European Journal of Statistics and Probability, 14 (1) 30-40, 2026
Print ISSN: 2055-0154(Print)
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

A Robustness Analysis of Logistic Quantile Regression Model with Contaminated Dataset

Besta Okey Onyegbuchulem[1], Chialuka Adline Onyegbuchulem[2], Johndamascene Chikere Imoh[3], Churchill Ugochukwu Nwokeforo[4]
[1,3 & 4]. Department of Mathematics/Statistics, Dr. B.U. Nzeribe State Polytechnic Awo-Omamma, Imo State, Nigeria
[2.] Department Mathematics, Alvan Ikoku Federal University of Education Owerri, Nigeria
Corresponding Author's Email Adress: bokey@imopoly.net
ORCID ID: 0000-0002-3507-3109

doi: https://doi.org/10.37745/ejsp.2013/vol14n13040 Published September 03, 2026

Citation: Onyegbuchulem B, O, Onyegbuchulem C.A., Imoh J.C., Nwokeforo C.U. (2026) A Robustness Analysis of Logistic Quantile Regression Model with Contaminated Dataset, *European Journal of Statistics and Probability*, 14 (1) 30-40

Abstract: *Traditional mean regression models may be inadequate when dealing with asymmetric response distributions. In such cases, quantile regression offers a more robust alternative, accommodating outliers and error distribution misspecification by characterizing the entire conditional distribution of the outcome variable. This paper proposes a robust logistic quantile regression model, which involves adding 0.5 and 0.05 differently to the numerator and denominator of the logit link model. The performance of these two models is evaluated and compared using goodness of fit, AIC, skewness, kurtosis, RMSE, and MSE. Results indicate that both models are robust, with the model adding 0.5 to both the numerator and denominator of the logit link function performing slightly better. This study underscores the importance of selecting an appropriate distribution in quantile regression analysis, particularly for modeling skewed and normal data.*

Keyword: Logistic quantile regression, Pseudo R-square, AIC, RMSE

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, Bottai, & Columbu.(2010). This model obviously was able to address some of the issues mentioned above regarding the choice of a measure of central tendency.

Quantile regression metamorphosed into Binary response quantile regression model which came as a result of the deficiencies inherent in both linear regression model as well the quantile regression model. The problem of modeling two alternatives in the response variable such as modeling probability of yes or no distribution in the response variables or modeling other situations where the response variable takes on only two values, this cannot yield continuous quantiles that can be modelled via linear regression (Manski, 1975; 1985, Koenker & Hallock (2001) and Kordas, 2006)

Empirically when there are bounded outcomes, classical standard regression models may not be practically sufficient to handle such data set. A more practical way to analyze such outcomes is to use logistic quantile regression. Bounded outcomes are outcomes with restricted range, typically between a minimum and maximum value. Logistic quantile regression for bounded outcomes is therefore used to model outcomes such as proportions: values between 0 and 1, percentages: values between 0 and 100, scores between 0 and 100, probabilities: values between 0 and 1 etc. the logistic quantile regression approach models the quantiles of the bounded outcome variables using logistic link function.

Logistic quantile regression is kind of binary response quantile regression model for modeling the conditional quantiles of a bounded outcome variable that uses

* Bounded Outcome: The dependent variable (Y) is bounded within a certain level.
* Quantile Modeling: Models the τ^{th} quantile of Y given predictors X :

$$Q_\tau\left(Y/X\right) = \frac{1}{1+e^{-X\beta(\tau)}}$$

* Logistic Link: Ensures predictions stay within the bounds of 0 and 1.

