A Cauchy Transformation approach to the Robustness of Quantile Regression Model to Outliers

Nwabueze, Joy C. [1],Onyegbuchulem, Besta O.[2], Nwakuya, Maureen T.[3] and Onyegbuchulem, Chialuka A.[4]

(1) Department of Statistics, Federal University of Agriculture Umudike
(2) Department of Maths/Statistics, Imo State Polytechnic Umuagwo.
(3) Department of Maths/Statistics, University of Port Harcourt, River State.
(4) Department of Mathematics.Alvan Ikoku Federal College of Education, Owerri.
Correspondence E-Mail: - bokey@imopoly.net

Abstract

The main advantage of quantile regression models had over OLS is their robustness to outliers. This is because quantile regression models are insensitive to outliers and skewed distributions. This very property of quantile regression model is the same with the sample quantile. This work was done to examine the robustness of quantile regression model to outliers. Data analysis was done usingreal life data set on fuel consumption (in miles per gallon), in highway driving as the response variable. Extreme values where inserted to create outliers in the response variable data set. Car weight, length, wheel base, width, Engine size and horse power are the explanatory variables used in the analysis with a sample size of 91. The standard Cauchy distribution was used to transform the quantile regression model. The results show that the graphs of the mean square errors clustered around the zero line in all the study quantiles, also the descriptive results show that the residual means is equal to the residual medians and equal to zero. The skewness of the residuals approximates to zero across all the study quantiles, while the kurtosis approximates to 3, both the residual standard deviations, mean square errors and root mean square errors approximate to zero across all the study quantiles. From the results of the analysis, it can be concluded that quantile regression model is insensitive to outliers.

Key words: Quantile Regression Model, Cauchit Quantile Regression Model, robustness to outliers and Mean Square Error

1 Introduction

A better alternative to conditional-mean modeling has the tendency of measuring the intercept of the median regression which obviously happened to be a key theorem about minimizing sum of the absolute deviation and a geometrical algorithm for constructing median regression and this was proposed by Ruder Josip Boskovic, a Jesuit Catholic Priest from Dubrovnik in 1790. Hence, the Conditional-median regression is type of quantile regression where the conditional 50th quantile is modeled as a function of dependent variable, other quantiles can obviously be used to determine the none central positions of a distribution, Koenker & Bassett (1978).
Binary response quantile regression model came as a result of the deficiencies inherent in both linear regression models as well the quantile regression model, such as the problem of modeling two alternatives in the response variable or ratio response variables (Manski, 1975; 1985). Some functions that can be used to form binary or ordinal quantile regression model include logit, probit, Negative log-log, Aranda-Ordaz, Complementary Log-log, Log-log, and Cauchit regression, (Bonat, Ribeiro and Zeviani, 2012).

Using Cauchy function to transform quantile regression is a predictive analysis, where the dependent variable is ordinal (statistically it is polytomous ordinal) and the independent variables are ordinal or continuous-level. Cauchy transformed quantile regression model is applied for two major reasons which include causal analysis and forecasting an effect. Obviously the major aim of Quantile Regression model was to correct some of the anomalies accompanied with Linear Regression such as stringent assumption of the linear regression model, difficulty in using the linear regression model on skewed distributed response data, and data heavily distributed with outliers. Therefore, Cauchy transformed quantile regression model is therefore proposed with the aim to manage outliers in the response variable at a reduced error term. This study is therefore aimed at examining the robustness or otherwise of the quantile regression model to outliers using Cauchy transformation.

