International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 11 Issue 01 January 2023, Page no. – 3220-3229

Index Copernicus ICV: 57.55, Impact Factor: 7.362

DOI: 10.47191/ijmcr/v11i1.15

IJMCR

International Journal Of Mathematics & Computer Research

VOLUME 11

YEAR 2023

Measuring the Impact of Tau vector on Parameter Estimates in the Presence of Heteroscedastic data in Quantile Regression Analysis

Ajao, I. O.[1], Obafemi, O. S.[2], Osunronbi, F.A.[3]

[1,2,3] Department of Mathematics and Statistics, The Federal Polytechnic, Ado-Ekiti, Ado-Ekiti, Nigeria.

ARTICLE INFO

Published Online:
31 January 2023

Corresponding Author:
Ajao, I. O.

ABSTRACT

The ordinary least squares (OLS) regression models only the conditional mean of the response and is computationally less expensive. Quantile regression on the other hand is more expensive and rigorous but capable of handling vectors of quantiles and outliers. Quantile regression does not assume a particular parametric distribution for the response, nor does it assume a constant variance for the response, unlike least squares regression. This paper examines the impact of various quantiles (tau vector) on the parameter estimates in the models generated by the quantile regression analysis. Two data sets, one with normal random error with non-constant variances and the other with a constant variance were simulated. It is observed that with heteroscedastic data the intercept estimate does not change much but the slopes steadily increase in the models as the quantile increase. Considering homoscedastic data, results reveal that most of the slope estimates fall within the OLS confidence interval bounds, only few quartiles are outside the upper bound of the OLS estimates. The hypothesis of quantile estimates equivalence is rejected, which shows that the OLS is not appropriate for heteroscedastic data, but the assumption is not rejected in the case of homoscedastic data at 5% level of significance, which clearly proved that the quantile regression is not necessary in a constant variance data. Using the following accuracy measures, mean absolute percentage error (MAPE), the median absolute deviation (MAD) and the mean squared deviation (MSD), the best model for the heteroscedastic data is obtained at the first quantile level (tau = 0.10).

KEYWORDS: quantiles, ordinary least square, estimates, heteroscedastic, accuracy measures.

INTRODUCTION

Standard linear regression techniques summarize the average relationship between a set of regressors and the outcome variable based on the conditional mean function E(y/x). This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of y. Quantile regression provides that capability.

Analogous to the conditional mean function of linear regression, we may consider the relationship between the regressors and outcome using the conditional median function Q_τ (y/x), where the median is the 50th percentile, or quantile τ , of the empirical distribution. The quantile $\tau \in$ (0, 1) is that y which splits the data into proportions τ below and 1 -τ above: $F(y_\tau) = \tau$ and $y_\tau = F^{-1}(\tau)$:for the median, τ = 0:5. The classic paper for quantile regression is Koenker and Bassett (1982). Koenker (2005) presents an extensive examination of the econometric theory related to a wide variety of quantile models. Some useful and accessible overviews of quantile regression analysis are presented in Buchinsky (1998) and Koenker and Hallock (2001). Buchinsky (1994, 1998) helped popularize the use of quantile regression analysis with highly influential papers on the distribution of wages. The approach has since been used quite extensively in labour economics. Some examples include Albrecht, Bjorklund, and Vroman (2003), Eide and Showalter (1999), Hartog, Pereira, and Vieira (2001), and Machado and Mata (2005). Examples from urban economics include Carillo and Yezer (2009), Chen, Kuan, and Lin (2007), Cobb-Clark and Sinning (2011), Craig and Pin (2001), Deng, McMillen, and Sing (2012), and Gyourko and Tracy (1999). The robustness of quantile regression makes it an attractive alternative for modeling the heavy-tailed behaviour of portfolio returns. Xiao, Guo, and Lam (2015) discuss an approach that uses an AR(1)–ARCH(7) quantile regression model for the return rate at time t. This paper however examines the impact of various quantiles (tau vector) on the parameter estimates in the models generated by the quantile regression analysis.

Methodology: Fitting Quantile Regression Models

The standard regression model for the average response is

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \qquad i = 1, \ldots, n$$

And the β_j's are estimated by solving the least squares minimization problem

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

In contrast, the regression model for quantile level τ of the response is

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \ldots + \beta_p(\tau)x_{ip}, \qquad i = 1, \ldots, n$$

And the $\beta_j(\tau)$'s are estimated by solving the least squares minimization problem

$$\min_{\beta_0(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau \left(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau) \right)$$

where $\rho_\tau(r) = \tau \max(r,0) + (1-\tau)\max(-r,0)$. The function $\rho_\tau(r)$ is referred to as the check loss, because its shape resembles a check mark.

