DIAGNOSTICS FOR ROBUST REGRESSION: LINEAR VERSUS NONLINEAR MODEL

Jan Kalina

Abstract

Robust statistical methods represent important tools for estimating parameters in linear as well as nonlinear econometric models. In contrary to the least squares, they do not suffer from vulnerability to the presence of outlying measurements in the data. Nevertheless, they need to be accompanied by diagnostic tools for verifying their assumptions. In this paper, we propose the asymptotic Goldfeld-Quandt test for the regression median. It allows to formulate a natural procedure for models with heteroscedastic disturbances, which is again based on the regression median.

Further, we pay attention to nonlinear regression model. We focus on the nonlinear least weighted squares estimator, which is one of recently proposed robust estimators of parameters in a nonlinear regression. We study residuals of the estimator and use a numerical simulation to reveal that they can be severely heteroscedastic also for data generated from a model with homoscedastic disturbances. Thus, we give a warning that standard residuals of the robust nonlinear estimator may produce misleading results if used for the standard diagnostic tools.

Key words: robust estimation, outliers, diagnostic tools, nonlinear regression, residuals

JEL Code: C14, C12, C21

1 Robust regression

This paper is devoted to diagnostic tools for robust regression methods in the linear as well as nonlinear model. Robust regression methods represent important tools for estimating parameters in a variety of econometric models, proposed with the particular aim to be resistant against the presence of outlying measurements (outliers) in the data (Ronchetti & Trojani, 2001; Baldauf & Silva, 2012).

First, we consider the linear regression model

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

where $Y_1, \dots, Y_n$ are values of a continuous response variable and $e_1, \dots, e_n$ are random errors (disturbances). The task is to estimate the regression parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$. Gradually, the concept of breakdown point is becoming one of crucial measures of robustness of regression estimators (Huber & Ronchetti, 2009), while a high breakdown point can be interpreted as a high resistance (insensitivity) against outlying measurements in the data.

The regression median (also called L_1 estimator) is defined as the argument of minimum of the sum of absolute values of residuals in (1). It belongs to the class of M-estimators (Huber & Ronchetti, 2009) and represents one of the most popular estimators in the linear regression. However, the regression median does not have a high breakdown point.

The least weighted squares (LWS) represents one of robust estimators for the linear regression model with a high breakdown point (Víšek, 2011). The estimator has appealing properties like other statistical methods based on ranks of observations (Saleh et al., 2012). It has asymptotically a 100 % efficiency of the least squares under Gaussian errors. Its relative efficiency was declared to be high based on numerical simulations (Víšek, 2011), compared to maximum likelihood estimators under various distributional models. Extensions of the idea of implicit weights assigned to individual observations turn out to yield promising results also in other models, e.g. robust correlation coefficient (Kalina, 2012a) or classification analysis (Kalina, 2012b).

This paper has the following structure. Section 2 presents a heteroscedasticity tests for the regression median, namely the Goldfeld-Quandt, which is derived as an asymptotic test based on the asymptotic representation for the estimator. Section 3 recalls the nonlinear least weighted squares estimator as one of highly robust nonlinear regression estimators. Our example investigates residuals of this robust nonlinear estimator and brings arguments against using standard diagnostic tools for the nonlinear least weighted squares estimator. Finally, Section 4 concludes the paper.

2 Goldfeld-Quandt test for regression median

Goldfeld-Quandt is a standard test commonly used for the least squares estimator in the linear regression model. It considers the null hypothesis of homoscedastic disturbances, which reflects assumptions or a prior knowledge on the form of heteroscedasticity (Greene, 2002). We propose the Goldfeld-Quandt test for the regression median in Section 2.1. To remove the heteroscedasticity from (1), we propose a specific estimation procedure for the regression median in Section 2.2.

2.1 Asymptotic Goldfeld-Quandt test

The Goldfeld-Quandt test (Goldfeld & Quandt, 1965) considers the null hypothesis

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

against the alternative hypothesis that the variance of the disturbances depends on some variable (or variables) in a monotone way. Formally, the alternative hypothesis

$$H_1: var\ e_i = \sigma^2 k_i, \ \ i = 1, \dots, n, \tag{3}$$

models the heteroscedasticity by means of given constants $k_1, \dots, k_n$, which are (formally) fixed and known, although they do not influence the test itself.