Quantiles and Transformation

The quantiles unlike the mean can analyze transformations of the outcome variable. For a numeric outcome Y, the cumulative distribution function CDF can be written as

$$F\left(y\right) = P\left(Y \le y\right)$$

And the p^{th} quantile of Y is given by

$$Q\left(p\right) = F^{-1}\left(y\right).$$

The inverse function of the CDF of F is non-increasing (non-increasing) and continuous. Some of these characteristics of quantiles are as follows:

$$Q_{h(y)}\left(p\right) = h\left[Q_y\left(p\right)\right] \tag{1}$$

Where $h(.)$ is a function that is non-decreasing. This implies that $g(.)$ is a monotonically decreasing function, the following equality must hold.

$$Q\left[g\left(Y\right)\right] = g\left[Q\left(Y\right)\right] \Leftrightarrow Q\left(Y\right) = g^{-1}\left(Q\left[g\left(Y\right)\right]\right)$$

Thus taking the inverse of $Q\left[g\left(Y\right)\right]$ will give us $g^{-1}\left(Q\left[g\left(Y\right)\right]\right)$, and the inverse of $g\left[Q\left(Y\right)\right]$ is $Q\left(Y\right)$. The equation (1) holds because if $h(.)$ is a monotone function, it then means that the following will be applied

$$P\left(Y\leq y\right)=P\left[h\left(Y\right)\leq h\left(y\right)\right]$$

Logistic Quantile Regression

In this section, we follow the description provided by Wong (2018): Suppose there are n sample observations and q explained variables,

$$x_i=\left(x_{i1},\ x_{i2},---,x_{iq}\right)$$

where $i=1,2,---,n,$ and a bounded continuous response variable y_i, which lies between the known interval $\left(y_{\min},y_{\max}\right)$. Note that $y_{\min}$ and $y_{\max}$ are constants which denote the limits of the outcome variable. Wong (2018) stated that Bottai et al. (2010) chose the following logistic link function:

$$h\left(y_i\right)=\log it\left(y_i\right)\ =\ \log\left(\frac{y_i-y_{\min}}{y_{\max}\ -y_i}\right)\ \left(Wong,2018\right) \tag{2}$$

which transforms the bounded interval to [0; 1]. The logit function of $Q_{yi}(P)$ will then be:

$$h\left[Q_{yi}\left(P\right)\right]=\log it\left(Q_{yi}\left(P\right)\right)\ =\ \log\left(\frac{Q_{yi}\left(P\right)-y_{\min}}{y_{\max}\ -Q_{yi}\left(P\right)}\right) \tag{3}$$

The next step is to solve $Q_{y_i}\left(p\right)$ from the following equality:

$$\log\left(\frac{Q_{yi}\left(P\right)-y_{\min}}{y_{\max}\ -Q_{yi}\left(P\right)}\right)=\beta_{p,0}\ +\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq} \tag{4}$$

$$\frac{Q_{yi}\left(P\right)-y_{\min}}{y_{\max}\ -Q_{yi}\left(P\right)}=\exp\left(\beta_{p,0}\ +\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right) \tag{5}$$

$$\left.\begin{aligned}&Q_{yi}\left(P\right)\left[1+\exp\left(\beta_{p,0}\ +\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)\right]\\&\quad=y_{\max}\exp\left(\beta_{p,0}\ +\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)+y_{\min}\end{aligned}\right\} \tag{6}$$

Let assume that the observations are "n" in number with $x_i=\left(x_{i1},x_{i2},\ldots,x_{iq}\right)$, where i = 1,2, . . . ,n, with y_i as response variable with bounded values that lies between the intervals $\left(y_{\min},\ y_{\max}\right)$.

Note also that $y_{\min}$ and $y_{\max}$ are constant which denote the limits of the response variable and ε is the value that will be added to the boundaries.

$$=\frac{(y_{\max}+\varepsilon)\exp\left(\beta_{p,0}+\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)+(y_{\min}+\varepsilon)}{1+\exp\left(\beta_{p,0}+\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)}\,(Wong,\,2018) \quad (7)$$

Following Wong (2018), logistic quantile regression with two different boundaries (0.5 and 0.05) is therefore proposed in other to examine the one that is most robust. The two models therefore become:

$$Q_{yi}(P)=\frac{(y_{\max}+0.5)\exp\left(\beta_{p,0}+\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)+(y_{\min}+0.5)}{1+\exp\left(\beta_{p,0}+\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)} \quad (8)$$

