Science Journal of Applied Mathematics and Statistics
2025, Vol. 13, No. 2, pp. 27-33
https://doi.org/10.11648/j.sjams.20251302.11

Research Article

A Robust Quantile Regression Model for Count Data: The Half Cauchy Transformation Approach

Runyi Emmanuel Francis[1, *], Maureen Tobe Nwakuya[2], Maxwell Azubuike Ijomah[2]

[1]Department of Statistics, Federal Polytechnic Ugep, Ugep, Nigeria

[2]Department of Mathematics and Statistics, University of Port Harcourt, Port Harcourt, Nigeria

Abstract

This paper introduces an innovative approach to modelling count data through the introduction of a robust quantile regression model, the Half Cauchy Quantile Regression (HCQR). Count data is frequently challenged by outliers and skewed distributions. By integrating the heavy-tailed properties of the Half Cauchy distribution into the quantile regression framework, the HCQR model offers reliable estimates, particularly in the presence of extreme values. Quantile regression models, including HCQR, typically exhibit greater robustness to such extremes compared to traditional methods. The study highlights the limitations of traditional count regression models, such as the Negative Binomial Regression (NBR), particularly their performance inadequacies within the quantile regression framework. A comparative analysis using real-world crime data illustrates that the HCQR model substantially outperforms the NBR model. By integrating the half Cauchy distribution into the quantile regression framework, the HCQR model was formulated. In the Half Cauchy Quantile Regression Model, the Half Cauchy quantile function is used to transform the traditional quantile regression outputs, accommodating the characteristics of the Half Cauchy distribution. This superiority is demonstrated through improved metrics such as lower Standard Deviation, Skewness, Kurtosis, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), establishing HCQR's enhanced robustness and predictive accuracy.

Keywords

Quantile Regression Model, Count Regression, Half Cauchy Distribution, Half Cauchy Quantile Regression Model, Robustness to Extreme Values

1. Introduction

In traditional linear regression, the primary focus is on estimating the conditional mean of the response variable, which represents the center of the data distribution. However, this statistical regression approach often fails to capture important information about the tails of the distribution, particularly in the presence of extreme values. Quantile regression addresses this limitation by modelling the conditional quantiles of the response variable. This allows quantile regression to estimate a broader range of statistical relationships, capturing not only the central tendency of the response variable but also its be-

[*]Corresponding author: runyiemmanuel@gmail.com (Runyi Emmanuel Francis)

Received: 27 March 2025; **Accepted:** 7 April 2025; **Published:** 29 April 2025

havior at various points in the data distribution. Quantile regression has gained recognition as a powerful tool for analyzing data beyond the mean, especially in the presence of outliers, skewness, and heavy-tailed distributions [5, 6]. However, existing quantile regression models often struggle to handle count response data, where outliers and extreme values are more pronounced [3]. Traditional count regression models like Poisson and Negative Binomial regressions, while effective, may not adequately address the presence of extreme outliers in the data [11]. This paper therefore presents the Half Cauchy Quantile Regression (HCQR) model, leveraging the robust properties of the Half Cauchy distribution. By incorporating the Half Cauchy distribution's heavy-tailed nature in the quantile regression framework, the proposed HCQR model enhances the robustness of quantile regression, particularly when modeling count data with outliers.

Quantile regression, introduced by [7], extends traditional regression by estimating conditional quantiles instead of focusing solely on the conditional mean [2]. It has been widely used to model data with heteroscedasticity, skewness, and extreme values [8]. [1, 4] provided a comparative analysis of OLS and quantile regression, demonstrating that quantile regression offers more detailed insights into the impact of covariates across the entire distribution of response variables.

There has been a paradigm shift where several approaches have been introduced which focused on extending QR to handle data that departs from normality in regression models, especially when the response variable follows an asymmetrical distribution. [10] proposed the Birnbaum-Saunders quantile regression using the Birnbaum-Saunders distribution. [12] proposed a new quantile regression using the unit Birnbaum–Saunders distribution. [9] proposed the Transmuted unit Rayleigh quantile regression model as an alternative to Beta and Kumaraswamy quantile regression models [13, 14] investigated the robustness of QR to outliers using a Cauchy transformation, proving its effectiveness in handling extreme values. Their study emphasized QR's superiority in providing robust estimations compared to traditional regression methods.