The test is based od dividing the data to three groups according to $k_1, \dots, k_n$, while this is commonly performed according to values of one of the regressors in the linear regression model or according to fitted values of the response. Let SSE_1 denote the residual sum of squares in the first group of the data computed for the regression median 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 parameters in the linear regression model.

The asymptotic test may be based on the following theorem. Its proof is analogous to the theoretical reasoning of Kalina (2011), using the asymptotic representation for the regression median (Knight, 1998).

Theorem 1. Let the test statistic F of the Goldfeld-Quandt test be computed using residuals of the regression median. Then, the statistic

$$F = \frac{SSE_3}{SSE_1} \cdot \frac{r_1 - p}{r_3 - p} \tag{4}$$

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.

2.2 Heteroscedastic regression

Heteroscedasticity can be removed from the linear regression model by means of a modified model

$$\frac{Y_i}{\sqrt{k_i}} = \frac{\beta_1 X_{1i}}{\sqrt{k_i}} + \dots + \frac{\beta_p X_{pi}}{\sqrt{k_i}} + \frac{e_i}{\sqrt{k_i}}, \quad i = 1, \dots, n. \tag{5}$$

This approach requires to specify the constants $k_1, \dots, k_n$, which was not however necessary within the testing procedure of Section 2.1. Therefore, let us discuss the choice of suitable values $k_1, \dots, k_n$ now.

One of typical choices is to take $\sqrt{k_i} = X_{ji}$ for a certain j ($j = 1, \dots, p$) and $i = 1, \dots, n$, where the variance of the errors is modeled to be directly proportional to the *j*-th regressor. Other examples include

$$\sqrt{k_i} = \sqrt{X_{ji}} \quad \text{or} \quad \sqrt{k_i} = \hat{Y}_i = b_1 X_{1i} + \cdots + b_p X_{pi}, \tag{6}$$

where $i = 1, \dots, n$. In the model (5), the regression parameters are estimated by the regression median and heteroscedasiticy should be tested again.

If the null hypothesis of homoscedasticity is not rejected in this transformed model, then (5) is preferable to the model (1). It holds namely under H_1 that

$$var\ \frac{Y_i}{\sqrt{k_i}} = \frac{1}{k_i} var\ Y_i = \frac{1}{k_i} var\ e_i = \frac{1}{k_i}\ \sigma^2 k_i = \sigma^2, \quad i = 1, \dots, n. \tag{7}$$

The approach (5) ensures homoscedasticity under the assumption that exactly (11) holds, while the true form of heteroscedasticity may deviate from (3) and may reduce its benefits.

3 Robust estimation in nonlinear regression

This section recalls the nonlinear least weighted squares (NLWS) estimator and presents a numerical simulation motivated by the need for diagnostic tools for the estimator.

3.1 Nonlinear least weighted squares

Let us consider the nonlinear regression model

$$Y_i = f\left(\beta_1 X_{1i} + \cdots + \beta_p X_{pi}\right) + e_i, \quad i = 1, \dots, n, \tag{8}$$

where $Y = (Y_1, \dots, Y_n)^T$ is a continuous response,

$$X_i = (X_{1i}, \dots, X_{pi})^T, \quad i = 1, \dots, n, \tag{9}$$

is the vector of independent variables observed for the *i*-th measurement, f is a given nonlinear function and $(e_1, \dots, e_n)^T$ is the vector of random regression errors (disturbances). The model (8) can be expressed as

$$Y_i = f(X_i^T \beta) + e_i. \tag{10}$$

The aim of the analysis is to estimate the regression parameters $\beta = (\beta_1, \dots, \beta_p)^T$. Nonlinear regression models have found numerous econometric applications, e.g. in the analysis of cross-section data or financial time series (Chang et al., 2002).