And

$$Q_{yi}(P)=\frac{(y_{\max}+0.05)\exp\left(\beta_{p,0}+\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)+(y_{\min}+0.05)}{1+\exp\left(\beta_{p,0}+\beta_{p,1}x_{i1}+---+\beta_{pq}x_{iq}\right)} \quad (9)$$

2.2 Data Description

The idea of generating data sets with outliers is to ascertain which of the two models of logistic quantile regression model is robost in handling outliers better. The response variable is generated as follows:

$$y = 1 + 3\times r + error \quad (10)$$

where

$r =$ normal generated random number of size n

$error =$ Random numbers Sampled without replacement from the elements of $(\text{error1 and error 2})$

$error1 =$ Normal generated random number for $(\text{n - n}\times\text{out.per})$ sample size

$error2 =$ Normal generated random number of size $(\text{n}\times\text{out.per})$ with mean $=(\text{maximum value of } (\text{error } 1)\times 8)$ and standard deviation $=1$

$\text{out.per} = 20$

$n = 3000$

The explanatory variables are generated using the cdf inverse function of the normal distribution shown by Nwakuya et al, (2019) and Onyegbuchulem et al, (2026):

$$x=\sqrt{2}\,erf^{-1}(2F-1),\quad p\in(0,1) \quad (11)$$

Were

$x =$ inverse function

erf = the error function

$F = cumulative\ density\ \ function\ (cdf\ )$

The explanatory variables data sets are generated for x_1 following the cdf inverse function in equation (11):

$$x_i = rnorm(n, \bar{x}, sd) \tag{12}$$

Where

$x_i =$ the covariates where $i = 1,\ 2$ and 3

n = the sample size (3000)

$\bar{x} = mean$

sd = standard deviation

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 x_3 are given as 145.6429 and 6.073667 respectively and these explanatory variables $(x_1, x_2\ and\ x_3)$ were all generated using normal distribution in R – code. The models to be compare are shown in equations (8 and 9).