The Half Cauchy distribution is also one of such robust method, which is particularly suitable especially when the response variable follows a non-normal distribution that is a count data exhibiting a heavy-tailed nature. Previous research has demonstrated the effectiveness of using the Cauchy distribution in quantile regression [12, 13], but the Half Cauchy distribution, a truncated version of the Cauchy distribution, is better suited for modeling count data.

Traditional count regression model, like Negative Binomial, assume specific distributions for count data. However, these models may struggle when the data exhibit extreme outliers. The proposed Half Cauchy Quantile regression (HCQR) model aims to overcome this limitation by incorporating the Half Cauchy distribution.

2. Material & Methods of Research

2.1. Count Regression Model - Negative Binomial Regression

Count data refers to data that represents the number of occurrences of an event within a specified time. In other words, count data are discrete (countable), non-negative integers (0, 1, 2, 3…) that represents the frequency of occurrence of an event or phenomenon.

Modelling count data with quantile regression can be more complex than other types of data due to its discrete nature, it's restriction to non-negative values, and the presence of overdispersion. Count regression models are specialized statistical models designed to handle count data. These models are usually used when the response variable is a count, and we are interested in understanding the relationship between the count variables and one or more predictor variables. One commonly used count regression model is the Negative Binomial regression model.

In the context of modelling count data, Negative Binomial Regression (NBR) model is widely used to handle overdispersed data, where the variance exceeds the mean. It is particularly suited for situations where traditional Poisson regression, which assumes equal mean and variance, is inadequate. The NBR is a generalization of the Poisson model. In practice, count data often exhibits overdispersion, meaning that the variance is greater than the mean.

The Negative Binomial regression model is given as:

$$\mathrm{Log}\left(\mathrm{E}(\mathrm{Y_i} \mid \mathrm{X})\right) = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_p \mathrm{X}_p$$

here $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients, and $X_1, X_2, \ldots, X_p$ are the predictors.

2.2. Quantile Regression with Jittered Data

Quantile regression is a statistical technique that extends the capabilities of ordinary least regression by estimating the conditional quantiles of the response variable. However, quantile regression typically assumes that the response variable is continuous. The challenge with applying continuous quantile regression to count data is that count data is inherently discrete (i.e. it consists of integer values (0, 1, 2, 3…) while quantile regression models are designed for continuous data. When dealing with count data and the assumption is violated, then transformation is necessary to transform the data into a continuous form. One of such transformation is jittering, which simply involves adding some amount of noise to the count data.

Jittering is a simple technique where a small amount of random noise is added to the count data, making the data more continuous while still retaining the characteristics of the original count variable. For this study, the noise is drawn from a half Cauchy distribution.

Given that the response variable (Y) is count, (i.e. $Y = y_1, y_2, \dots, y_n$) where y_i represents the count for the i^{th} observation, we apply jittering by adding a random noise ϵ_i which is drawn from the half Cauchy distribution. This transformed data $Y_{jittered}$ is given as:

$$Y_{\text{jittered}} = Y + \epsilon$$

where $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ are the jittering noise values, drawn from the half Cauchy distribution.

For each observation y_i, the jittered data takes the form,

$$y_i^{\text{jittered}} = y_i + \epsilon_i$$

where $\epsilon_i \sim$ Half Cauchy. This means that each ϵ_i is a random value drawn from the Half Cauchy distribution. To ensure that the transformed values are non-negative, we use the condition:

$$y_i^{\text{jittered}} = \max(y_i + \epsilon_i, 0)$$

This condition ensures that the jittered data remains non negative (as count data cannot be negative).

2.3. Half Cauchy Distribution

The half Cauchy distribution is the folded (truncated) form of the standard Cauchy distribution around the origin so that only positive (nonnegative) values are observed and detected. The half-Cauchy distribution is a heavy-tailed distribution characterized by its robustness against outliers, making it suitable for modeling data with extreme values.

