RT-MAE 2012-07

ANALYSIS OF RESIDUALS IN QUANTILE REGRESSION: AN APPLICATION TO INCOME DATA IN BRAZIL

by

Bruno R. Santos
and
Silvia N. Elian

Palavras-Chave: Analysis of Residuals; Quantile Residuals; Quantile Regression; Income; Equivariance Property.

Classificação AMS: 62J05.

- Setembro de 2012 -

Analysis of Residuals in Quantile Regression: An Application to Income Data in Brazil

Bruno R. Santos and Silvia N. Elian

Departamento de Estatística
Instituto de Matemática e Estatística,
Universidade de São Paulo
CP 66281 - São Paulo, Brasil
E-mails: bramos@ime.usp.br
selian@ime.usp.br

Abstract

Analysis of residuals is a very important analysis usually performed in the classical regression diagnostics framework. In this paper, we propose a similar kind of analysis, but in quantile regression models. We make use of quantile residuals defined by Dunn and Smyth (1996) to verify the assumption of asymmetric Laplace distribution (Yu and Zhang, 2005) to the errors in a quantile regression model. To illustrate the method we used data from the National Household Sample Survey, performed in Brazil. We were able to visualize a better approximation of the asymmetric Laplace assumption only in the log-linear model fitted to describe income as a function of other variables.

Keywords: Analysis of Residuals; Quantile Residuals; Quantile Regression; Income; Equivariance Property.

1 Introduction

Analysis of residuals has been a common way of verifying some model assumptions in the classic regression analysis, usually the normal distribution assumption for the errors. In this article, we propose a similar method using the asymmetric Laplace distribution and a definition of quantile residuals to analyze residuals of a quantile regression model fit.

In Section 2, we give a brief summary of the asymmetric Laplace distribution used in this article, then in Section 3 we provide the main concepts of quantile regression models and in Section 4 we define the quantile residuals used in the analysis of residuals. We finish this paper with an application of this method in Section 5 and in Section 6, we give our last remarks on the subject.

2 Asymmetric Laplace Distribution

We shall consider throughout this text the definition of the asymmetric Laplace distribution (ALD) of Yu and Zhang (2005). In this way, we must consider that if $Y \sim \text{ALD}(\mu, \sigma, \tau)$, then its distribution function is given by

$$F(y;\mu,\sigma,\tau) = \begin{cases} \tau \exp\left(\dfrac{1-\tau}{\sigma}(y-\mu)\right), & \text{if } y \leqslant \mu, \\ 1-(1-\tau)\exp\left(-\dfrac{\tau}{\sigma}(y-\mu)\right), & \text{if } y > \mu. \end{cases}$$

where $0 < \tau < 1$ is the skew parameter, $\sigma > 0$ is the scale parameter and $-\infty < \mu < \infty$ is the location parameter.

We will see in the next section the relation between this distribution and the quantile regression framework.

3 Quantile Regression

Since its definition by Koenker and Bassett (1978), quantile regression has been used in several kinds of studies (see, e.g., Yu et al., 2003) as an alternative to the least squares method. First, we should assume the following linear model to describe the relation between $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$

$$Y = \beta_0(\tau) + \beta_1(\tau)x_1 + \cdots + \beta_p(\tau)x_p + \epsilon,$$

where the τth quantile of ϵ is zero.

Using the asymmetric Laplace distribution, we have that if $\epsilon \sim \text{ALD}(0, \sigma, \tau)$, so its τth quantile is equal to zero, in agreement with the assumption of the model.

Beyond that, it is known that the quantile regression estimator, $\hat{\beta}(\tau)$, for the parameters of the model above is obtained by finding

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$.

Nevertheless, considering the asymmetric Laplace distribution for the errors, we have that $\hat{\mu}_i = x_i'\hat{\beta}(\tau)$ is the maximum likelihood estimator (MLE) for the conditional location parameter, the τth conditional quantile of Y,

$Q_Y(\tau|x)$, since $\hat{\beta}(\tau)$ is the MLE for $\beta(\tau)$. Therefore, we have a consistent estimator for $Q_Y(\tau|x)$.