The most common estimator of parameters in the nonlinear model (1) is the nolinear least squares (NLS) estimator defined as the argument of

$$\min \sum_{i=1}^{n} \left(Y_i - f(X_i^T b)\right)^2 \tag{11}$$

over all possible values of

$$b = (b_1, \dots, b_p)^T \in \mathbb{R}^p. \tag{12}$$

Using the NLS estimator should be accompanied by verifying its assumptions and its diagnostic tools are well known (Seber & Wild, 2003). Neverthless, the estimator suffers from a high vulnerability with respect to the presence of outliers in the data. While various robust regression estimators (Huber & Ronchetti, 2009) are available for the linear regression model, most of them do not allow to be extended to the nonlinear model (8).

The principle of the LWS estimation can be extended to the nonlinear model. Then, let us call the estimator as the nonlinear least weighted squares (NLWS). In order to give the formal definition of the NLWS estimator, we will use the notation $u_{(i)}(b)$ for the residual corresponding to the *i*-th observation for a given estimator (12) of β. We consider the residuals arranged in ascending order in the form

$$u_{(1)}^2(b) \leq \dots \leq u_{(n)}^2(b). \tag{13}$$

We define the least weighted squares estimator of the parameters in the model (20) as

$$argmin \sum_{i=1}^{n} w_i u_{(i)}^2(b), \tag{14}$$

where the argument of the minimum is computed over all possible values of $b = (b_1, \dots, b_p)^T$ and where $w_1, \dots, w_n$ are magnitudes of weights determined by the user.

The arguments for the high robustness of the LWS estimator with respect to outliers are valid also for the NLWS estimator thanks to the construction of the estimator, i.e. it can be explained as a consequence of the implicit weights assigned to individual observations An approximative algorithm for the computation of the NLWS estimator can be used in a direct analogy to the available LWS algorithm.

Just like the least squares estimator and other regression estimators (including the LWS), the NLWS estimator assumes i.a. uncorrelated and homoscedastic disturbances. Therefore, it is important to have tests for verifying these two assumptions. They will be based on residuals, which are defined as $(u_1, \dots, u_n)^T$, where

$$u_i = Y_i - f\left(X_i^T \hat{\beta}\right) = Y_i - \hat{Y}_i, \quad i = 1, \dots, n, \tag{15}$$

where $\hat{Y}_1, \dots, \hat{Y}_n$ are fitted values computed using the NLWS estimator. Possible weighting schemes include linearly decreasing weights or weights generated by a non-increasing function, such as a logistic curve.

3.2 Simulation

We randomly generate 70 observations following the Gompertz curve model

$$Y_i = \beta_1 + \beta_2 exp\{\beta_3 + e^{\beta_4 x_i}\} + e_i, \quad i = 1, \dots, n, \tag{16}$$

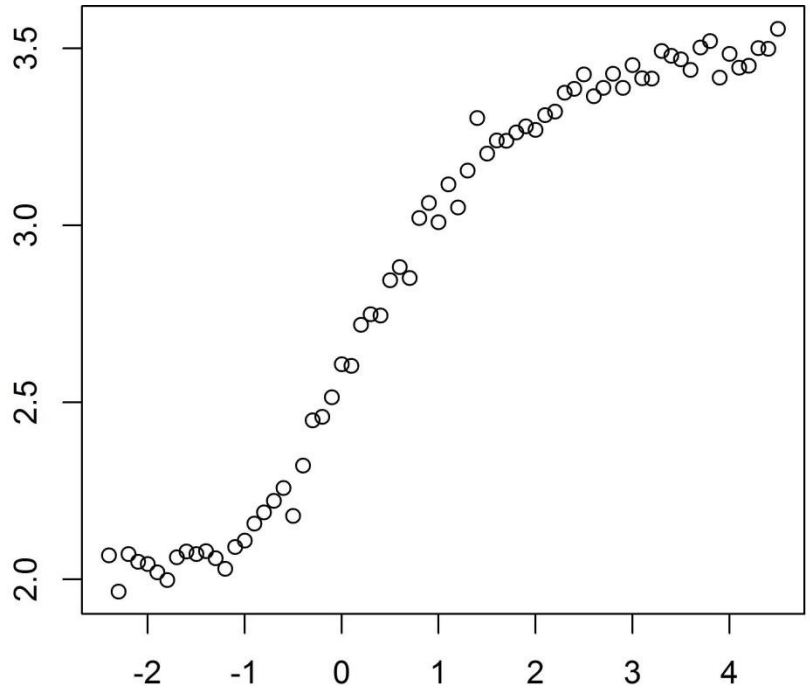
to illustrate the performance of the NLWS estimator with $\beta = (2, 1.5, -1, -1)^T$. The disturbances are generated as independent identically distributed random variables following a normal distribution $N(0, \sigma^2)$ with $\sigma = 0.05$. Figure 1 contains the plot of the response depending on the single regressor. The Gompertz growth curve is known as a model suitable e.g. for modeling of economic growth or as a consumption curve.