2.0 Methodology

The Cauchy distribution with a common notation as X ˜ Cauchy(θ, λ) where θ is the location parameter and $\left(\lambda > 0\right)$ is the scale parameteris coined after Augustin Cauchy, and it belongs to the family of stable distributions that is closed under the formation of sums of independent random variables, its expected value, the variance, skewness and kurtosis do not exist but its median is given as θ (Alzaatreh et. al; 2016). Cauchy distribution has been applied in various fields like electrical theory, physical anthropology, measurement problems, mechanical theory, risk and financial analysis. It was applied by Stigler (1989) to derive an expression that is explicit for $P\left(Z_1 \leq 0, Z_2 \leq 0\right)$, where $\left(Z_1, Z_2\right)'$ follows the standard bivariate normal distribution, Johnson et al. (1994).also applied it to model the condition of effect of a fixed straight line of particles released from a area source, in physics, it is called a Lorenzian distribution, where it is defined as the energy of an unstable state distribution in quantum mechanics. Generally, the general pdf of the Cauchy distribution is defined as:

$$(\pi\alpha)^{-1}\left[1+\left\{\frac{y-\theta}{\alpha}\right\}^{2}\right]^{-1}, \quad -\infty<y<\infty \qquad (1)$$

$$\left(Norman \ \ et\,al.,\ 2005\right)$$

While the standard probability density function (PDF) of the Cauchy distribution is:

$$f(y) \quad = \ \pi^{-1}\left(1+y^{2}\right)^{-1},\ for\ \theta=0,\ \alpha=1, -\infty<y<\infty \qquad (2)$$

$$\left(Norman \ \ et\,al.,\ 2005\right)$$

The main characteristics of the Cauchy distribution can be said to be the non-existence of the mean, variance, skewness and kurtosis. Normanetal.(2005) also stated that there is no standardized form of the Cauchy distribution, this is simply because, it is not possible to standardize without using (finite) values of mean and standard deviation which do not exist in Cauchy distribution. In this case, however, a standard form need to be obtained by substituting θ=0, λ=1 which makes it to be the same with the student's t distribution with one degree of freedom, Normanetal. (2005). From equation (1),
let

$$\left(y-\theta\right)=u$$

$$f(y) \ = (\pi\alpha)^{-1}\left[1+\left\{\frac{u}{\alpha}\right\}^{2}\right]^{-1}, \ -\infty<y<\infty \qquad (3)$$

$$\left(Norman \ \ et\,al.,\ 2005\right)$$

$$f(y) \ = (\pi\alpha)^{-1}\left[1+\frac{u^{2}}{\alpha^{2}}\right]^{-1}$$

$$= \frac{1}{\pi\alpha}\left[\frac{\alpha^{2}+u^{2}}{\alpha^{2}}\right]^{-1}$$

$$= \frac{\alpha}{\pi}\left[\frac{1}{\alpha^{2}+u^{2}}\right] \qquad (4)$$

the Cumulative density function of the Cauchy distribution is derived as:

$$F(u) \ = \ \int_{-\infty}^{m} f(u)\,du$$

(5)

$$= \int_{-\infty}^{m} \frac{\alpha/\pi}{\alpha^{2}+u^{2}}\,du$$

$$= \frac{\alpha}{\pi}\int_{-\infty}^{m} \frac{1}{\alpha^{2}+u^{2}}\,du$$

$$= \frac{\alpha}{\pi}\left\{\frac{1}{\alpha}\ \tan^{-1}\left(\frac{u}{\alpha}\right)\right\}\Bigg|_{-\infty}^{(y-\theta)}$$

$$= \ \frac{\alpha}{\pi}\times\frac{1}{\alpha}\left\{\tan^{-1}\frac{(y-\theta)}{\alpha} - \tan^{-1}\left(\frac{-\infty}{\alpha}\right)\right\} \tag{6}$$

$$= \ \frac{1}{\pi}\left\{\tan^{-1}\frac{(y-\theta)}{\alpha} - \tan^{-1}(-\infty)\right\}$$

$$= \ \frac{1}{\pi}\left\{\frac{\pi}{2} + \tan^{-1}\left(\frac{(y-\theta)}{\alpha}\right)\right\}$$

$$F(y) = \ \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{(y-\theta)}{\alpha}\right) \tag{7}$$

$$F(\infty) = \ \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(\infty) = 1 \tag{8}$$

Equation (8) is the general cumulative density function (cdf) of the cauchy distribution**.** Therefore Let,

