Journal of the Nigerian Association of Mathematical Physics
Volume27, (*July, 2014*), pp 227 – 234

The Application of Quantile Regression Estimator as Alternative to Ordinary Least Squares

A. H. Bello

Department of Statistics, School of Sciences, Federal University of Technology, Akure – Ondo State. Nigeria

Abstract

This research work attempts to study the robustness of quantile regression as alternative to ordinary least squares method. Whereas the sum of squared errors is minimized in ordinary least squares regression, the median regression estimator minimized the sum of absolute errors, in quantile regression. The remaining conditional quantile functions are estimated by minimizing an asymmetrically weighted sum of absolute errors. A linear multiple regression model was fitted for both ordinary least squares and quantile regression, the results of the parameters estimated from analysis of economic growth rate of Gross Domestic Product (GDP) in Nigeria for thirty one year's which were computed with the use of Stata (12.0), MS Excel and SPSS were compared and it was found that the existence of outliers and non-normality assumption are violated when Ordinary least squares estimates(OLS) was used and the result can be misleading. Quantile regression gives true relationship between response variables and covariate for different subsections of the sample and therefore gives a better and efficient estimate of the relationship among random variables..

Keywords: Quantile Regression, Ordinary least squares, Outliers, Gross Domestic Product, Nigeria.

1.0 Introduction

In the past, only a little more than no person could with certain confident present or predict the future of any two or more related entities or variables e.g. tea and bread, investment and profit etc. Analyses were made basically with thoughts using the mind rather than with facts and proofs using the model. Today, several hundreds of individuals have in one time or the other visited the future of some variables with a known confidence with models. This model for transporting (visiting) the future is called Regression. Before the advent of regression, assumptions were made with ifs' today assurances are made with proof's which comprises of several assumption in a model.

In the classical methodology of ordinary least squares the conditional mean function, the function that describes the response variable y changes with covariates (x), is almost all we need to know about the relationship between the dependent variable and independent variables. For the ordinary least squares to give the best linear unbiased estimates, certain assumptions must exist:

1. $\in (e_i) = 0$
2. Homoscedasticity: $\mathcal{V}(e_i) = \sigma^2$(constant variance)
3. No autocorrelation $cov(e_i, e_j) = 0, \quad i \neq j$

The question is, if one or more of these assumptions are violated, what will become of the result obtained from OLS method, as it is interesting to know that some data from fields including biostatistics, econometrics, survival analysis etc. violate the assumption of homoscedasticity?

Corresponding author: A. H. Bello, E-mail:bellab_2011@yahoo.com Tel.: +2347033420672

"On the average" has never been a satisfactory statement with which to conclude a study on heterogeneous populations. Characterization of the conditional men constitutes only limited aspect possibly more extensive changes involving the entire distribution [1]

1.1 Aim And Objectives

The aim of the study is to compare Quantile Regression estimates with OLS and determine which one is better. The objectives are to:

1. Outlining the effects of outliers on regression model
2. Bringing to fruition the importance of employing quantile regression method in the analysis of data
3. Establishing that quantile regression is a good alternative to ordinary least squares.

1.2 Characteristics of Quantile Regression

The following characteristics feature differentiates quantile regression method from other regression methods.

1. The entire conditional distribution of the dependent variable y can be characterized through different values of $\tilde{\tau}$
2. Heteroscedascity can be detected
3. If the data is heteroscedastic, median regression estimate can be more efficient than mean regression estimates.
4. The minimization problem

$Min\beta\epsilon IR\Sigma\varepsilon\tilde{\tau}\left(Y_i - (x_i, \beta)\right)$can be solved efficiently by linear programming methods.

5. Quantile functions are also equivalent to monotone transformations. That is

$$Qn(Y/X)(x\tau) = h(Q(Y/X)(x\tau))$$

6. Quantiles are more robust with regards to outliers [2]

2.0 Literature Review

Ordinary Least Squares is claimed to produce parameters with desirable characteristics – best, linear, unbiased estimators (Blue). This means that parameters estimated using OLS have the smallest variance, model a linear relationship between response and outcome variables, and resemble value of parameters in population. However, these characteristics only hold if there are not serious violations of the model assumptions or presence of influential outliers [3 – 5]. Heteroscedasiticity will make parameter estimates not longer BLUE, while the presence of influential outliers will cause the regression line to be leveraged in the outliers direction. Furthermore, the presence of outliers also violates the one-model assumption that one regression line is sufficient to model the relationship between variables for the whole distribution.