Data analysis and Results

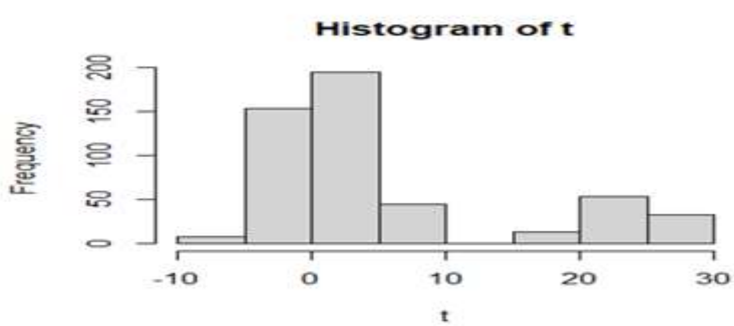

Fig.1: Histogram of the simulated

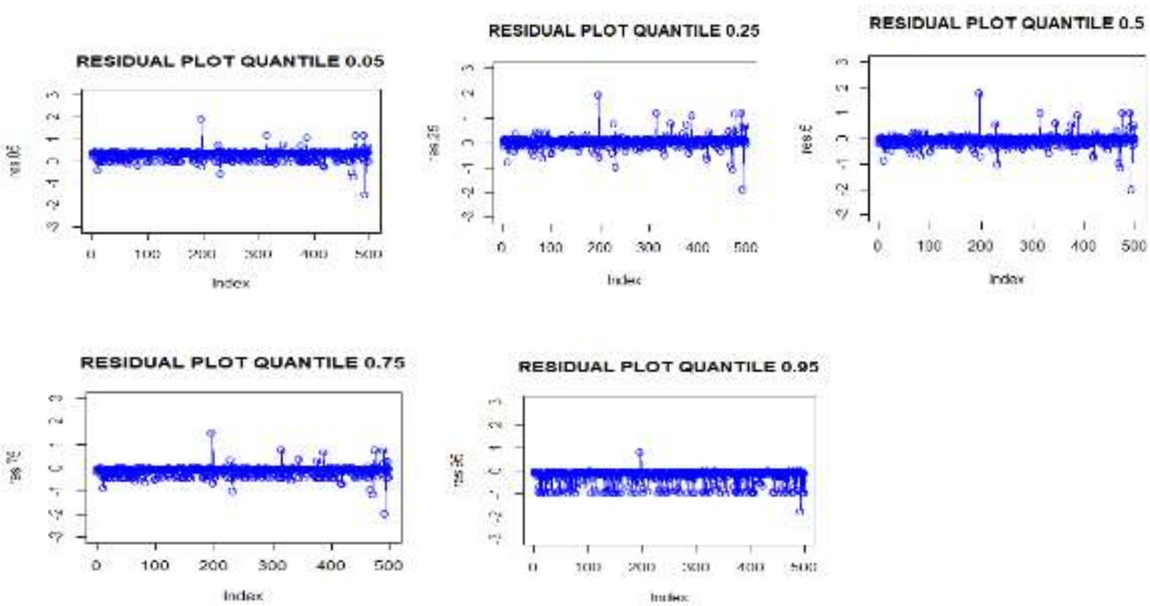

Fig.2 Residual plot for Logistic quantile regression for the various quantiles of equation

Table 1: Results of the intercepts and slopes for the various quantiles of equation 8

	0.05	0.25	0.5	0.75	0.95
constant	-1.7597	-1.4605	-1.3461	-1.3230	-1.3165
X1	0.1294	0.1456	0.0994	0.1234	0.1454
X2	-0.0001	-0.0341	0.0293	0.0156	0.0167
X3	0.0129	0.0193	0.0048	0.0023	0.0037

The plot of table 1 shows that at quantile 0.05 y is expected to deceased at the intercept term by 1.7597 and slope b_2 by approximately 0 while there are an increase at slopes b_1 and b_3 by 0.1294 and 0.0129 respectively. At quantile 0.25 is an upward improvement with a decrease of the intercept by 1.4605 and slope b_2 by approximately 0.0341 while there are an increase at slopes b_1 and b_3 by 0.1456 and 0.0193 respectively. For quantile 0.5, there is a significant upward growth of -1.3461 for the intercept and 0.0994, 0.0293 and 0.0023 for slopes $b1$, b_2 and b_3 respectively. For quantile 0.75, the intercept continued to decrease by 1.3230 and the slopes continued to increase with 0.1234, 0.0156 and 0.0023 for b_1, b_2 and b_3 respectively. And lastly for quantile 0.95 the intercept decrease by 1.3165 and the slopes increase by 0.1454, 0.0167 and 0.0037 for b_1, b_2 and b_3 respectively.

Table 2: Diagnostic test results for the various quantiles of equation 8

Quantiles	0.05	0.25	0.5	0.75	0.95
Pseudo R	0.7310	0.8052	0.8625	0.9151	0.9181
AIC	394.4056	-125.7974	-364.4642	-471.1442	-200.6331
Skewness	-0.6792	0.1163	-0.0517	-0.6311	-1.4038
Kurtosis	18.9084	24.1752	24.6659	19.7960	4.0331
MSE	0.1546	0.0641	0.0558	0.0612	0.1651
RMSE	0.3932	0.2531	0.2361	0.2474	0.4063
Resid mean	0.3240	0.0806	-0.0470	-0.1081	-0.2344
Resid media	0.3956	0.1148	0.0000	-0.0391	-0.0730