We used the NLWS estimation with linearly decreasing weights to estimate regression parameters of the model (16) based on the simulated data set. The estimated values, which are shown in Table 1, are close to the true values of the parameters. Figure 2 reveals however a controversial property of the residuals. The horizontal axis shows fitted values of the response, obtained as

$$\hat{Y}_i = b_1 + b_2 exp\{b_3 + e^{b_4 x_i}\}, \quad i = 1, \dots, n, \tag{17}$$

where $(b_1, b_2, b_3, b_4)^T$ denotes the NLWS estimate of $(\beta_1, \beta_2, \beta_3, \beta_4)^T$. The vertical axis shows the residuals of the NLWS estimate. We can see that residuals depend heavily on the fitted value of the response. In other words, their variability depends on the shape of the nonlinear function f. On the other hand, the random errors in (16) were constructed as homoscedastic, i.e. independent on the value of the response. Besides, we need to point out that the NLWS estimator is biased (just like the NLS), which makes the residuals not to be centered around zero. Particularly, the mean of the NLWS residuals is 0.015 in our example.

Fig. 1: Data set from Section 3.2. The data are randomly generated following the Gompertz growth model. The horizontal axis shows the regressor uniformly distributed over [-2.5, 4.5]. The vertical axis shows the response generated using (16).



Source: own computation

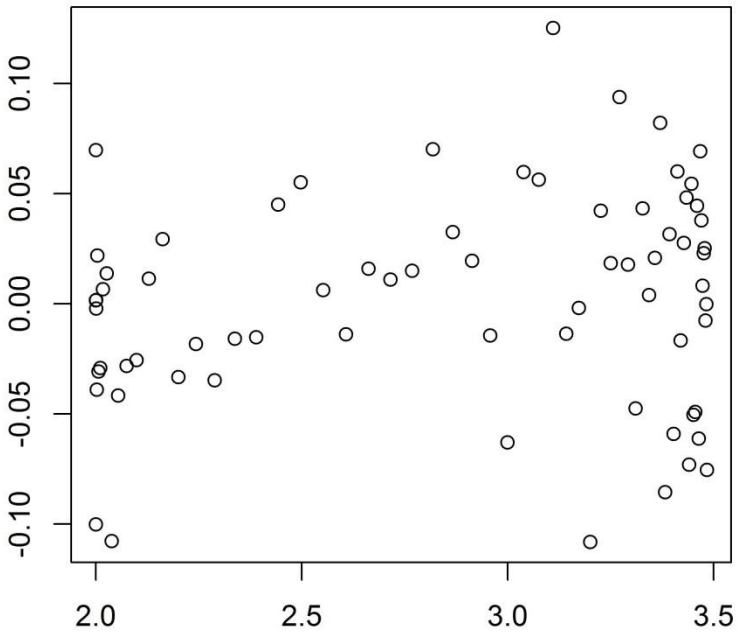
Tab. 1: True and estimated values of parameters for data in Figure 1. The estimate is obtained by the NLWS with linearly decreasing weights.

Parameter	True value	NLWS estimate	Standard error of the NLWS estimate
β_1	2	1.98	0.012
β_2	1.5	1.54	0.018
β_3	-1	-0.96	0.027
β_4	-1	-0.92	0.035

Source: own computation

We may conclude from Figure 2 that the (standard) residuals are not adequate for testing heteroscedasticity for the NLWS estimator. This is in accordance with the recommendations of Cook & Tsai (1985), who discouraged from using diagnostic tests for the nonlinear least squares.

Fig. 2: Illustration of the controversial behavior of residuals of the NLWS regression. The horizontal axis shows fitted values of the response (17) and the vertical axis the (standard) residuals (15).