OLS still has inherent disadvantages even when procedures to overcome effects of violation assumptions and outliers are applied. Models suggested by OLS cannot be immediately extended to other locations in the distribution that may be more interesting to be investigated in other studies. For example, the study of educational achievement focuses on over-achieving or under achieving students.

OLS also assumes that response variables only affect the location shift of the conditional distribution, while response variables may affect other parameters of the distribution in some instances. This means that OLS provides limited information about the relationship between variables [6 – 7]. OLS may give inaccurate information about the nature of the relationship between variables. When heteroscedasticity occurs and the slope of the regression line on the conditional mean is zero, OLS or related approaches to overcome heretroscedasticity will suggest no relationship between variables, although there are relationships between variables on non-central locations or on other distributional parameters

3.0 Methodology

3.1 Multiple Linear Regressions (OLS)

A typical equation of the general linear regression model can be written as:

$$Yi = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \ldots + \beta_K x_K + \mu \quad (1)$$

The general regression model $Y = X\beta + \mu$ can be written as

$e = Y - \hat{Y}$

$= Y - Xb$

Thus $e'e = (Y - Xb)'(Y - Xb)$

$e'e = (Y' - b'X')\ (Y - Xb) = S$

$S = Y'Y - Y'Xb - b'X'Y + b'X'Xb$

$= Y'Y - 2b'X'Y + b'X'Xb[since Y'Xb = b'X'Y]$
$\frac{\delta s}{\delta b} = -2X'Y + [X'Xb + b'X'X]$
$= -2X'Y + 2X'XB[since X'Xb = b'X'X]$
Is at minimum when
$\frac{\delta s}{\delta b} = -2X'Y + 2X'Xb = 0$
i.e. when $X'Xb = X'Y$
$\Rightarrow b = (X'X)^{-1}X'Y$
The estimate $b_0 of \beta_0$ can be calculated from the formula

$$b_0 = \frac{1}{T}[\Sigma Y - b_1 \Sigma X_1 - b_2 \Sigma X_2 \ldots - b_k \Sigma X_k]$$

3.2 Quantile Regression Method

Quantile Regression essentially transforms a conditional distribution function into a conditional quantile function by slicing it into segments. These segments describe the cumulative distribution of a conditional dependent variable Y given the explanatory variable x_i.
For a dependent variable Y given the explanatory variable $X = x$ and fixed τ, $0 < \tau < 1$, the conditional quantile function is defined as the $\tau - th$ quantile $Q_{Y|X}(\tau|x)$ of the conditional distribution function $F_{Y|X}(y|x)$. For the estimation of the location of the conditional distribution function, the conditional median $Q_{Y|X}(0,5|x)$ can be used as an alternative to the conditional mean [8 - 9].
One can nicely illustrate Quantile Regression when comparing it with OLS. In OLS, modeling a conditional distribution function of a random sample ($y_1, \ldots, y_n$) with a parametric function $\mu(x_i, \beta)$ where x_i represents the independent variables, β the corresponding estimates and μ the conditional mean, one gets following minimization problem:
$Min\beta \epsilon R \Sigma \varepsilon \tilde{\tau}(Y_i - \mu(x_i, \beta))^2$ (2)
One thereby obtains the conditional expectation function $E[Y|x_i]$. Now, in a similar fashion one can proceed in Quantile Regression. Central feature thereby becomes ρ_τ, which serves as a check function.
This check-function ensures that
1. All ρ_τ are positive
2. The scale is according to the probability τ

ere, as opposed to OLS, the minimization is done for each subsection defined by ρ_τ where the estimate of the τ^{th}-quantile function is achieved with the parametric function $\sum (Xi, \beta)$. In Quantile Regression the conditional function of $Q_{Y|X}(\tau|x)$ is segmented by the τ^{th}-quantile [10].
In the analysis, the τ^{th}-quantiles are:
$\tau \in \{0,05; 0,1; 0,25; 0,5; 0,75; 0,9; 0,95\}$
The τ sample quantile can be obtained by solving the following minimization problem

$$\hat{q}_\tau = \arg\min_{q \in R} \sum_{i=1}^{n} \rho_\tau (y_i - q), \quad (3)$$

4.0 Data Analysis and Results

In order to demonstrate the effect of outliers on a regression model and the analytical power of quantile regression, SPSS and Stata statistical software was used to analyze the growth rate of GDP in Nigeria for period of 31 years (1981-2011), using Total GDP as the dependent variable, and investigates the effects of the independent variables which were grouped into Eight (8): Agriculture, Crude Oil & Gas, Building & Construction, whole sales and retail trade, Telecommunication, Financial Institution, Education and health. A multiple linear regression has been assumed for simplicity.

Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \mu$$

(for OLS Estimate)

$$Y = \beta_{0\tau} + \beta_{1\tau} x_1 + \beta_{2\tau} x_2 + \beta_{3\tau} x_3 + \beta_{4\tau} x_4 + \beta_{5\tau} x_5 + \beta_{6\tau} x_6 + \beta_{7\tau} x_7 + \beta_{8\tau} x_8 + \mu$$

(forQuantile regression estimate)

Where τ = conditional quantile Function, Y = Dependent variable (Total GDP)

x_1= Agriculture, x_2 = Crude Oil & GAS, x_3 = Building &$Construction$,

x_4 = whole sales and retail trade, x_5 = Telecommunication, x_6 = Financial Institution, x_7 = Education, x_8 = Health.

β_0= intercept, β_i= Slopes

μ = Error term

Table 1: OLS Estimate for the original Data Computed via statistical software stata 12.1 and SPSS Data Analysis

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	R^2
53729.9	1.29799	1.00127	1.7773	1.97182	3.70704	-6.7724	-29.037	9.18595	1.000

From the test of outlier that was carried out it was observed that in the year 1991, 1992, 1996 -2011 are outliers as the probability of those years are greater than $\propto$ level (0.05).

4.1 Effect of Outliers on the Model

Table 2: Compared Estimate of the inherent (Original Data) and Outlying Observation removed (1991, 1992, 1993, 1996-2011) Computed via statistical software stata 12.1 and SPSS Data Analysis.

	Original Data	**Outlier removed**
β_0	0.7485	21.71467
β_1	0.4558	-0.2388839
β_2	0.2901	0.0678989
β_3	0.1147	-0.0781037
β_4	0.1304	0.9279655
β_5	-0.0099	0.2157055
β_6	-0.0264	0.594569
β_7	0.0364	-34.94631
β_8	0.0041	35.04013
R^2	0.9998	1.000

The result of the analysis shows that of original data and when outlier has been removed.

Conclusion: Observations in the year (1991, 1992, and 1996 -2011) are outliers but influential since R^2 does not reduce.

Table 3: Ordinary Least Square Estimates Models Computed via statistical software stata 12.1 and SPSS Data Analysis

Parameter	Estimate	P-Values
β_0	0.7485	0.000
β_1	0.4558	0.000
β_2	0.2901	0.000
β_3	0.1147	0.024
β_4	0.1304	0.113
β_5	-0.0099	0.599
β_6	-0.0264	0.518
β_7	0.0364	0.510
β_8	0.0041	0.851
R^2	0.9998	

Model: $Y = 0.7485 + 0.4558x_1 + 0.2901x_2 + 0.1147x_3 + 0.1304x_4 - 0.0099x_5 - 00264x_6 + 0.0364x_7 + 0.0041x_8$.

Test for individual Significance of the Model

Hypothesis Statement

H_0: $\beta_I = 0$

H_1: $\beta_I \neq 0$

Significant level $\propto = 0.05$

Critical region: Reject Ho if the significance probability i.e. P- values less than the $\propto$ level (0.05) otherwise do not reject.

Decision: Since P-values for $\beta_0, \beta_1, \beta_2, \beta_3$ (0.000, 0.000, 0.000, and 0.024) less than $\propto$ level (0.05), there is statistical reason to reject H_0. While for $\beta_4, \beta_5, \beta_6, \beta_7, \beta_8$, the P-values are (0.113, 0.599, 0.518, 0.510 and 0.851) is greater than $\propto$ level, there is no statistical reason to reject H_0.

Conclusion: Variables ($\beta_0, \beta_1, \beta_2, and \beta_3$) are significance to the growth rate of GDP in Nigeria, while variables ($\beta_4, \beta_5, \beta_6, \beta_7, \beta_8$) are insignifance or having little or no significance to the growth rate of Nigeria GDP. Hence, the new model is:

$Y = 0.7485 + 0.4558x_1 + 0.2901x_2 + 0.1147x_3$

Table 4: Quantile Regression Estimate on Data Computed Via statistical software Stata 12.1 and SPSS Data Analysis