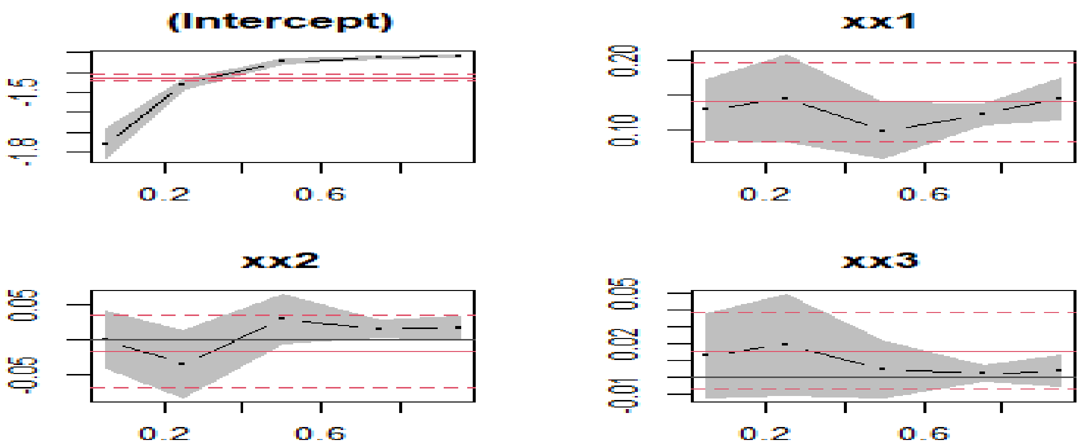

Fig.3: Plot of Logistic quantile regression for the various quantiles estimates of equation 9

Table 3: Results of the intercepts and slopes for the various quantiles of equation 9

	0.05	0.25	0.5	0.75	0.95
constant	-1.8703	-1.5202	-1.3929	-1.3675	-1.3577
X1	0.1363	0.1544	0.1040	0.1290	0.1606
X2	-0.0006	-0.0395	0.0311	0.0158	0.0167
X3	0.0137	0.0210	0.0039	0.0028	0.0050

The plot of table 3 shows that at quantile 0.05 y is expected to decease at the intercept term by -1.8703 and slope b_2 by approximately 0 while there are an increase at slopes b_1 and b_3 by 0.1363 and 0.0137 respectively. At quantile 0.25 is an upward improvement with a decrease of the intercept by 1.5202 and slope b_2 by approximately 0.0395 while there are an increase at slopes b_1 and b_3 by 0.1544 and 0.0210 respectively. For quantile 0.5, there is a significant upward growth of -1.3929 for the intercept and 0.1040, 0.0311 and 0.0039 for slopes $b1$, b_2 and b_3 respectively. For quantile 0.75, the intercept continued to decrease by 1.3675 and the slopes continued to increase with 0.1290, 0.0158 and 0.0028 for b_1, b_2 and b_3 respectively. And lastly for quantile 0.95 the intercept decreased by 1.3577 and the slopes increased by 0.1606, 0.0167 and 0.0050 for b_1, b_2 and b_3 respectively.

Table 4: Diagnostic test results for the various quantiles of equation 9

QUANTILES	0.05	0.25	0.5	0.75	0.95
PSEUDO R	0.6789	0.7767	0.8410	0.8974	0.8859
AIC	693.3808	81.8535	-160.2722	-226.8405	220.6706
SKEWNESS	-0.5705	0.1196	-0.0461	-0.5093	-1.3960
KURTOSIS	69.1910	69.2105	72.6852	70.9693	12.4409
MSE	0.2571	0.1412	0.1291	0.1317	0.3066
RMSE	0.5070	0.3758	0.3593	0.3629	0.5537
RESID MEAN	0.3746	0.0904	-0.0526	-0.1203	-0.2979
RESID MEDIAN	0.4577	0.1285	0.0000	-0.0428	-0.0909

