be computed using residuals of the quantile regression estimator with a parameter α. Then F has asymptotically Fisher's F-distribution with r_3 - p and r_1 - p degrees of freedom under the null hypothesis of homoscedasticity and assuming normal distribution of disturbances in the linear regression model.

Theorem 4. Let the regression quantile estimator with parameter α be computed in the linear regression model.

Let the test statistic of Breusch-Pagan test χ^2 be computed as one half of regression sum of squares in the model

$$\frac{u_i^2}{s^2} = \alpha_0 + \alpha_1 Z_{1i} + \ldots + \alpha_K Z_{Ki} + v_i, \qquad i=1,\ldots,n, \qquad (14)$$

where $u = (u^1,\ldots,u^n)^T$ is the vector of residuals of the regression quantile estimator and s^2 is the estimator of σ^2. Then the test statistic χ^2 is asymptotically χ^2_K distributed assuming the null hypothesis of homoscedasticity and normal distribution of disturbances in the linear regression model.

Heteroscedasticity for trimmed least squares

We recall the definition of the trimmed least squares (TLS) estimator, which is a robust estimator in the linear regression model based on regression quantiles. Its asymptotic representation derived by Jurečková and Sen (1996) allows to derive analogous asymptotic heteroscedasticity tests also for the TLS estimator.

The estimator depends on two parameters. These are fixed values α_1 and α_2 between 0 and 1. Let us denote by $\hat{\beta}(\alpha_1)$ and $\hat{\beta}(\alpha_2)$ the regression quantiles corresponding with the parameters α_1 and α_2 . Let us assume $0 < \alpha_1 < ½ < \alpha_2 < 1$. To define the TLS estimator let us introduce weights $w = (w_1,\ldots,w_n)^T$ where w_{1i} defined by 1, if the fitted value of the i-th observation by the quantile regression with parameter α_1 is smaller than Y_i and at the same time the fitted value of the i-th observation by the quantile regression with parameter α_2 is greater than Y_i. Let W denote the diagonal matrix with diagonal elements $w_1,\ldots,w_n$. The TLS estimator $b_{TLS}(\alpha_1, \alpha_2)$ is defined by

$$b_{LTS}(\alpha_1,\alpha_2) = (X^TWX)^{-1}X^TWY. \qquad (15)$$

Theorems 1 and 2 are valid also for the trimmed least squares estimator. The asymptotic representation requires the assumption that errors *e* come from a continuous distribution with a density function symmetric around 0.

Robust regression with heteroscedastic errors

If the null hypothesis of equality of variances in the model (1) is rejected by one of the previously described tests, we recommend to transform the model (1) to another model in order to suppress the negative consequences of heteroscedasticity. This is valid for any of the robust regression estimators. The estimation of regression parameters in the transformed model is called heteroscedastic regression. We discuss the procedure on the example of the LWS regression.

Assumptions or a prior knowledge on the form of heteroscedasticity should be incorporated within the process of removal heteroscedasticity. This is the case of the Goldfeld-Quandt test in the formula (5). Using the same notation we work with the model

$$\frac{Y_i}{\sqrt{k_i}} = \frac{\beta_1 X_{1i}}{\sqrt{k_i}} + \ldots + \frac{\beta_p X_{pi}}{\sqrt{k_i}} + \frac{e_i}{\sqrt{k_i}} L, \qquad (16)$$

One of typical examples is the choice $\sqrt{k_i} = x_{ji}$ for a certain j and for i=1,...,n, where the variance of the errors

is modeled to be directly proportional to the *j*-th regressor. Other examples include $\sqrt{k_i} = \sqrt{X_{ji}}$ or $\sqrt{k_i} = \hat{Y}_i = b_1X_{1i} + ... + b_pX_{pi}$, where i=1,...,n. In the model (16) we estimate the regression parameters by the least weighted squares method and heteroscedasiticy should be tested again. If the null hypothesis of homoscedasticity is not rejected this time, then the model (16) is considered to be preferable to the model (1). Therefore we consider only the results of the transformed model (16) including not only the point estimates of β, but also confidence intervals and hypothesis tests of β based on the asymptotic distribution of the LWS estimator, the value of the robust coefficient of determination and other statistics.