$\alpha = 1$ and $\theta = 0$ in equation (8) , the standard cumulative density function (cdf)of the cauchy distribution becomes

$$F(y) = \frac{1}{2} + \pi^{-1}\tan^{-1}(y), -\infty < y < \infty \tag{9}$$

$$(Norman\ \ et\,al.,\ 2005)$$

Let

$$y \ = h\left[Q_y(\tau)\right]$$

Then the cdf inverse $\left(F^{-1}\right)$ function of the general Cauchy distribution becomes:

$$\frac{1}{2} + \pi^{-1}\tan^{-1}\left\{\frac{h\left[Q_y(\tau)\right]-\theta}{\alpha}\right\} \ = \ Q_y(\tau) \tag{10}$$

$$\pi^{-1}\tan^{-1}\left\{\frac{h\left[Q_y(\tau)\right]-\theta}{\alpha}\right\} \ = \ Q_Y(\tau) - \frac{1}{2}$$

$$\tan^{-1}\left\{\frac{h\left[Q_y(\tau)\right]-\theta}{\alpha}\right\} \ = \ \pi\left\{Q_y(\tau) - \frac{1}{2}\right\}$$

$$\frac{h\left[Q_y(\tau)\right]-\theta}{\alpha} \ = \ \tan\left[\pi\left\{Q_y(\tau) - \frac{1}{2}\right\}\right]$$

$$h\left[Q_y(\tau)\right]-\theta \quad = \quad \alpha\tan\left[\pi\left\{Q_y(\tau)-\frac{1}{2}\right\}\right]$$

$$h\left[Q_y(\tau)\right] \quad = \quad \alpha\tan\left[\pi\left\{Q_y(\tau)-\frac{1}{2}\right\}\right]+\theta \qquad (11)$$

2.1 Cauchy Transformed Quatile Regression

Cauchy transformed quantile regression belong to the family of Cauchy distribution. Eugene et al., (2002) introduced the beta – generated family of distribution where the authors used the beta distribution as the base line distribution, this was followed by Alshawarbeh etal; (2013) who introduced the beta – Cauchy distribution which was extended to $T-R(W)$ family by Alzaartreh et. al; (2013), where the authors gave the $T-R(W)$ Cumulative distribution function as $G(x)=\int_{\alpha}^{W(F(x))} r(t)\,dt$ where $r(t)$ denotes the probability density function of the random variable T with support (a,b) for $-\infty \le a < b \le \infty$. The authors used the random variable T as the transformer to modify the random variable into an entirely new family of the generalize distribution of a random variable. In this article, Cauchy transformed quantile regression is introduced where the quantile function of Cauchy distribution is used as the transformer to transform the quantile regression into entirely new model that can handle ordinal response data and binary response data, as well manages outliers in any distribution. The general probability density function of a Cauchy distribution is given in equation (1), the general cdf is given in equation (8) while the cdf inverse $\left(F^{-1}\right)$ or the Probit function of the Cauchy distribution that will be used for data simulation is derived from the cdf of Cauchy distribution in equation (11).

The next step is to equate the cdf inverse $\left(F^{-1}\right)$ of the Cauchy function of equation (11) to the quantile regression model and solve simultaneously for the cauchit quantile regression model.

$$h\left[Q_y(\tau)\right] = \left(\beta_0^{(\tau)} + \beta_1^{(\tau)}x_i\right) \qquad (12)$$

$$\alpha\tan\left[\pi\left\{Q_y(\tau)-\frac{1}{2}\right\}\right]+\theta = \left(\beta_0^{(\tau)} + \beta_1^{(\tau)}x_i\right)$$

$$\alpha\tan\left[\pi\left\{Q_y(\tau)-\frac{1}{2}\right\}\right] = \left(\beta_0^{(\tau)} + \beta_1^{(\tau)}x_i\right)-\theta$$