Source: own computation

4 Conclusions

This paper is devoted to diagnostics for two robust estimators, namely the regression median for the linear regression model and for the nonlinear least weighted squares estimator.

In the linear regression, we derived an asymptotic Goldfeld-Quandt test of heteroscedasticity for the regression median estimator. It is valid asymptotically in the same form as it is routinely used for the least squares. While the test is exactly valid (i.e. for a small number of observations) for the least squares, it can be recommended for the regression median for a large number of observations. In addition, other diagnostic tests could be derived for the residuals of the regression median, including White test or tests of the Szroeter's class. Nevertheless, deriving the tests requires to assume the normal distribution of the disturbances, which limitates the new tools, because under normality without outliers the regression median loses its efficiency compared to the least squares.

In the nonlinear regression, however, the situation is much more complex. We illustrated the intricate properties of residuals of the nonlinear least weighted squares on simulated data. The results on the data reveal that the residuals are far from homoscedasticity, even if the assumption of homoscedastic disturbances in the regression model is fulfilled. Thus, we find

residuals to be unsuitable for making conclusions about the disturbances (random errors). While tests from the linear regression are no longer valid for the NLWS estimator, we do not recommend to use residuals even for a subjective diagnostics concerning the disturbances.

Important limitations of the robust nonlinear estimation include a non-robustness to the specification of the alternative hypothesis for the heteroscedasticity tests as well as specification of the nonlinear function f in the model (Baldauf & Silva, 2012).

On the whole, constructive results of this paper may allow the robust methods to become more popular in econometric applications. On the other hand, we must admit that available robust regression methods still suffer from serious shortcomings for several main reasons:

- They require various tuning constants with a difficult interpretation;
- Various robust methods yield rather different results;
- Computational intensity;
- Robustness only with respect to outliers but not to a misspecification of the model.

Acknowledgment

The research was supported by the Czech Science Foundation project No. 13-01930S and by the Neuron Fund for Support of Science.

References

1. Baldauf, M. & Silva J.M.C.S. (2012): On the use of robust regression in econometrics. *Economic Letters*, 114 (1), 124-127.
2. Breusch, T.S. & Pagan, A.R. (1979): Simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47 (5), 1287-1294.
3. Chang, Y., Park, J.Y. & Phillips, C.B. (2001): Nonlinear econometric models with cointegrated and deterministically trending regressors. *Econometric Journal*, 4 (1), 1-36.
4. Cook, R.D. & Tsai C.-L. (1985): Residuals in nonlinear regression. *Biometrika*, 72, 23-29.
5. Goldfeld, S.M. & Quandt, R.E. (1965): Some tests for homoscedasticity. *Journal of the American Statistical Association*, 60 (310), 539-547.
6. Greene, W.H. (2002): *Econometric Analysis.* 5th edn. New York: Macmillan.
7. Huber, P.J. & Ronchetti E.M. (2009): *Robust Statistics*. 2nd edn. New York: Wiley.
8. Kalina, J. (2011): Some diagnostic tools in robust econometrics. *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica*, 50 (2), 55-67.

9. Kalina, J. (2012a): Implicitly weighted methods in robust image analysis. *Journal of Mathematical Imaging and Vision*, 44 (3), 449-462.
10. Kalina, J. (2012b): Highly robust statistical methods in medical image analysis. *Biocybernetics and Biomedical Engineering*, 32 (2), 3-16.
11. Knight, K. (1998): Limiting distributions for L1 regression estimators under general conditions. *Annals of Statistics*, 26, 755-770.
12. Ronchetti, E. & Trojani, F. (2001): Robust inference with GMM estimators. *Journal of Econometrics*, 101, 37-69.
13. Saleh, A.K.M.E., Picek, J., & Kalina, J. (2012): R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika, 75,* 311-328.
14. Seber, G.A.F., & Wild C.J. (2003): *Nonlinear Regression.* Wiley, New York.
15. Víšek, J. Á. (2011): Consistency of the least weighted squares under heteroscedasticity. *Kybernetika, 47,* 179-206.

Contact

Jan Kalina
Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
kalina@cs.cas.cz