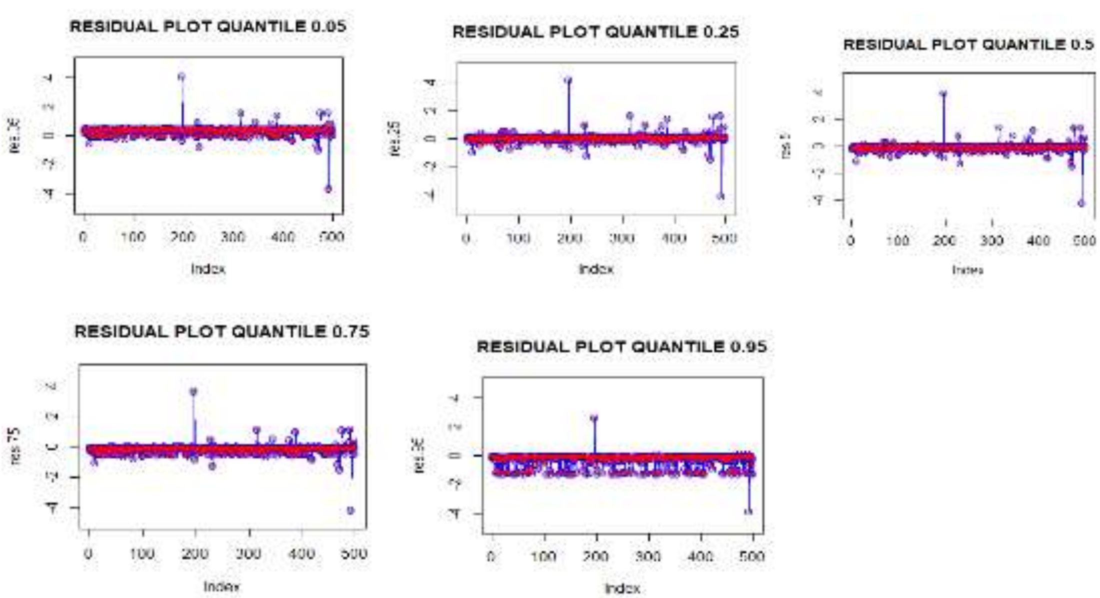

Fig.4: Residual plot for Logistic quantile regression for the various quantiles of equation 9

The pseudo R- square results in Table 4 indicate that the predictor variables explained 67.89% of the variation at the 0.05 quantile, 77.67% at the 0.25 quantile, 84.10% at the 0.50 quantile, 89.74% at the 0.75 quantile and 88.59% at the 0.95 quantile of the response variable Y. given the relatively high pseudo R-square values across these five quantiles, the model demonstrates a good fit and explains a substantial proportion of the variation at each quantile.

Model diagnostic Test

The results of the residual analysis in table 2 and 4 show that their skewness are moderately skewed for both models with -0.67 and -0.57 indicating potential under estimation, slightly right skewed with 0.12 and 0.11 for quantile 0.25 indicating minor overestimation tendency and approximately symmetry, it showed approximately symmetry residuals at quantile 0.50 with -0.05 for both models indicating well-behaved and normally distributed residuals. At quantile 0.75 the models are moderately left-skewed and asymmetric for both models with -0.63 and -0.51 indicating potential underestimation tendencies for both models. At quantile 0.95 the models showed extremely left skewed residuals with -1.40 for both models suggesting underestimation tendency and extreme asymmetry, potentially indicating model issues or outliers.

The kurtosis values of logistic quantile regression with the addition of 0.5 in table 2 show quantile 0.05, 0.25, 0.5, and 0.75 have kurtosis values of 18.91, 24.17,24.67 and 19.80 indicating leptokurtic distributions which suggest heavy tails with many extreme values or outliers and quantile 0.95 with low kurtosis value of 4.03 indicating platykurtic distribution which suggest light tails with fewer extreme values. The kurtosis of logistic quantile regression model with addition of 0.05 in table 4 show quantile 0.05, 0.25, 0.5 and 0.75 produced extreme kurtosis values of 69.19, 69.19, 72.68, and 70.97 respectively which indicate that the residuals of the quantile are leptokurtic distributed, which suggest heavy tails with many extreme values but quantile 0.95 has a low kurtosis value of 12.44 indicating a platykurtic distribution which suggest light tails with fewer extreme values. It can be observed that the two model (addition of 0.5 and 0.05) share similar characteristics.

For the mean and median in the two models in Table 2 and 4, it can be observed that they are equal for the both models at quantile 0.05, 0.25 and 0.5 suggesting no skewness and approximately normally distributed residual values. But for quantile 0.75 and 0.95, as can be seen in both table 2 and 4 showed unequal mean and median. This suggest that the residual distribution at quantile 0.75 and 0.95 are asymmetric which indicate heavy tails or outliers. We can equally observe that the mean and median of the two model at quantile 0.5 as can be seen in Table 2 and 4 are approximately zero suggesting that the distribution of the residuals are centered at zero and are symmetric around zero. This indicates normal or approximately normal residual distribution.

