American Journal of Applied Mathematics and Statistics, 2015, Vol. 3, No. 2, 86-88
Available online at http://pubs.sciepub.com/ajams/3/2/8

DOI:10.12691/ajams-3-2-8

Robustness of Quantile Regression to Outliers

Onyedikachi O. John*

Department of Physical Sciences, Rhema University, Aba
*Corresponding author: johnkady@yahoo.com

Received February 27, 2015; Revised March 29, 2015; Accepted April 22, 2015

Abstract Sensitivity of an estimator to departures from its distributional assumptions is a very important issue that is worth considering. The influence function, which describes the effect of an infinitesimal contamination at point, y, on the estimator we are seeking, standardized by the mass, ε, of the contamination, is bounded for the median. This property of the median is enjoyed by the other quantile points. Quantile regression inherits this robustness property since the minimized objective functions in the case of sample quantile and in the case of quantile regression are the same. This robustness is investigated by analyzing the quarterly implicit price deflator using quantile regression. The coefficients for the median and other quantiles remain unchanged even when outlier is added to the data.

Keywords: *breakdown points, infinitesimal contamination, influence function, quantile regression, robustness, outliers*

Cite This Article: Onyedikachi O. John, "Robustness of Quantile Regression to Outliers." *American Journal of Applied Mathematics and Statistics*, vol. 3, no. 2 (2015): 86-88. doi: 10.12691/ajams-3-2-8.

1. Introduction

An outlier is an extreme observation. Typically, points further than, say, three or four standard deviations from the mean are considered as 'outliers'. Outliers occur frequently in real data, and can cause one to misinterpret patterns in plots, and may also indicate that model fails to capture the important characteristics of the data. Deleting outliers from the regression model can sometimes give completely different results. Thus the sensitivity of an estimator to departures from its distributional assumptions is a very important issue. The sample mean being a superior estimate of the expectation under normality of the error distribution can be adversely affected even by a single observation if it is sufficiently far from the rest of the data points, Cizek, [2]. On the other hand, the performance of the median can be superior in the presence of outlying observations; a point stressed by many authors including, remarkably, Kolmogorov, [9].

The modern view of this, strongly influenced by Tukey, [1], is framed by the sensitivity curve, or influence function of the estimators, and perhaps to a lesser degree by their finite sample breakdown points, [8]. The influence function, introduced by Hampel, [5], is a population analogue of Tukey's empirical sensitivity curve. The idea of contaminating a distribution with a small amount of additional data has a long history in statistics and the investigation of robust estimators.

The median has a bounded influence function, implying that the effect of an outlier on a sample median is bounded no matter how far the outlying observation is. This robustness of the median is of course overweighed by lower efficiency in some cases. Other quantiles enjoy similar properties. Quantile regression inherits this robustness property since the minimized objective functions in the case of sample quantiles and in the case of quantile regression are the same.

2. Influence Function

The influence function is the directional derivative of T(F) at F, and it measures the effect of a small perturbation in F on T(F), Essama-Nssah, [4]. Suppose T is a functional of F. Δy is the probability measure which assigns mass 1 to {y}. The influence function is then defined by

$$IF(y;T;F) = \lim_{s\to 0}\frac{T(s\Delta y+(1-s)F-T(F))}{s} \quad (1)$$

It describes the effect of an infinitesimal contamination at the point, y, on the estimator: in mixed distribution, εΔy + (1-ε)F, it is as if an observation is randomly sampled from distribution F with probability (1-ε) and from Δy with probability ε.

A. Influence Function for the Mean

$$T(F_e) = \int y dF_e = \varepsilon y + (1-\varepsilon)T(F) \quad (2)$$

Where $F_e = \varepsilon\Delta y + (1-\varepsilon)F$.

So the influence function

$$IF(y;T;F) = y - T(F). \quad (3)$$

This implies that, as y gets larger, its influence on the mean becomes larger.

B. Influence Function for Quantile Points

For the τ^{th} quantile points, the influence function,

$$IF(y;T;F)=\begin{cases}\dfrac{\tau}{f\left(F^{-1}(\tau)\right)}; & y>F^{-1}(\tau)\\ \dfrac{(\tau-1)}{f\left(F^{-1}(\tau)\right)}; & y\le F^{-1}(\tau)\end{cases} \quad (4)$$

This implies that the influence of contamination at y on the median, and generally on the other quantile points is bounded provided that the sparsity at τ is finite.

C. Influence Function for Quantile Regression

Following the idea expressed in Koenker and Portnoy, [8], the influence function can be extended to regression. Let F represent the joint distribution of the pairs (x,y). Writing dF in the conditional form

$$dF = dG(x)\, f\left(y|x\right)dy \quad (5)$$

and again assuming that f is continuous and strictly positive when needed we have,

$$IF\left((y,x),\hat{\beta}_{F(\tau)},F\right)=Q^{-1}x\,\mathrm{sgn}\left(y-x'\hat{\beta}_F(\tau)\right) \quad (6)$$

where

$$Q=\int xx'f\left(x'\hat{\beta}_F(\tau)\right)dG(x).$$

Again we see that the estimator has bounded influence in y since y appears only clothed in the protective sgn (·) function. This can also be illustrated in the following theorem, see [6].

