Nigerian Journal of Scientific Research (NJSR)
ISSN-0794-0378

COMPARING GENERALIZED LEAST SQUARES - QUANTILE REGRESSION WITH ORDINARY LEAST SQUARES IN HANDLING BOTH HETEROSCEDASTICITY AND OUTLIERS SIMULTANEOUSLY

OLASUPO, A.O.*, EFUWAPE, B.T. AND TAIWO, A.I.
Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria

ABSTRACT

Modeling in the presence of outliers and heteroscedasticity is a substantial problem in the realm of robust statistical analysis, both of which are essential components for drawing valid conclusions. This study focuses on the challenge of resolving problems arising from homoscedasticity violations when outliers are present, as conventional regression models can become unstable in these situations. The goal of this study is to create a robust regression model that successfully tackles these issues by simultaneously assessing Generalized Least Squares (GLS) and Quantile Regression (QR). Real life dataset for some macroeconomic indicators were obtained from the website of the Central Bank of Nigeria for period, 1981 to 2020. The regression models were obtained using Statistical Package for Social Sciences, Number Cruncher Statistical Systems and R program. Poverty Reduction is the dependent variable, while Money Supply, Government Expenditure, Government Revenue, and Financial Inclusion are the independent variables. The estimator of GLS and QR was combined in other to have GLS-QR. The proposed method exhibits better resistance against outliers and heteroscedasticity. These strategies are useful in addressing complex data, as demonstrated by real- life applications. This research provides a helpful toolkit for practitioners and academics to model complex datasets more accurately and robustly.

Keywords: GLS-Quantile Regression, Heteroscedasticity, Least Squares, Outliers
***Correspondence:** ahmed.olasupo@oouagoiwoye.edu.ng, 08063421603

INTRODUCTION

One of the primary objectives of statistics is to determine the best way to transform a sample so that an outcome variable may be linked to a group of predictor variables [1]. According to Lakshimi [2], the most widely used method for modeling an outcome variable as a function of the predictor factors is linear regression. Regression analysis is a statistical method used in many domains, such as science, engineering, and management science, to fit a model to a set of data [3].

The estimator minimizes the sum of square distances between the regression surface and the observed data points [4]. The Gauss-Markov theorem, according to Phillips [5], states that ordinary least squares (OLS) is always the best linear unbiased estimator (BLUE). Assuming a normally distributed error ε with a mean of 0 and variance of $\sigma^2 I$, OLS is considered the uniformly smallest variance unbiased estimator. The assumption of homoscedasticity is one of the basic requirements for a regression model [6]. Once homoscedasticity is assumed, inference techniques like confidence intervals, prediction intervals, and hypothesis tests gain significant strength. However, the OLS parameter estimates and conclusions may be incorrect if the residual term ε is not normally distributed. When the outcome variable has a probability distribution with a variance that is functionally connected to the mean, this assumption is frequently violated [7]. This state is called heteroscedasticity, and according to Cattaneo *et al*. [8], it can also negatively impact coefficient estimates obtained with OLS. Real-life data frequently differs from the assumptions that researchers make, especially in practical applications like economics, engineering, science, and medical research. The precision and dependability of statistical analysis might be strongly impacted by this divergence.

The OLS estimator will continue to be impartial and stable in the face of heteroscedasticity. But the OLS's parameter covariance matrix (CM) would be the most negatively impacted by heteroscedasticity. Consequently, the diagonal matrix's elements that are used to calculate the regression coefficient's standard error start to exhibit bias, unreliability, and inconsistent behavior. Furthermore, the t-tests performed for individual coefficients typically show an excessive degree of conservatism or leniency, contingent on the type of heteroscedasticity and bias present in the

confidence intervals. Thus, according to Midi *et al*. [9], the OLS estimator is no longer BLUE.

An additional area that needs to be considered and addressed is the existence of an outlier in the data collection. Given that outliers occur in almost all datasets to some extent and that they remain one of the most significant problems in regression analysis, outliers have long been the focus of statistical research. Outliers can sometimes dramatically distort the findings of statistical research, while other times, their effects might not be as obvious. Any observation that deviates from the rest of the data is called an outlier, and 10% of data sets typically contain outliers [10]. A data set may contain outliers due to a variety of reasons, such as errors in measurement or data entry, inadvertently including observations from unrelated populations, or the occurrence of a believable event.