The probability density function (PDF) of the half-Cauchy distribution is given by:

$$f(x{:}\,\sigma) = \frac{2}{\pi\sigma}\left[1 + \left(\frac{x-\theta}{\sigma}\right)^2\right]^{-1} ;x > 0$$

θ is the location parameter (center of the peak)

σ is the scale parameter (controls the width of the distribution)

when $\theta = 0$ and $\sigma = 1$, we have a standard Half Cauchy distribution as

$$f(x) = \frac{2}{\pi}\left(\frac{1}{1+x^2}\right)$$

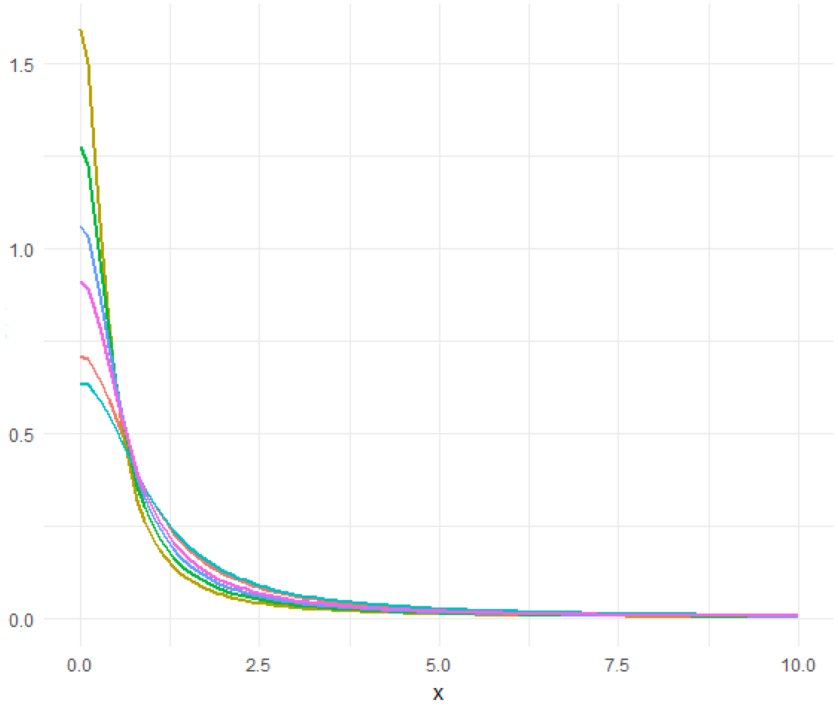

Figure 1. Half Cauchy Distribution Probability Density Function plot.

The cumulative distribution function (CDF) can be expressed as:

$$F(x) = \frac{2}{\pi}\arctan(x)\,, x \geq 0$$

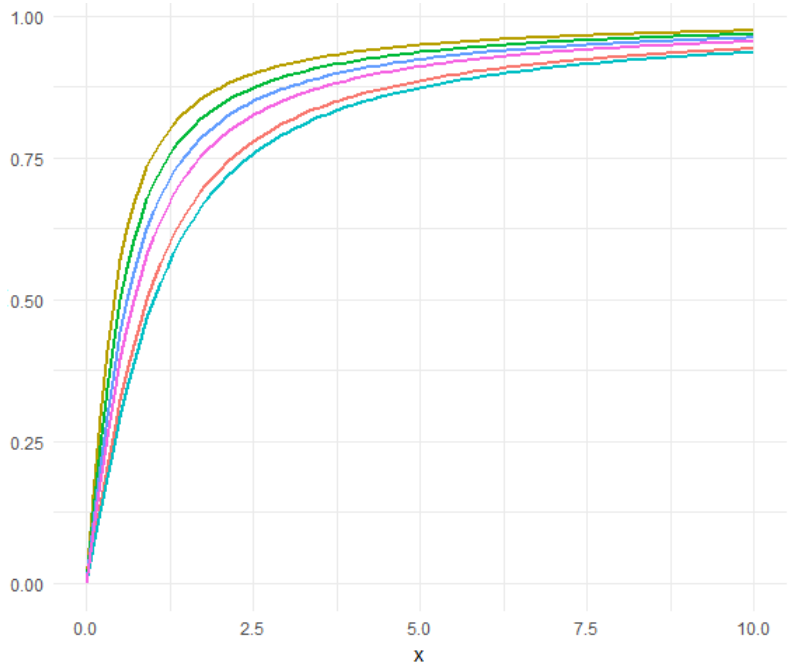

Figure 2. Half Cauchy Distribution Cumulative Density Function plot.