Theorem 1*: Let* D be an $n\times n$ diagonal matrix with nonnegative elements and $\hat{\varepsilon}=y-X\hat{\beta}(\tau;y,X)$ be the residual vector of the τ^{th} quantile regression fit with $\hat{\beta}(\tau;y,X)$ the τ^{th} quantile regression estimate of the model $y_i=x_i'\beta+\varepsilon_i$, y the vector of observed dependent variable and X the design matrix. Then

$$\hat{\beta}(\tau;y,X)=\hat{\beta}\left(\tau;X\hat{\beta}(\tau;y,X)+D\hat{\varepsilon},X\right). \quad (7)$$

The above theorem indicates that the quantile regression estimate is not affected by any change in the values of the dependent variable for some observations as long as the relative positions of the observation points to the fitted plane are maintained. Intuitively, the breakdown point of an estimator is the proportion of incorrect observations (arbitrarily large observations) an estimator can handle before giving an incorrect (arbitrarily large) result. The higher the breakdown point of an estimator, the more robust it is. The median has a breakdown point of 50%.

3. Analysis and Discussion

To demonstrate the robustness of quantile regression to outlying observations, we consider data from Central Bank of Nigeria, [3], with the Quarterly Implicit Price Deflator as the dependent variable, and Agriculture, Industry, Building and Construction, Wholesale and Retail, Services, as independent variables. The data cover from 2000 to 2012. R package is used for this analysis, and the result is as follows:

Table 1 and Table 2 give the OLS and quantile regression results for the original data. X1: agriculture, x2: industry, x3: building and construction, x4: wholesale and retail, x5: services.

Table 1. Ordinary Least Squares Estimates

(intercept)	246.33071
X1	0.25348
X2	0.19621
X3	0.05155
X4	0.11198
X5	0.35762

Table 2. Quantile Regression Estimates

	tau=0.1	tau=0.3	tau=0.5
(Intercept)	168.88776	234.89627	256.10545
X1	0.29482	0.22849	0.21231
X2	0.18295	0.19922	0.20529
X3	0.04120	0.02064	0.03499
X4	0.15598	0.14307	0.13905
X5	0.31105	0.36521	0.35490

Table 3.Ordinary Least Squares Estimates

(intercept)	-520.6084
X1	2.5606
X2	-0.2934
X3	-0.2558
X4	0.3747
X5	-0.5697

Table 4. Quantile Regression Estimates

	tau=0.1	tau=0.3	tau=0.5
(Intercept)	168.88776	234.89627	256.10545
X1	0.29482	0.22849	0.21231
X2	0.18295	0.19922	0.20529
X3	0.04120	0.02064	0.03499
X4	0.15598	0.14307	0.13905
X5	0.31105	0.36521	0.35490

An outlier is added to the data on the implicit price deflator by multiplying an observation by 5, and the resulting data analyzed. Table 3 and Table 4 give the OLS and quantile regression result for this contaminated data.

Again an outlier is added to the original data on the implicit price deflator by multiplying another observation by 7, and the resulting data analyzed. Table 5 and Table 6 respectively give the OLS and quantile regression of the data.

It is clearly seen from the results that the estimates for the conditional mean changed drastically when an outlier is introduced to the original data, giving entirely different representation of the original data, with industry, wholesale and retail, and services showing negative relationship to the implicit price deflator. However, the quantile regression results remain the same even in the presence of the outlying observations.

Table 5. Ordinary Least Squares Estimates

(intercept)	515.9725
X1	0.2045
X2	-0.2255
X3	-2.3464
X4	3.0652
X5	0.4680

Table 6. Quantile Regression Estimates

	tau=0.1	tau=0.3	tau=0.5
(Intercept)	168.88776	234.89627	256.10545
X1	0.29482	0.22849	0.21231
X2	0.18295	0.19922	0.20529
X3	0.04120	0.02064	0.03499
X4	0.15598	0.14307	0.13905
X5	0.31105	0.36521	0.35490

4. Conclusion

The influence function is an indispensible tool, designed to measure the sensitivity of estimators to infinitesimal perturbation of the nominal model. Quantile regression, introduced by Koenker and Basset, [7], inherits its robustness property from median regression, and can produce good and reliable estimates even in the presence of extreme outliers.

References

[1] D. F. Andrew, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Roger, J. W. Tukey, "Robust Estimate of Location: survey and advances," Princeton, Princeton U. Press, 1972.

[2] P. Cizek, "Quantile Regression in XploRe Application Guide," ed. W. Hardle, Z. Hlavka, S. Kline, Berlin, MD Tech Springer, 2003.

[3] Central Bank of Nigeria, "2012 Statistical Bulletin: Domestic Product, Consumption and Prices," Stabull 004, 2013.

[4] B. Essama-Nssah and P. J. Lambert, "Influence Functions for Distributional Statistics," Society for the study of Econoic Inequality, ECINEQ Working Paper Series, 2011.

[5] F. Hampel, "The Influence of Curve and its Role in Robust Estimator," J. of the American Statistical Association, 1974, 69, p.383-393.

[6] R. Koenker, "Quantile Regression" New York, Cambgridge University Press, 2005,p.138-141

[7] R. Koenker and G. Basset, "Regression Quantiles" Econometrica, 1978, 46, p.33-50.

[8] R. Koenker and S. Portnoy, "Quantile Regression," University of Illinois, Urban Champaign, 1999. Available at http:/www.econ.uiuc.edu/roger/.

[9] A. N. Kolmogorov, "The method of the median in the Theory of Errors," Mat. Sb. Reprinted in selected works of A. N. Kolmogorov, vol II, A. N. Shirayev, (ed), Kluwer: Dordrecht, 1931.