τ	0.1	P-Values	0.3	P-Values	0.5	P-Values	0.7	P-values	0.9	P-values
β_0	18736.54	0.397	24599.19	0.010	24527.1	0.610	66896.97	0.134	68864.36	0.000
β_1	1.352366	0.000	1.352366	0.000	1.281864	0.000	1.243877	0.000	1.146241	0.000
β_2	1.009682	0.000	1.009118	0.000	1.026883	0.000	1.031432	0.000	1.072827	0.000
β_3	3.13522	0.000	3.302396	0.000	2.429544	0.452	1.127738	0.745	-0.927269	0.306
β_4	1.648825	0.000	1.613736	0.000	1.661437	0.052	1.987877	0.019	2.110061	0.000
β_5	4.190017	0.000	4.138692	0.000	2.870639	0.350	3.47866	0.389	2.992421	0.000
β_6	-6.070974	0.000	-5..37858	0.000	-4.615802	0.393	-6.491758	0.289	-5.840793	0.002
β_7	-32.63557	0.000	-31.65569	0.000	-23.13082	0.215	-21.68269	0.194	-7.002535	0.068
β_8	11.68743	0.000	11.507	0.000	8.865202	0.005	6.933774	0.008	2.897132	0.000
R^2	0.9950		0.99522		0.9956		0.9965		0.9976	

Model for $\tau = 0.1$

$$Y = 18736.54_{0.1} + 1.352366_{0.1}x_1 + 1.009682_{0.1}x_2 + 3.13522_{0.1}x_3 + 1.648825_{0.1}x_4 + 4.190017_{0.1}x_5 - 6.070974_{0.1}x_6 - 32.63557_{0.1}x_7 + 11.68743_{0.1}x_8.$$

Test for individual Significance of the Model at ($\tau = 0.3$)
Hypothesis Statement
H_0: $\beta_I = 0$
H_1: $\beta_I \neq 0$
Significant level $\propto$=0.05
Critical region: Reject H_0 if the significance probability i.e. P- values less than the $\propto$ level (0.05) otherwise do not reject.
Decision: Since P-values forβ_i less than $\propto$ level (0.05), there is statistical reason to reject Ho.
Conclusion: All the Variables are significance to the growth rate of GDP in Nigeria.

Table 5: Comparing OLS and Quantile Regression Standard Error of The Estimators

	OLS	**QUANTILE REGRESSION**				
Variables		**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
Agriculture	0.0340	0.0173	0.1	0.0889	0.0619	0.0123
Crude Oil & Gas	0.0135	0.0049	0.0033	0.0315	0.0248	0.0048
Building Construction	1.0107	0.6466	0.4823	3.1756	3.4183	0.8852
Wholesales & Retail	0.2298	0.1411	0.116	0.8087	0.7833	0.2419
Telecommunication	1.0323	0.6143	0.3853	3.0083	3.9605	0.5642
Financial Inst	1.6406	0.9226	0.6865	3.5005	5.9706	1.6186
Education	8.2253	3.668	2.8557	18.1365	16.192	3.6419
Health	2.0299	0.7005	0.3637	2.8162	2.396	0.4152
Constant	13335.38	21703.05	8579.921	47357.89	42991.91	11382.21

4.2 Results of the Analysis

Nigeria Gross Domestic Products Data (GDP) is not normally distributed which violated one of the assumptions of OLS and if an observation is an outlier but influential, such observation should not be removed before fitting the appropriate model.

Intuitively, Ordinary least squares indicates that the independent variable (Agriculture) has a positive influence on the dependent variable (Total GDP) with an estimate of β_1 =1.2980 This indicates that the growth rate of Total GDP increases as the growth rate of Agriculture increases. Quantile regression also confirms this and gives a better understanding of the influence of the underlying covariate. The influence is at 0.1th (1.352366) and 0.3th (1.347126) respectively but drops at 0.5^{th}, 0.7^{th} and 0.9^{th}, however seem to be stronger than is suggested by OLS. In testing for the significance of Agriculture to GDP OLS output shows that Agriculture is significant to the growth rate of GDP, quantile regression also supported the statement at all the quantiles conditions.

Considering the output of OLS on Crude Oil and Gas it suggested a positive influence on the growth rate of GDP with an estimate of β_2= (1.00127). Quantile Regression is also suggesting the same result at 0.1^{th} and 0.3^{th} .It increases at 0.9^{th} (1.072827). In all the quantiles conditions the independent variable (Crude oil) was significance to the model, OLS also showed that Crude was significant to the growth rate of the GDP.