For each quantile level τ, the solution to the minimization problem yields a distinct set of regression coefficients. Note that τ = 0.5 corresponds to median regression and $2\rho_{0.5}(r)$ is the absolute value function.

Simulation of data

The intuition behind quantile regression is easy to illustrate using a simple simulated data set. The raw data are shown in fig. 1. To make the graphs easier to read, the single explanatory variable, x, is limited to the set of integers from 1 to 100. Each integer occurs at least one time in the simulated data set, leading to 200 observations in total. Normal random error with non-constant variances were generated for the first case while a constant variance was used for the second case.

DATA AND DATA ANALYSIS

Heteroskedastic data: Quantile regression becomes more interesting when the errors are not homoskedastic. For this paper, the data used was obtained through simulation. All data simulation and analyses were done using *R* - 3.5.2 and Minitab 18.

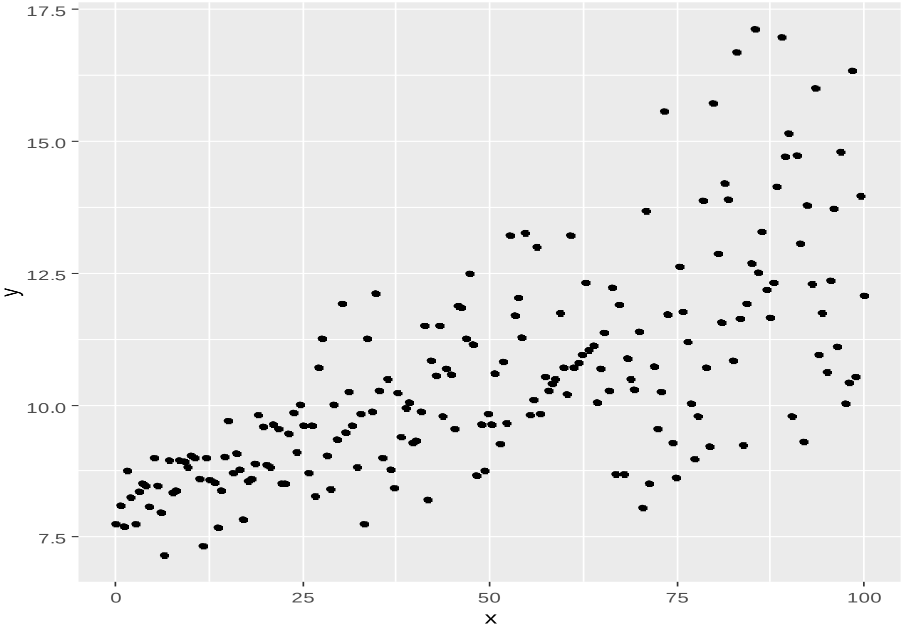

Fig.1: Simulated non-constant variance data set

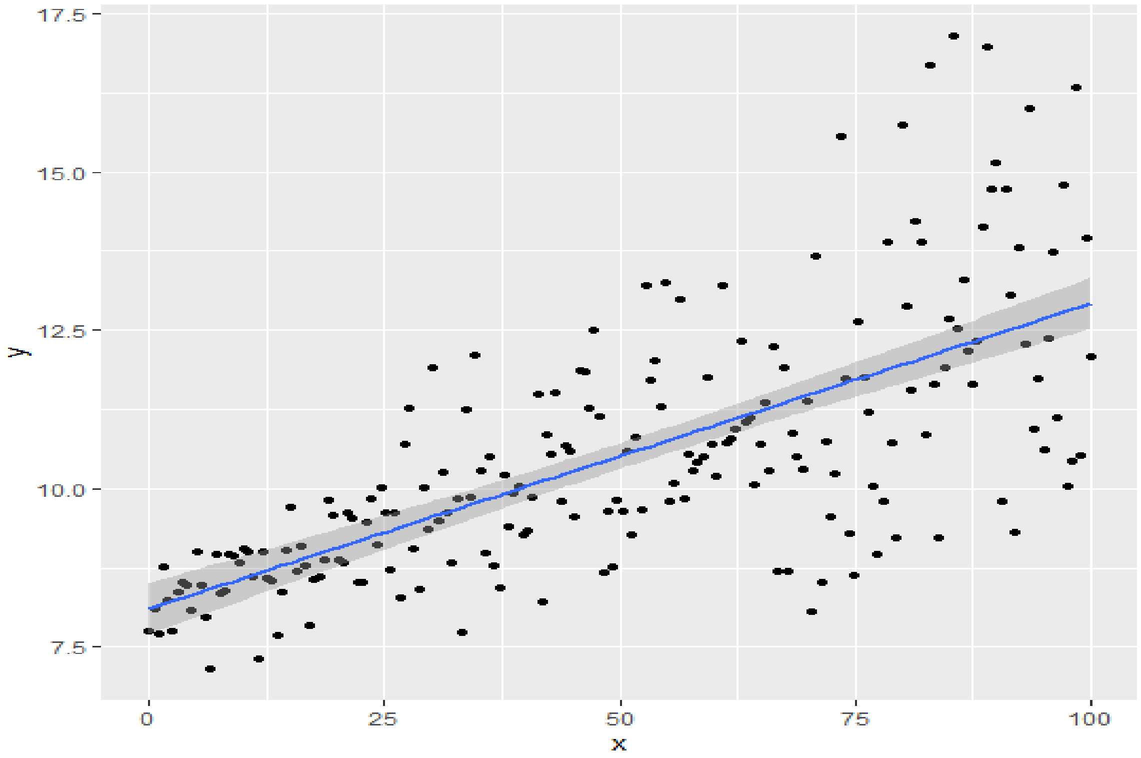

Fig.2: Plot of the fitted line with confidence interval for non-constant variance

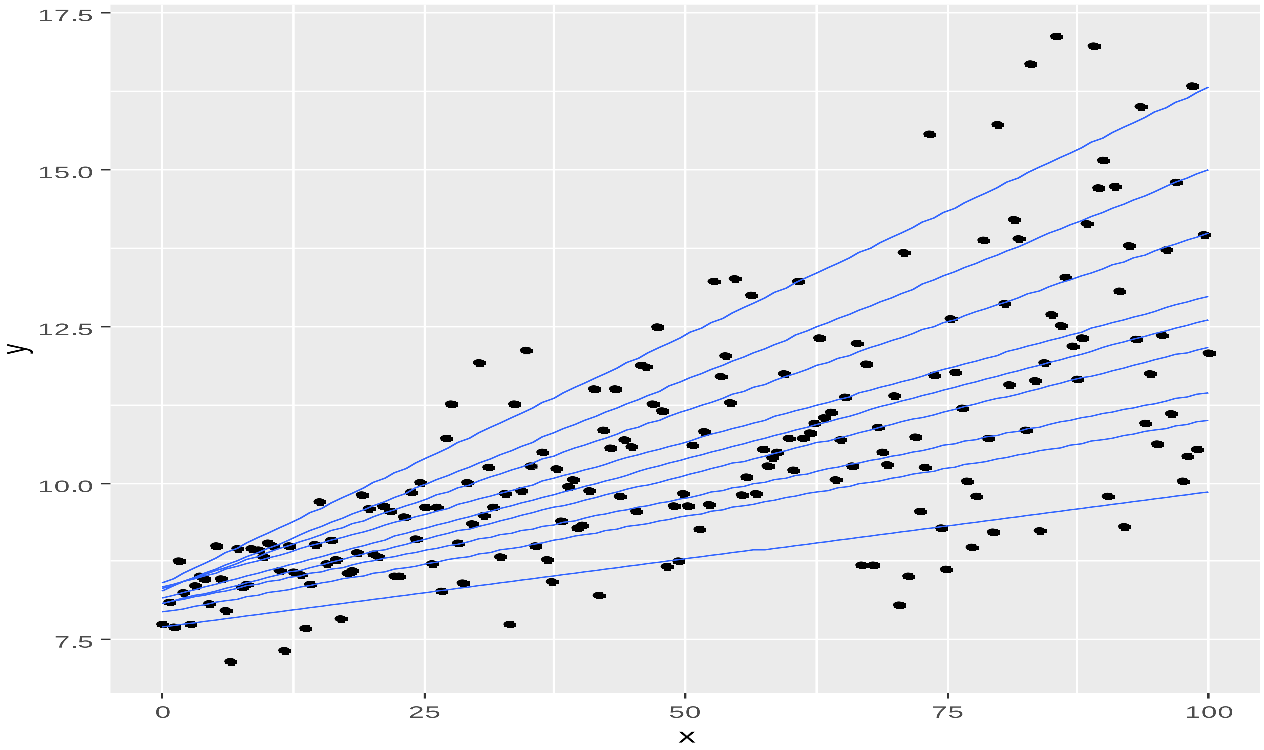

Fig.3: The intercept estimate and the steadily increasing slopes

The intercept estimate doesn’t change much but the slopes steadily increase.