Inverse CDF (Quantile Function): The quantile function (i.e. Q(p)) is the inverse of the CDF, solving for x in terms of the cumulative probability $p = F(x)$, where $p \in [0,1]$.

$$Q(p) = \tan\left(\frac{\pi}{2}p\right), \ (0 \le p \le 1)$$

Since $Q(p) = \tan\left(\frac{\pi}{2}p\right)$ for $0 \le p \le 1$, then let τ be the quantiles in quantile regression, where $0 \le \tau \le 1$. We can now relate equation as:

$$Q(\tau) = \tan\left(\frac{\pi}{2}\tau\right) \ (0 \le \tau \le 1)$$

The proposed HCQR model extends traditional quantile regression by integrating the Half Cauchy distribution's quantile function. The mathematical formulation involves transforming the response variable using the quantile function of the Half Cauchy distribution, which allows for robust estimation of the conditional quantiles.

$$Q(\tau) = Q_\tau(Y|X)$$

$$\tan\left(\frac{\pi}{2}\tau\right) = \beta_0^{(\tau)} + \beta_1^{(\tau)}X_1 + \beta_2^{(\tau)}X_2 + \cdots + \beta_p^{(\tau)}X_p$$

to solve for τ, we apply the inverse tangent (arctangent) function to both sides of the equation to get:

$$\frac{\pi}{2}\tau = \arctan(\beta_0^{(\tau)} + \beta_1^{(\tau)}X_1 + \beta_2^{(\tau)}X_2 + \cdots + \beta_p^{(\tau)}X_p)$$

solve for τ by multiplying both sides of the equation by $\frac{2}{\pi}$ to get:

$$\tau = \frac{2}{\pi}\arctan(\beta_0^{(\tau)} + \beta_1^{(\tau)}X_1 + \beta_2^{(\tau)}X_2 + \cdots + \beta_p^{(\tau)}X_p)$$

recall that τ is the quantile range where $\tau \in [0,1]$ which essentially corresponds to the cumulative probability (quantiles) based on the predictors $X_1, X_2, \dots, X_p$ and their coefficient $\beta_0^{(\tau)}, \beta_1^{(\tau)}, \beta_2^{(\tau)}, \dots \beta_p^{(\tau)}$.

Since τ represents a quantile, it can be expressed as $Q(\tau)$, representing the quantile function. Therefore, we can rewrite the above equation 43 as:

$$Q(\tau) = \frac{2}{\pi}\arctan(\beta_0^{(\tau)} + \beta_1^{(\tau)}X_1 + \beta_2^{(\tau)}X_2 + \cdots + \beta_p^{(\tau)}X_p)$$

Now, the quantile function $Q_y(\tau)$ for the response variable Y similarly relates to the quantile function $Q(\tau)$.

The regression model can be expressed as follows:

$$Q_\tau(Y) = \frac{2}{\pi}\arctan(\beta_0^{(\tau)} + \beta_1^{(\tau)}X_1 + \beta_2^{(\tau)}X_2 + \cdots + \beta_p^{(\tau)}X_p)$$

$Q_\tau(Y)$ = the τ^{th} quantile of the response variable Y to be estimated.

$X_1, X_2, \dots, X_p$ = The predictor variables (independent variables) that influence the quantile of the response variable Y.

$\beta_0^{(\tau)}$ = The intercept (slope) term for the quantile regression model at quantile τ.