4 Quantile Residuals

Following the paper of Dunn and Smyth (1996), we will use the quantile residuals defined as

$$r_{q,i} = \Phi^{-1}\left\{F(y_i, \hat{\mu}_i, \hat{\sigma}, \tau)\right\},$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal and F is as defined in Section 2. According to the authors, apart from sampling variability in the estimators of μ and σ, the $r_{q,i}$ are exactly normal, implying that if μ and σ are consistently estimated, then the distribution of $r_{q,i}$ converges to the standard normal distribution. It is important to notice that the above definition is a special case of Cox and Snell (1968) "crude" residuals.

We argued in the last section that μ is consistently estimated by its MLE, with the quantile regression estimator. Using the same idea, it is easy to prove that the MLE for σ is

$$\hat{\sigma} = n^{-1}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{y}_i).$$

Considering these results, we can use the quantile residuals of a quantile regression model to determine if the assumption of the asymmetric Laplace

distribution is confirmed after the model fit. For this, we can analyze graphics such as a histogram or a QQ-plot of the quantile residuals.

5 Application

In this part of the article, we will consider data from the National Household Sample Survey, which took place in Brazil, in 2009, to model income as a function of other variables. This type of model is often studied with quantile regression (see Buchinsky, 1994 and Yu et al., 2005).

This survey is done every year by the Brazilian Institute of Geography and Statistics (IBGE). We limited our sample to people who earned at least one third of the minimum wage in 2009, who were between 18 and 80 years old and who worked at least 40 hours/week during the period of the survey. With this filter, we selected 122.727 people.

Our response variable of interest, Y_i, is the real gross monthly income. For the independent variables, we will consider gender, age, age squared, education and a dummy variable indicating whether the person is single or not. We will use the following linear model, and its respective log-linear formulation,

$$y_i = \beta_0(\tau) + \beta_1(\tau)\mathrm{G}_i + \beta_2(\tau)\mathrm{A}_i + \beta_3(\tau)\mathrm{A}_i^2 + \beta_4(\tau)\mathrm{E}_i + \beta_5(\tau)\mathrm{S}_i + u_i,$$

where G_i is equal to 1 for men and 0 for women, A_i is the age in years, E_i is the years of schooling, and S_i is equal to 1 for single individuals and 0

otherwise. In both cases, we will assume $u_i \sim \text{ALD}(0, \sigma, \tau)$. For the sake of brevity, we will analyze the results only for $\tau = 0.5$, regarding the conditional median of Y.

We refer to Table 1 for the estimates of the fitted models, using the linear formulation and also the log-linear formulation.

Table 1: Estimates for the fitted models, $\tau = 0.5$

Variables	Linear Model	Log-linear Model
(Intercept)	-688.44	4.69
Gender	218.00	0.28
Age	27.81	0.04
Age^2	-0.16	-0.00
Education	68.96	0.09
Single	-102.26	-0.14

Now, we must verify if the assumption of the asymmetric Laplace distribution of the errors is reasonable in each formulation. It is expected, and concluded with the analysis of Figure 1, that the errors, and consequently the response variable, are not distributed according to a symmetric Laplace distribution, which is the case when τ is equal to 0.5. It is well known that income has an asymmetric distribution, with greater concentration in lower incomes. In agreement with this idea, the quantile residuals demonstrate, in both graphics, a bigger concentration in lower values.

On the other hand, with the log transformation of the response variable, as we can visualize with the quantile residuals in Figure 2, we believe that a better approximation is achieved. Using the log-linear model, the quantile residuals show a symmetric behavior, as we can notice in the histogram. In the qq-plot, we can see a bigger difference between the theoretical and

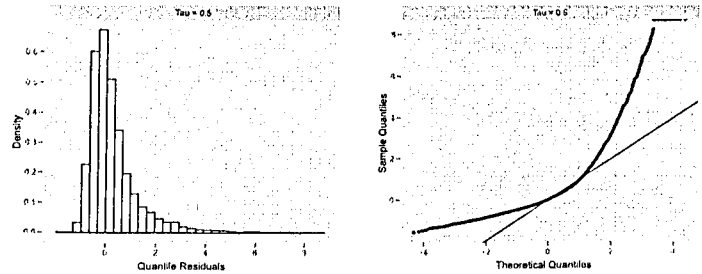

Figure 1: Histogram and QQ-plot for the quantile residuals in the linear model.