(13)

$$\tan\left[\pi\left\{Q_y(\tau)-\frac{1}{2}\right\}\right] = \alpha\left(\beta_0^{(\tau)} + \beta_1^{(\tau)}x_i\right)-\theta$$

$$\pi\left\{Q_y(\tau)-\frac{1}{2}\right\} = \tan\left\{\alpha\left(\beta_0^{(\tau)} + \beta_1^{(\tau)}x_i\right)-\theta\right\}$$

$$\left.\begin{array}{l} Q_y(\tau) = \pi^{-1}\tan^{-1}\left\{\alpha\left(\beta_0^{(\tau)} + \beta_1^{(\tau)}x_i\right) - \theta\right\} + \frac{1}{2}, \\ -1 < Q_y(\tau) < 1, \ -\infty < x_i < \infty, \ 0 < \tau < 1 \end{array}\right\} \quad \textbf{(14)}$$

$Q_y(\tau)$ = the response variable and the cdf inverse $\left(F^{-1}\right)$ of the distribution to be estimated

x_i = the covariates to be simulated

β_0 = the intercept parameter

β_i = the unknown parameters

τ = specified quantiles of the model. This research examines the following quantiles: 0.05, 0.25, 0.5, 0.75, 0.95

Equation (14) is the proposed Cauchy transformed quantile regression model.

2.2 Data Simulation and Analysis

The simulation experiments were adopted from the Hao and Naiman (2007). The simulation were performed as follows: a design matrix x_i for the explanatory variables which is an $n \times k$, $k = 3$ is a fixed number of explanatory variable for a sample size n, drawn from the cdf inverse function of independent normal distribution where $n = 150$. In order to estimate the standard error, the confidence interval and the p-values, the simulation experiment by Efron (1979) was followed by bootstrapping the sampled values 200 times. The data were generated using the the R – code of $x_i = rnorm\left(n, \bar{x}, sd\right)$, where $x_i = \left(x_i, x_2\right)$ while x_3 = (1.2+(1.5*xc)) where xc is $rnorm\left(n, \bar{x}, sd\right)$. The mean and standard deviation for the explanatory variables were adopted from Nwakunya et. al (2019) which gave the mean and standard deviation for x_1 as 131.7143 and 2.728419 respectively, mean and standard deviation of x_2 are given as 50.57143 and 1.361518 respectively while the mean and standard deviation of xc are given as 145.6429 and 6.073667 respectively. The response variable for the experiment has the design matrix of $\left(n \times 1\right)$ where n is 150 but in order to estimate the standard error, confidence interval and the p-values, it was bootstrapped 200 times. Some degree of outliers were infused in the simulated data using $y_i = 1 + 3 \times r + error$ but in order to examine the equivariance of the location parameter the response variable is simulated as $y_i = sample\left(c\left(error1, error2\right)\right)$ where $r = \left(n, \bar{x}, sd\right)$, $error$ are random numbers sampled without replacement from $\left(error1 + error2\right)$, $error1$ is $\left(n - n \times out.per\right)$ sample size, $out.per = 20$ and

$error\,2 = rnorm\left(n \times out.per, \max\left(error\,1\right) \times 8,1\right)$, n is 150 then bootstrapped 200 times so as to estimate the standard error, the confidence interval and the p-values.

3.0 Results

The result of Cauchit Quantile Regression model shows that all the parameters in Tables 1 are not significant for all the study quantiles except for the intercept of the 5th, 25th, and 95thquantilesand x_3 of the 95th quantile. Figure 4, shows that the residual graphs cluster round 0 and are uniformly spread around the negative and positive axis.