Building and construction affects the Total GDP positively as suggested by OLS while at $0.9^{th\ (-0.9272695)}$Quantile Regression suggested a negative effects. However, the influence is stronger felt at 0.1, 0.3 and 0.7 quantiles with estimate β_3= (3.13522, 3.302396, 1.127738) respectively than is suggested by OLS estimate β_3= (1.777301). OLS reveals that Building and Construction was insignificance to the growth rate of GDP this cannot be generalized when compared to Quantile regression, it was significant at 0.1th, 0.30th but at Median i.e. $0.5^{th,}$ 0.9^{th} and 0.7^{th}quantile it was insignificant, hence policy recommendation based on OLS model may not be reliable depending on the quantile the policy will be based upon.

On Whole Sales and Retail Trade, OLS suggest a positive influence to the growth rate of GDP with an estimate β_4= 1.971819. Quantile Regression is also suggesting this. The influence is the same at 0.9^{th} and 0.7^{th}, then drops sharply to a much weaker influence at 0.1th and 0.3th with estimate β_4= (1.648825, 1.613736). The positive influence suggested by OLS was significant, whereas, quantile regression suggestion was significant at 0.1th, $0.3^{th\ and}$ 0.9^{th}quantiles while insignificant at 0.5th and 0.7^{th}quantiles respectively; the decision on its significance depends majorly on the quantile the researcher is considering.

According to OLS an increase in the growth rate of Telecommunication (GDP), will bring about increase in the growth rate of Total GDP in regarding the data. This is also being suggested by quantile regression. This shows that the independent variable, Telecommunication has a positive influence on the dependent variable Total GDP. However, in testing for the significance of the model on positive contribution of Telecommunication as suggested by OLSwas significant but quantile regression indicates that it was significant at 0.1th, 0.3th and 0.9^{th}quantiles respectively. This implies that, the general recommendation from OLS that if the growth rates of Telecommunication increases, the growth rate of Total GDP would increase cannot be generalized. Thus, a policy recommendation based on OLS estimate could therefore be grossly misleading.

In the same instance, one cannot generalize the result given by OLS estimate with respect to influence of Financial Institution on the Total GDP because of ambiguity. OLS suggest a negative influence with an estimate β_6= (-6.772368). Quantile regression suggested both positive and negative influence, significant and insignificant effects depend majorly on the quantiles one is interested in.

5.0 Conclusion

Quantile regression is offering a comprehensive strategy for completing the regression picture as it goes beyond this primary goal of determining only the conditional mean, and enables one to pose the question of relationship between the response variable and covariate at any quantile of the conditional distribution function. Quantile regression overcomes various problems that OLS is confronted with frequently; error terms are not constant across a distribution, thereby violating the axiom of homoscedasticity. Also, by focusing on the mean as a measure of location, information about the tails of a distribution is lost. And last but not least, OLS is sensitive to extreme outliers, which can distort the results significantly. As indicated in the data of Nigeria GDP growth rate, the major contributory sector to GDP growth rate in Nigeria is Agriculture and Oil and Gas sector of the Economy, sometimes a policy based upon an OLS analysis might not yield the desired result as a certain subsection of the years does not react as strongly to this recommendation or even worse, negative influence between the response variable and the explanatory variables was not indicated by OLS. Quantile regression is a robust alternative to OLS from the data analyzed.

References

[1] Buchinsky, M. (1994): Changes in the U.S wage structure 1963-1987: Application of quantile regression. Econometrica, 62, 405-405.

[2] Karlsson, A. (2006):"Estimation and inference for Quantile Regression of Longitudinal Data withApplications in Biostatistics".

[3] Berry, W. D. (1993). *Understanding regression assumptions.*Thousand Oaks, CA: Sage Publications Inc.

[4] Hao, L., and Naiman, D. Q. (2007). *Quantile Regression*. Thousand Oaks: Sage Publications Inc.

[5] Cade, B. S., and Noon, B. R. (2003): A gentle introduction to quantile regression for ecologists.Frontiers in Ecology and the Environment, 1, 412-420.

[6] Eide, Eric , Showalter and Mark H. (1999): Factors Affecting the Transmission of Earnings across Generations: A Quantile Regression Approach. The Journal of Human Resources. 34(2): 253 - 267.

[7] Chen, W. (2005): A reliability case on estimating extremely small percentiles of strength data for continuous improvement of medium density fiberboard product quality. M.S. Thesis. University of Tennessee. Knoxville, TN.

[8] Abrevaya, J. (2001): "The effects of demographics and maternal behavior on the distribution of birth outcomes", in Economic Application of Quantile Regression, New York.

[9] Gilchrist, W. (2000): Statistical modelling with quantile functions. Boca Raton, FL: Chapman and Hall/CRC.

[10] Moore, D. S. (2007). The basic practice of statistics(4th ed.). New York: W.H. Freeman and Co.