The mean square errors and the root mean square errors of the residual of the first and second modeled logistic quantile regression shows significant low error. For the MSE of the first model that has the addition of 0.5 quantiles 0.05, 0.25, 0.5, 0.75 and 0.95 have their MSE as 0.1546, 0.0641, 0.0558, 0.0612 and 0.1651 respectively. This shows a very low average square errors, indicating good model fit across the quantiles. Also, the Root Mean Square error (RMSE) of the residual for quantiles 0.05, 0.25, 0.5, 0.75 and 0.95 have their RMSE as 0.3932, 0.2531, 0.2361, 0.2474 and 0.4063 respectively between the predicted and the actual value this equally suggest a moderately good fit.

The mean square errors and the root mean square errors for the second logistic model with the addition of 0.05 has 0.2571, 0.1412, 0.1291, 0.1317 and 0.3066 for quantiles 0.05, 0.25, 0.5, 0.75 and 0.95 respectively this suggests a relatively low average square errors. Its Root mean square errors has 0.5070, 0.3758, 0.3593, 0.3629 and 0.5537 for quantiles 0.05, 0.25, 0.5, 0.75 and 0.95 respectively. This shows that the first model with the addition of 0.5 is a little more robust then the model with the addition of 0.05.

CONCLUSION

In conclusion, this paper has proposed a logistic quantile regression model tailored to fit bounded response variables. We have demonstrated the robustness of this parametric quantile regression model, which effectively accommodates outliers and provides a flexible and robust framework for estimating interval or strictly positive response variables. The model's ease of interpretation has been showcased in the application section, and its accessibility via user-friendly functions in an R package has been highlighted. The main function for logistic quantile regression offers automatic model selection, full inference, residual and envelope plots for goodness-of-fit assessment, AIC and residuals plots for model evaluation. Additionally, we have introduced a robust logistic quantile regression model that incorporates adjustments of 0.5 and 0.05 to the numerator and denominator of the logit link model. Comparative analysis using goodness of fit, AIC, skewness, kurtosis, RMSE, and MSE has shown that both models are robust, with the 0.5 adjustment yielding slightly better performance. Theoretically, the transformation methodology presented herein can be effortlessly extended to other quantile regression models, including mixed effects models for longitudinal data and interval censored responses, which are prevalent in fields such as biology, medicine, and pharmacology. This underscores the versatility and broad applicability of the proposed logistic quantile regression model.

REFERENCES

Bottai, M., & Columbu, S.(2010), Cognitive deterioration in cancer patients using logistic quantile regression, Statistics in Medicine 23, 309–317.

Koenker, R. (1994). Confidence intervals for regression quantiles. In *Proceedings of the 5th Prague symposium on asymptotic statistics* (pp. 349–359). New York: Springer-Verlag.

Koenker, R., & Hallock, K. F. (2001). Quantile regression: An introduction. *Journal of EconomicPerspectives, 15,* 143–156.

Kordas, G. (2006), Smoothed binary regression quantiles, journal of Applied econometrics,https://doi.org/10.1002/jae.843

Manski, C.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics3,* 205–28.

Print ISSN: 2055-0154(Print)
Online ISSN 2055-0162(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

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.

Onyegbuchulem, B.O., Nwakuya, M.T.., Nwabueze, J.C. & Out, C.O. (2019). Evaluation of the Assumptions of the Linear Regression Model on the Quantile Regression Model, International Journal of Advancements in Research & Technology, Vol 2, Issue 9, September-2019.

Onyegbuchulem, B.O., Onyegbuchulem, C.A., Ogwo, O. & Oliwe, E. (2026). Diagnostic tests for quantile regression with weibull response variables, Royal Statistical Society Nigeria Local Group, Conference Proceeding, https://publications.funaab.edu.ng/index.php/RSS

Wong, V (2018) Logistic quantile regression to evaluate bounded outcomes, Mathematical Statistics Stockholm University,http://www.math.su.se