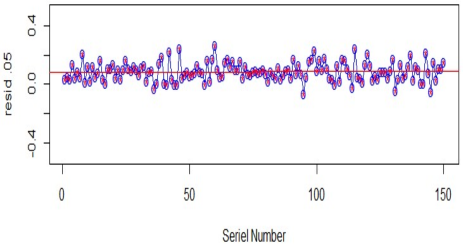
Quantiles	parameters	coefficient	Std error	t-value	Pr(>|t|)
0.05	**intercept**	-0.10582	0.03783	-2.79741	0.00585
	X_1	0.00490	0.00594	0.82543	0.41047
	X_2	-0.02644	0.02362	-1.11965	0.26470
	X_3	0.00255	0.00796	0.32018	0.74929
0.25	**intercept**	-0.05300	0.01883	-2.8152	0.00555
	X_1	0.00258	0.00279	0.92380	0.35711
	X_2	-0.01203	0.01443	-0.83380	0.40575
	X_3	-0.00228	0.00465	-0.4903	0.62467
0.5	**intercept**	-0.01672	0.02250	-0.74326	0.45852
	X_1	0.00175	0.00410	0.42770	0.66950
	X_2	-0.00327	0.01576	-0.20765	0.83579
	X_3	-0.00164	0.00401	-0.4101	0.68236
0.75	**intercept**	0.04291	0.02887	1.48633	0.13935
	X_1	-0.00092	0.00458	-0.20186	0.84031
	X_2	-0.00803	0.01472	-0.54564	0.58615
	X_3	-0.00111	0.00692	-0.16052	0.87270
0.95	**intercept**	0.13181	0.04557	2.89237	0.00441
	X_1	0.00370	0.00650	0.56902	0.57022
	X_2	-0.02862	0.02447	-1.16996	0.24392
	X_3	-0.02182	0.00969	-2.25218	0.02580

Table 1: Estimated Parameters of 150 Simulated Data for the Cauchy transformed QR Model

Table 2, show that the skewness for all the study quantiles approximate to zero while the kurtosis for all the study quantiles are a little above 3, the means are equal to the medians for all the study quantiles, the mean square errors (MSE), the root mean square errors (RMSE) and the standard deviation (SD) approximate to zero.

quantiles	Skewness	kurtosis	mean	Median	RMSE	MSE	SD
0.05	0.3853	3.4741	0.10293	0.092044	0.12328	0.0152	0.06810
0.25	0.0767	3.73875	0.04326	0.0473	0.0763	0.0058	0.0630
0.5	0.0556	3.7494	0.00000	0.0000	0.0628	0.0039	0.0630
0.75	-0.0773	3.7927	-0.0416	-0.0416	0.0756	0.0057	0.0634
0.95	-0.1531	3.5195	-0.0952	-0.0959	0.1156	0.0133	0.0658

Table 2: Descriptive Analysis for the Residuals of the Cauchit Quantile Regression Model



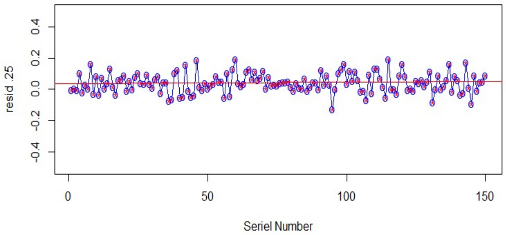
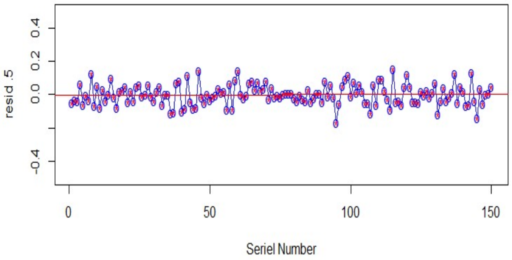


Figure 4$_A$: Residual Plot of Quantile 0.05; Figure 4$_B$: Residual Plot of Quantile 0.25



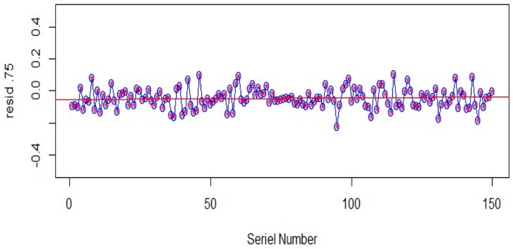


Figure 4$_C$: Residual Plot of Quantile 0.5; Figure 1$_D$: Residual Plot of Quantile 0.75