$\beta_1^{(\tau)}, \beta_2^{(\tau)}, \dots \beta_p^{(\tau)}$ = The regression coefficients for the predictor variables at quantile τ

τ = specified quantiles of the model (0.05, 0.25, 0.50, 0.75, 0.95)

The proposed half Cauchy quantile regression model will be applied on a jittered count (discrete) data as well as manage extreme values and outliers.

3. Results

A detailed comparison and discussion of the results obtained from the implementation of the Negative Binomial Regression (NBR) model and the proposed Half Cauchy Quantile Regression (HCQR) model was applied to a real life (crime data) where factors such as population density, income level, unemployment rate and police presence was used to investigate contributing factors to rise of crime rate. The above-mentioned models are evaluated based on results obtained on their model diagnostics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Skewness, Kurtosis, Standard Deviation, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Table 1. Model Comparison for NBR, CQR and proposed HCQR for the jittered data.

Model		MSE	RMSE	MAE	Skewness	Kurtosis	Standard Deviation	AIC	BIC
Negative Binomial Regression		104.837	10.239	5.755	8.220	75.270	10.293	3004.587	3190.536
Half Cauchy Quantile Regression	0.05	95.969	9.796	3.147	7.245	68.616	9.289	2973.108	2993.066
	0.25	96.044	9.800	3.190	7.218	68.284	9.283	2973.510	2993.467
	0.50	93.296	9.659	2.982	7.220	68.392	9.287	2960.729	2980.686
	0.75	91.823	9.582	2.862	7.248	68.688	9.286	2953.902	2973.859
	0.95	90.245	9.500	2.773	7.303	69.549	9.260	2946.694	2966.651

4. Discussion

The table above presents a comparison of the Negative binomial regression (NBR) and the proposed Half-Cauchy quantile regression (HCQR Each metric evaluates the model's performance with specific quantiles highlighted, leading to a holistic view of model accuracy, error distribution and fit quality. The MSE, RMSE and MAE values of the proposed model is lower than that of the NBR, suggesting it predicts with the least error, gives slightly more precise prediction under this model and performs well across different sections of the data compared to the Negative Binomial Regression model. Also, the skewness and kurtosis value for the proposed model is lower compared to the Negative Binomial Regression model. This suggests and reveals a more asymmetric error distribution of the data, leading to a more reliable prediction having heavy tails with fewer extreme values or outliers in its prediction. The proposed model has lower values in the standard deviation as compared with the other models considered in the study, indicating more closeness to the median (mean). Finally, the AIC and BIC values for the proposed model is also observed to be lower than the Negative Binomial Regression model. This indicates that it provides a better balance between model goodness of fit and performance. Across all metrics, the proposed Half Cauchy quantile regression model consistently outperforms the performance of the Negative Binomial Regression model, especially at higher quantiles which are crucial for understanding extremes values (outliers) in life data. This suggest that the Half Cauchy quantile regression model can be particularly effective in applications were predicting outliers (extreme value) accurately is crucial especially when the error distribution follows a non-normal distribution.

5. Conclusion

From the analysis conducted, the proposed HCQR model formulated with the Half Cauchy distribution, was found have performed well in handling count (discrete) and robust against extreme values (outliers). The proposed model's performance evaluations demonstrated the superior robustness of HCQR compared to Negative Binomial regression model. The performance metrics such as RMSE, MSE, MAE, AIC, BIC, Skewness, Kurtosis and standard deviation demonstrated the robustness and suitability of the proposed HCQR model in handling count (discrete) data. The proposed HCQR model outperformed the Negative Binomial Regression model in terms of robustness and goodness-of-fit, particularly in challenging datasets.

This study also demonstrates the potentials of the proposed HCQR in addressing critical gaps in statistically modeling count (discrete) data, laying a foundation for broader adoption in statistical and applied research.

6. Recommendations

Based on outcomes presented in the results obtained, we recommend that researchers working with count data are encouraged to consider the half Cauchy quantile regression model, especially in situations where traditional count regression models fail to account for outliers present in the count data. The proposed HCQR model's robustness makes it a strong alternative for handling real world datasets where heavy – tailed distributions are present.