MATERIALS AND METHODS

Generalized least square

GLS is a generalization of OLS. It is often employed when the assumption of homoscedasticity is not met.

Linear regression is given as:

$$y_i = \beta_0 + \beta_i X_i + e_i \quad (1)$$

Equation (1) can be re written as:

$$y_i = \beta_0 X_{0i} + \beta_i X_i + e_i \quad (2)$$

Where $X_{0i} = 1$ for each i.

Assuming heteroscedastic variances σ_i^2 are known, divide equation (2) by σ_i

$$\frac{y_i}{\sigma_i} = \beta_0 \left(\frac{x_{0i}}{\sigma_i}\right) + \beta_1 \left(\frac{x_i}{\sigma_i}\right) + \frac{e_i}{\sigma_i} \quad (3)$$

Equation (3) is re written as:

$$y_i^* = \beta_0^* x_{0i}^* + \beta_1^* x_i^* + e_i^* \quad (4)$$

To obtain the GLS estimators minimize equation (4)

$$\sum e_i^{2*} = \sum (y_i^* - \beta_0^* x_{0i}^* - \beta_1^* x_i^*)^2 \quad (5)$$

Equation (5) is re written as:

$$\sum \left(\frac{e_i}{\sigma_i}\right)^2 = \sum \left[\left(\frac{y_i}{\sigma_i}\right) - \hat{\beta}_0^* \left(\frac{x_{0i}}{\sigma_i}\right) - \beta_1^* \left(\frac{x_i}{\sigma_i}\right)\right]^2 \quad (6)$$

$$\beta_1 = \frac{(\sum w_i)\sum(w_i x_i y_i) - (\sum w_i x_i)(\sum w_i y_i)}{(\sum w_i)(\sum w_i x_i^2) - \sum(w_i x_i)^2} \quad (7)$$

Where $w_i = \frac{1}{\sigma_i}$

Quantile regression

Quantile Regression is a statistical technique utilized to model how variables relate to one another, focusing on different quantiles of the variables including the median, 0.25- quantile point and the 0.75- quantile point.

$$y_t = x_t' \beta_q + e_t \quad (8)$$

where y_t is the dependent variable of the time series data, x_t' is the independent variable of the time series data, β_q is the vector of unknown parameter related to the qth quantile and e_t is the error term.

$$\sum q e_t + \sum (1 - q) e_t \quad (9)$$

Minimize equation (8)

$$e_t = y_t - x_t' \beta_q \quad (10)$$

$$\sum e_t^2 = \sum \left(y_t - x_t' \beta_q\right)^2 \quad (11)$$

$$\beta_q = \frac{\delta Q_q(y/x)}{\delta x_k} \quad (12)$$

Substituting equation (10) in to (9)

$$q \sum \left(y_t - x_t' \beta_q\right) + 1 - q\left(y_t - x_t' \beta_q\right) \quad (13)$$

Where $0 < q < 1$

Generalized least squares-quantile regression

GLS-QR addresses a variety of concerns that OLS has; for example, error terms are not constant over a distribution, which breaks the homoscedasticity axiom. Furthermore, when relying on the mean as a metric of position, information about a distribution's tails is lost.

Combining equation (7) and (12) gives β_{GLSQ_0} and β_{GLSQ_1}

The Proposed GLS-QR deterministic regression model can be written as:

$$\hat{Y} = \hat{\beta}_{GLSQ_0} + \hat{\beta}_{GLSQ_1} X_1 + \hat{\beta}_{GLSQ_2} X_2 + \hat{\beta}_{GLSQ_3} X_3 + \cdots + \hat{\beta}_{GLSQ\,p-1} X_{p-1} \quad (14)$$

RESULTS AND DISCUSSION

Real-life data on certain macroeconomic variables were taken from the Central Bank of Nigeria's (CBN) 1981–2020 Statistical Bulletin. The R software, IBM Statistical Packages for Social Sciences (SPSS), and Number Cruncher Statistical Systems (NCSS) were used to create the regression models. In this study, Poverty Reduction (Y), Money Supply ($\boldsymbol{X_1}$), Government Expenditure ($\boldsymbol{X_2}$), Government Revenue ($\boldsymbol{X_3}$), and Financial Inclusion ($\boldsymbol{X_4}$) were used for the analysis. A regression model was fitted to characterize the association between four independent variables and poverty reduction.