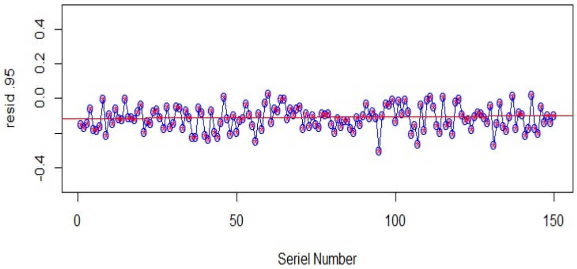


Figure 1$_E$: Residual Plot of Quantile 0.95

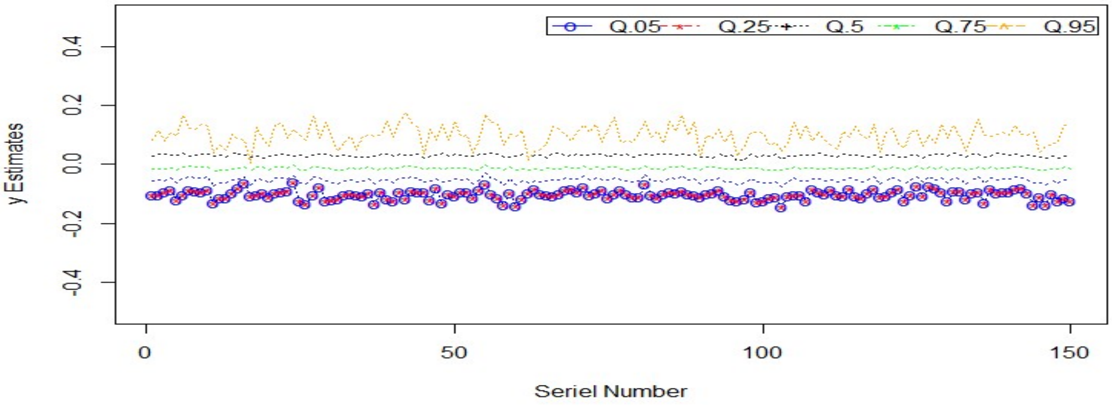


Figure 2: Plot for the Estimate of the Cauchy transformed Quantile RM

3.0 Conclusion

Cauchy transformed Quantile Regression model shows ability to manage outliers when it is applied to simulated data set in Table 2. Its residual graphs in Figures 1 shows little presence of outliers, the summary results of the residuals for the Cauchy transformed Quantile Regression model in Tables 2, show that both its skewness and kurtosis are closer to 0 and 3 respectively. Its median and mean are equal, its root mean square error and standard deviation are smaller and closer to zero. Based on the above remarks we therefore conclude that Quantile Regression model can handle data with outliers when transformed with Cauchy distribution. Hence it can be recommended for Cauchy transformed Quantile Regression model to be used for data with outliers.

References

Alshawarbeh, E, Famoye, F, Lee, C; (2013), Beta-Cauchy distribution: some properties and applications. *J. Stat. Theory Appl. 12(4), 378–391*

Alzaatreh, A, Lee, C, Famoye, F: A & Ghash, I (2016), The generalized Cauchy family of distributions with applications, *Statistical Distributions and Applications, 3:12, DOI 10.1186/s40488-016-0050-3*

Bonat, W.H., Ribeiro, P.J.,& Zeviani, W. M. (2012),Regression models with responses on the unity interval: specification, estimation and comparison; *Rev. Bras. Biom.*, S~ao Paulo, v.30, n.4, p.415-431, 2012

Efron, B. (1979), Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Eugene, N, Lee, C, Famoye, F; (2002), Beta-normal distribution and its applications. *Commun. Stat. Theory Methods 31(4), 497–512*

Hao L. &Naiman,D.Q., (2007). Quantile Regression*; 01-Hao.qxd. 3/13/2007.3.28*

Koenker, R., & Bassett, J., G. (1978). Regression quantiles. *Econometrica, 46,* 33-50.

Manski, C.F (1985),Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator, *Journal of Econometrics, 27(3),*313-333

Nwakuya M. T., Nwabueze J. C., Onyegbuchulem, B. O., & Imoh, J. C. (2019), Response variable transformation for quantile regression model, *International Journal of Scientific & Engineering Research (10)* 8.