Also, researchers are encouraged to refine and extend the framework, particularly by introducing Bayesian methods for parameter estimation. Expanding the model will make it more applicable to a wider range of real-world dataset.

Abbreviations

OLS	Ordinary Least Squares
NBR	Negative Binomial Regression
CQR	Cauchit Quantile Regression
HCQR	Half Cauchy Quantile Regression
MSE	Mean Square Error
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
AIC	Akaike Information Criteria
BIC	Bayesian Information Criterion

Acknowledgments

We express our profound gratitude to the referees, the associate editor, and the consulting editor for their invaluable contributions to this work. Their meticulous and insightful comments have significantly enhanced the quality of our manuscript. This research has been enriched by their rigorous reviews and thoughtful feedback, for which we are deeply appreciative.

Author Contributions

Runyi Emmanuel Francis: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Writing – original draft

Maureen Tobe Nwakuya: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Writing – original draft

Maxwell Azubuike Ijomah: Methodology, Supervision, Validation, Writing – review & editing

Funding

The authors wish to clarify that there was no external funding received for this research. All aspects of this study were conducted independently and without financial support from any grants or sponsorship entities.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Acquah, H. D. (2018). Comparing OLS and quantile regression estimation techniques for production analysis: An application to Ghanaian maize farms. *RJOAS*, 8(80), 388. https://doi.org/10.18551/rjoas.2018-08.52

[2] Allen, D. E., Gerrans, P., Powell, R., & Singh, A. K. (2009). Quantile regression: its application in investment analysis. *Jassa*, (4), 7-12.

[3] Congdon, P. (2017). Quantile regression for overdispersed count data: A hierarchical method. Journal of Statistical Distributions and Applications, 4(18). https://doi.org/10.1186/s40488-017-0073-4

[4] Francis, R. E., & Nwakuya, M. T. (2022). A comparative analysis of ordinary least squares and quantile regression estimation technique. *American Journal of Mathematical and Computer Modelling*, 7(4), 49-54. https://doi.org/10.11648/j.ajmcm.20220704.11

[5] Hao, L., & Naiman, D. Q. (2007). Quantile regression. SAGE Publications, Inc., https://doi.org/10.4135/9781412985550

[6] Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis,* 91(1), 74-89. https://CRAN.R-project.org/package=quantreg

[7] Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. Econometrica: *Journal of the Econometric Society*, 46(1), 33-50.

[8] Koenker, R., & Mizera, I. (2004). Penalized trigrams: Total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 66(1), 145-163.

[9] Korkmaz, M. Ç., Chesneau, C., & Korkmaz, Z. S. (2021). Transmuted unit Rayleigh quantile regression model: Alternative to beta and Kumaraswamy quantile regression models. *UPB Scientific Bulletin, Series A: Applied Mathematics and Physics* 83(3): 149-158.

[10] Leiva, V., Sánchez, L., Galea, M., & Saulo, H. (2020). Birnbaum-Saunders quantile regression and its diagnostics with application to economic data. *Applied Stochastic Models in Business and Industry*, 36(5), 956-973. https://doi.org/10.1002/asmb.2556

[11] Machado, J., & Santos Silva, J. (2005). Quantiles for counts. *Journal of the American Statistical Association,* 100(472), 1226-1237. https://www.jstor.org/stable/27590667

[12] Mazucheli, J., Leiva, V., Alves, B., & Menezes, A. (2021). A new quantile regression for modeling bounded data under a unit Birnbaum–Saunders distribution with applications in medicine and politics. *Symmetry*, *13*(4), 682. https://doi.org/10.3390/sym13040682

[13] Nwabueze, C. J., Onyegbuchulem, B. O, & Nwakuya, M. T. (2022). On the Equivariance of Location Reparameterization of Quantile Regression Model using Cauchy Transformation. *Mathematical Theory and Modeling*, 12(2), 2022.

[14] Nwabueze, J. C., Onyegbuchulem, B. O., Nwakuya, M. T., & Onyegbuchulem, C. A. (2021). A Cauchy transformation approach to the robustness of quantile regression model to outliers. Royal Statistical Society Nigeria Local Group Conference Proceedings, 2021.