The findings are displayed in Table 1. In the fitted model, the equation is:

$Poverty\ Reduction = 34.1718 - 3.3642e^{-05}\ Money\ Supply - 0.0008\ Government\ Expenditure + 0.0003 Government\ Revenue + \ 3.1053\ Financial\ Inclusion$

Table 1: OLS model

Model		**Unstandardized Coefficients**		**Standardized Coefficients**	**T**	**Sig.**
		B	Std. Error	Beta		
	(Constant)	34.172	2.608		13.102	.000
	MS	-3.364E-5	.001	-.043	-.042	.967
	GEX	-.001	.003	-.282	-.238	.813
	GRV	.000	.001	.144	.381	.706
	FII	3.105	4.317	.119	.719	.477
a. Dependent Variable: PGR						

Table 2: GLS-quantile regression (tau = 0.25)

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
C	21.84383	11.44937	1.907864	0.0598
MS	0.135001	0.033788	3.995584	0.0001
GEX	-0.009672	0.001953	-4.951334	0.0000
GRV	0.058917	0.112842	0.522121	0.6029
FII	0.049349	0.172870	0.285470	0.7760
Pseudo R-squared	0.491096	Mean dependent var		29.09783
Adjusted R-squared	0.455174	S.D. dependent var		5.359719
S.E. of regression	3.711242	Objective		65.77578
Quantile dependent var	26.00000	Restr. Objective		129.2500
Sparsity	5.861363	Quasi-LR statistic		115.5121
Prob (Quasi-LR stat)	0.000000			

Table 3: GLS-quantile regression (tau = 0.50)

Method: Quantile Regression (Median)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	21.31144	13.77735	1.546846	0.1256
MS	0.091820	0.057207	1.605039	0.1122
GEX	-0.010907	0.002373	-4.595478	0.0000
GRV	0.148465	0.169558	0.875600	0.3837
FII	0.101303	0.215694	0.469661	0.6398
Pseudo R-squared	0.489947	Mean dependent var		29.09783
Adjusted R-squared	0.453943	S.D. dependent var		5.359719
S.E. of regression	3.123155	Objective		92.06464
Quantile dependent var	28.00000	Restr. Objective		180.5000
Sparsity	6.481211	Quasi-LR statistic		109.1591
Prob (Quasi-LR stat)	0.000000			

Table 4: GLS-quantile regression (tau = 0.75)

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
C	30.93398	18.11162	1.707963	0.0913
MS	0.036084	0.070567	0.511347	0.6104
GEX	-0.012347	0.002135	-5.783905	0.0000
GRV	0.073948	0.199665	0.370363	0.7120
FII	0.303778	0.281386	1.079575	0.2834
Pseudo R-squared	0.502751	Mean dependent var		29.09783
Adjusted R-squared	0.467651	S.D. dependent var		5.359719
S.E. of regression	3.556108	Objective		81.42460

Quantile dependent var	31.00000	Restr. Objective	163.7500
Sparsity	8.322224	Quasi-LR statistic	105.5172
Prob (Quasi-LR stat)	0.000000		

GLS-quantile regression

The relationship between the four independent factors and the decrease in poverty has been explained by a GLS-QR model. The distribution's 25th, 50th, and 75th percentiles are displayed, respectively, in Tables 2, 3, and 4. The fitted model contains the following equations:

25th Percentile GLS-QR Model

$$\text{Poverty reduction} = 21.84383 + 0.135001\ Money\ Supply - 0.009672\ Government\ Expenditure + 0.058917 Government\ Revenue + 0.049349\ Financial\ Inclusion$$

50th Percentile GLS-QR Model

$$\text{Poverty reduction} = 21.3114 + 0.0918\ Money\ Supply - 0.0109\ Government\ Expenditure + 0.1485 Government\ Revenue + 0.1013\ Financial\ Inclusion$$

75th Percentile GLS-QR Model

$$\text{Poverty reduction} = 30.9339 + 0.0361\ Money\ Supply - 0.0123\ Government\ Expenditure + 0.0739\ Government\ Revenue + 0.3038\ Financial\ Inclusion$$

Comparison of OLS and GLS-QR

Adjusted coefficient of determination of OLS is - 0.05 shows that the model does not fit the data well due to the presence of heteroscedasticity and outliers. Inappropriate model specification and outliers is affecting the result. The GLS-QR models shows a consistent performance across the 25th, median and 75th quantiles all with adjusted coefficient of determination values around 0.45 - 0.47.