However sometimes the variability of the disturbances is modeled in a more complicated way, just like in formula (7) in the Breusch-Pagan test. Then we describe a possible procedure for the removal of heteroscedasticity in two stages. In the first stage the regression parameters in the model (1) are estimated by the least weighted squares method and squares of the LWS residuals u_i^2 are computed. Then the regression parameters in the auxiliary regression model

$$u_i^2 = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_K Z_{Ki} + v_i, \quad i=1,...,n, \qquad (17)$$

are estimated by the LWS estimator, where $v_1,\ldots,v_n$ are random disturbances. Thus we obtain estimates $\hat{\alpha}_0,\hat{\alpha}_1,...,\hat{\alpha}_K$ for regression parameters $\alpha_0, \alpha_1,\ldots, \alpha_K$. In the second stage the fitted values of u_i^2, which are computed as

$$\hat{u}_i^2 = \hat{\alpha}_0 + \hat{\alpha}_1 Z_{1i} + ... + \hat{\alpha}_K Z_{Ki}, \quad i=1,...,n, \qquad (18)$$

are used as the constants $k_1,\ldots,k_n$ for the transformed model (16), in which the estimators are computed using the LWS estimation procedure.

The White test (White, 1980) is often understood as a general method, which does not contain any recommendation about a possible removal of the heteroscedasticity (Greene, 2002). However since it is a special case of the Breusch-Pagan test, it also allows the heteroscedastic regression to be used in the same spirit.

Conclusion

This work studies the phenomenon of heteroscedasticity in robust regression. Assuming the standard linear regression model, the consequences of heteroscedasticity for robust regression are described and asymptotic heteroscedasticity tests for the least weighted squares, regression quantiles and trimmed least squares estimators are derived.

We also describe two possible ways of removing heteroscedasticity from the linear regression model. Both are based on a transformation of the original model and take into account such variables, which could possibly explain the variability of the disturbances. In other words this approach models the heteroscedasticity in a particular way. In practice such modeling is based on prior assumptions or knowledge. There exists no heteroscedasticity test optimal uniformly over all situations, but rather different tests have different properties. Therefore it is not possible to select the optimal heteroscedasticity test for a given data set. Another possibility is to use a robust regression estimator consistent also under the assumption of heteroscedastic disturbances (Víšek, 2010).

Acknowledgment

This research is supported by the grant 402/09/0557 (Robustification of selected econometric methods) of the Grant Agency of the Czech Republic.

References

Breusch, T.S., Pagan A.R. Simple test for heteroscedasticity and random coefficient variation. Econometrica 47 (5), 1287-1294, 1979.

Goldfeld, S.M., Quandt R.E. Some tests for homoscedasticity. Journal of the American Statistical Association 60 (310), 539-547, 1965.

Greene, W.H. Econometric analysis. Fifth edition. New York: Macmillan, 2002.

Hekimoglu, S., Erenoglu, R.C., Kalina J. Outlier detection by means of robust regression estimators for use in engineering science. Journal of Zhejiang University Science A 10 (6), 909-921, 2009.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H., Lee, T.C. The theory and practice of econometrics. New York: Wiley, 1985.

Jurečková, J., Sen, P.K. Robust statistical procedures: Asymptotics and interrelations. New York: Wiley, 1996.

Kalina, J. Heteroscedastic regression in robust econometrics. In Paganoni, A.M., Sangalli, L.M., Secchi, P., Vantini, S. (Eds.): Proceedings S.Co.2009, Sixth Conference Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction. Politecnico di Milano, Milano, 231-236, 2009.

Kalina, J. Asymptotic Durbin-Watson test for robust regression. Bulletin of the International Statistical Institute 62, 3406-3409, 2007.

Koenker, R. Quantile regression. Cambridge: Cambridge University Press, 2005.

Víšek, J.Á. Heteroscedasticity resistant robust covariance matrix estimator. Bulletin of the Czech Econometric Society 17 (27), 33-49, 2010.

Víšek, J.Á. Regression with high breakdown point. In Antoch, J., Dohnal, G. (Eds.): Proceedings of ROBUST 2000, Summer School of JČMF. Prague: JČMF and Czech Statistical Society, 324-356, 2001.

White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48 (4), 817-838, 1980.