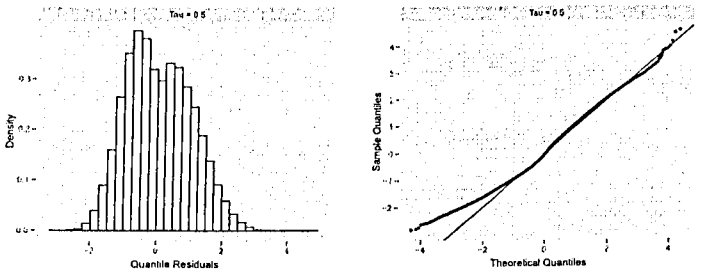

Figure 2: Histogram and QQ-plot for the quantile residuals in the log-linear model.

the sample quantiles in the left tail than it was expected. Despite this, we believe that the asymmetric Laplace distribution is a good approximation for the error distribution in this example.

Another consideration of these models, that we could address in this example, is the equivariance property that allows us to infer about the conditional median of income with the log-linear model (Koenker, 2005). In this way, we could give the following approach about inference with quantile regression considering this analysis of residuals. One could use monotone transformations until finding the proper asymmetric Laplace distribution for

the error distribution, when this is not obtained in the quantile regression model with the response variable not transformed. This is possible remembering that if $Q_y(\tau|x)$ is the τth conditional quantile of Y given X and we use $h(.)$, a monotone nondecreasing function, in Y, then by the equivariance property we have that

$$Q_{h(Y)}(\tau|x) = h(Q_Y(\tau|x)).$$

6 Concluding Remarks

Koenker and Machado (1999) and He and Zhu (2003) discuss evaluation methods of goodness of fit and lack-of-fit, respectively, in quantile regression models. We believe that our approach is connected with this kind of analysis, as we verify if τth quantile of the errors is equal to zero using the asymmetric Laplace distribution, which is a very important assumption of these models. We found that, in the Brazilian example, this distribution was reasonable to explain the error in the log-linear model with income as function of other variables. We used the equivariance property of quantile regression models to note possible inferences about the conditional median using the log-linear model.

The same way as the least squares estimator is closely related to the MLE of regression models with normal distribution for the errors, we described a similar connection between the estimator of quantile regression and MLE with errors distributed according to an asymmetric Laplace distribution. For this reason, we think that the graph analysis proposed in this article could

have the same importance as the residuals analysis usually performed in the classical regression analysis.

Although this was not dealt here, these residuals could also be used to verify other assumptions of quantile regression models, such as linearity and homocedasticity, for example, when plotted together with the fitted values of the model.

Acknowledgments

Special thanks to CAPES for financial support of the first author, during which he was able to complete the Master's Program in Statistics at the University of Sao Paulo, when he studied the main subjects of this article.

References

Buchinsky, M. (1994). Changes in US Wage Structure 1963-87: An Application of Quantile Regression. *Econometrica*, **62**, 405–458.

Cox, D. and Snell, E. (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society, Series B*, **30**, 248–275.

Dunn, P. and Smyth, G. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

He, X. and Zhu, L. (2003). A Lack-of-Fit Test for Quantile Regression. *Journal of the American Statistical Association*, **98**, 1013–1022.

Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33–50.

Koenker, R. and Machado, J. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*, **94**, 1296–1310.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile Regression: Application and Current Research Areas. *The Statistician*, **52**, 331–350.

Yu, K., van Kerm, P. and Zhang, J. (2005). Bayesian Quantile Regression: An Application to the Wage Distribution in 1990s Britain. *Sankhyā - The Indian Journal of Statistics*, **67**, 359–377.

Yu, K. and Zhang, J. (2005). A Three-Parameter Asymmetric Laplace Distribution and Its Extension. *Communications in Statistics - Theory and Methods*, **34**, 1867–1879.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2012-01 - BUENO, V.C. Asymptotic reliability of a coherent system in its signature representation. 09p. (RT-2012-01)

2012-02 - BUENO, V.C., RAMOS, J.A. Signature Processes. 16p. (RT-MAE-2012-02)

2012-03 - BUENO, V.C. Marked Point Signature Procesess. 10p. (RT-MAE-2012-03)

2012-04 - MIRANDA, J.C.S.; FRANCO FILHO, A.P. A class of dilation integral equations. 06p. (RT-MAE-2012-04)

2012-05 - BUENO, V.C. Marked point signature processes and redundancy allocation for a coherent system. 18P. (RT-MAE-2012-05)

2012-06 - KOLEV, N., PINTO, J. Extended Marshall-Olkin Model and Applications 13p. (RT-MAE-2012-06)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05314-970 - São Paulo, Brasil