Table 5: OLS and GLS-QR multiple regression results

	OLS		**GLS-QR (0.25)**		**GLS-QR (0.50)**		**GLS-QR (0.75)**	
Parameter	**Estimate**	**P-Value**	**Estimate**	**P-Value**	**Estimate**	**P-Value**	**Estimate**	**P-Value**
Constant	34.1718	<0.0001	21.84383	0.0598	21.31144	0.1256	30.93398	0.0913
MS	$-3.36423e^{-5}$	0.9670	0.135001	0.0001	0.091820	0.1122	0.036084	0.6104
GEX	−0.0007932	0.8129	-0.009672	0.0000	-0.010907	0.0000	-0.012347	0.0000
GRV	0.00028993	0.7059	0.058917	0.6029	0.148465	0.3837	0.073948	0.7120
FII	3.10528	0.4767	0.049349	0.7760	0.101303	0.6398	0.303778	0.2834
Important Regression Statistics								
	R^2	0.057178	R^2	0.49109	R^2	0.489947	R^2	0.502751
	Adjusted R^2	-0.050573	Adjusted R^2	0.45517	Adjusted R^2	0.453943	Adjusted R^2	0.467651

CONCLUSION

Robust statistical analysis requires tackling the problems caused by heteroscedasticity and outliers. The problems caused by homoscedasticity violations in the presence of outliers, which can make traditional regression models unstable, are the main emphasis of this paper. Generalized Least Squares and Quantile Regression, two simultaneous evaluation approaches, are used in this study in an effort to create a robust regression model. The study made use of actual macroeconomic data from the Central Bank of Nigeria, covering the years 1981 to 2020. The dependent variable was the poverty reduction, while the independent variables were the money supply, government expenditure, government revenue, and financial inclusion. In contrast, the GLS-QR model demonstrates higher coefficients of determination across different quantiles; specifically, for GLS-QR (0.25), GLS-QR (0.50), and GLS-QR (0.75), the coefficients of determination are 0.49, 0.50, and 0.57

for OLS, which shows that only a small portion of the variance in the response variable is explained by the explanatory variables. The reported results from comparing the GLS-QR and OLS models show significant differences in the adjusted coefficient of determination. GLS-QR model shows the efficacy in improving resistance against heteroscedasticity and outliers by integrating the estimators of GLS and QR. This study's approach provides insightful information about how to handle complexity in practical applications.

REFERENCES

1. GELMAN, A. & HENNIG, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society*, **180**(4): 967-1033.
2. LAKSHIMI, K. (2021). A study on mathematical and statistical aspects of linear models. *Turkish Journal of Computer and Mathematics Education*, **12**(4): 1328-1338.
3. ROBERT, M.B., BENJAMIN, S.D. & THOMAS, L.B. (1995). *Statistical methods for Engineers and Scientists*, 3rd ed. Marcel Dekker Inc., Newyork. 144pp.
4. RASHEED, B.A. & MALAYSIA, U. (2017). Linear Regression for Data having Multicollinearity, Heteroscedasticity and Outliers.
5. PHILLIPS, C. (2021). When is an Expectile the Best Linear Unbiased Estimator. *SSRN*. 87pp.
6. SCHMIDT, A.F. & FINAN, C. (2018). Linear Regression and Normality Assumption. *Journal of Clinical Epidemiology*, **98**: 146-151.
7. ERNST, A.F. & ALBERS, C.J. (2017). Regression assumptions in clinical psychology research practice- a systematic review of common misconception.
8. CATTANEO, M.D., JANSSON, M. & NEWEY, W.K. (2018). Inference in Linear Regression Model with many Covariates and Heteroscedasticity. *Journal of the American Statistical Association,* **113**(523): 1350-1361.
9. MIDI, H., RANA, S. & IMON, A. (2009). The performance of robust weighted least squares in the presence of outliers and heteroscedastic error. *WSEAS Transactions on Mathematics*, **7**(8): 351-361.
10. ZIMEK, A. & FILZMOSER, P. (2018). There and back again: Outliers detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **8**(6): 1-64.