Table 1: Summary of models with various quantiles (tau) in heteroskedastic data

tau	parameter	coef.	std. error	p-value	lower bound	upper bound
0.1						
	constant	7.7019	0.19382	0.0000	7.3337	7.8654
	slope	0.0216	0.00468	0.0000	0.0163	0.0313
0.2						
	constant	7.9378	0.15631	0.0000	7.7504	8.1423
	slope	0.0306	0.00479	0.0000	0.0234	0.0359
0.3						
	constant	8.0836	0.14606	0.0000	7.7713	8.1808
	slope	0.0336	0.00407	0.0000	0.0294	0.0419
0.4						
	constant	8.0777	0.11032	0.0000	7.9229	8.1741
	slope	0.0409	0.00333	0.0000	0.0369	0.0443
0.5						
	constant	8.1562	0.12948	0.0000	8.0138	8.3581
	slope	0.0445	0.00322	0.0000	0.0393	0.0493
0.6						
	constant	8.3376	0.15183	0.0000	8.0918	8.4967
	slope	0.0464	0.00546	0.0000	0.0425	0.0557
0.7						
	constant	8.3165	0.11039	0.0000	8.1475	8.4650
	slope	0.0567	0.00386	0.0000	0.0503	0.0611
0.8						
	constant	8.2872	0.11517	0.0000	8.1596	8.4228
	slope	0.0671	0.00591	0.0000	0.0580	0.0730
0.9						
	constant	8.4039	0.18047	0.0000	8.2278	8.6533
	slope	0.0791	0.00571	0.0000	0.0704	0.0846
OLS	R-square	0.4778				
	constant	8.1039	0.2060	0.0000	7.6977	8.5101
	slope	0.0482	0.0036	0.0000	0.0412	0.0552

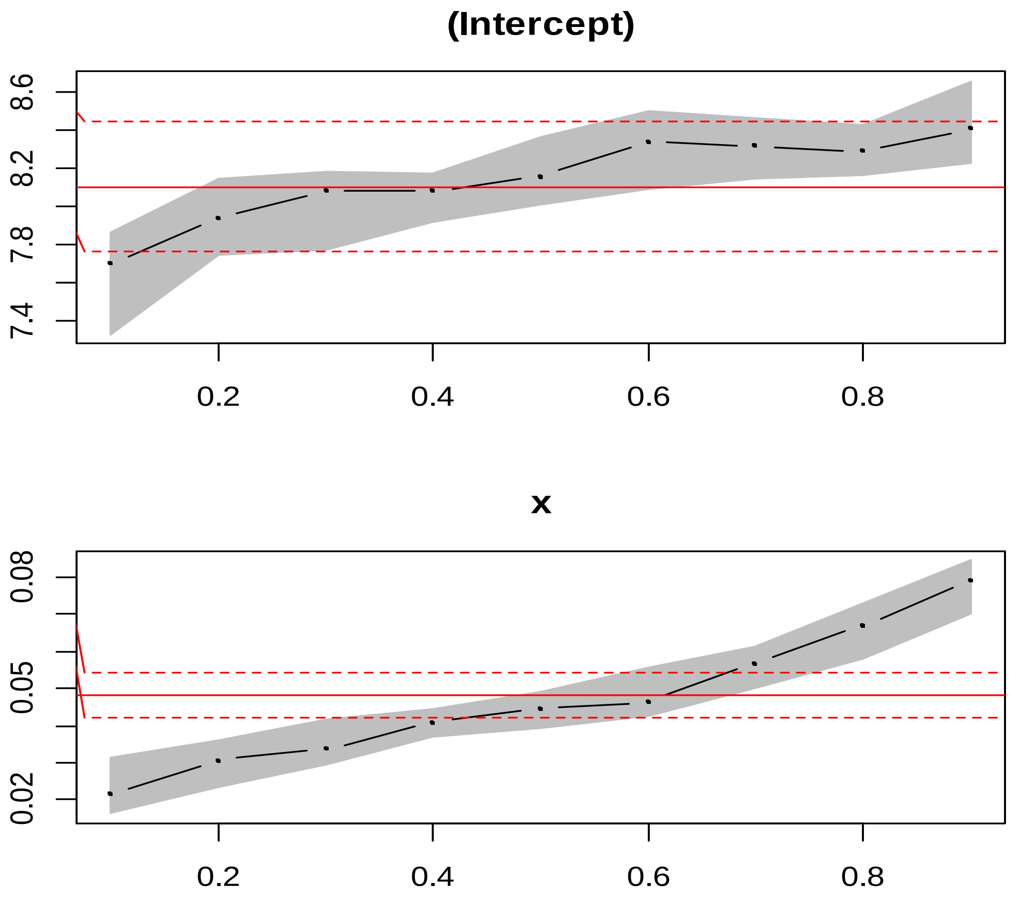

Fig.4: (a) OLS and quantile estimated intercepts for heteroskedastic data. (b) OLS and quantile estimated slopes for heteroskedastic data

The plots visualize the change in quantile coefficients along with confidence intervals. The intercept estimate does not change much but the slopes steadily increase. Each black dot is the slope coefficient for the quantile indicated on the x axis. The red lines are the least squares estimates and its confidence interval. It can be seen that the lower and upper quartiles are well beyond the least squares estimate. This shows that the OLS may not be appropriate for establishing the relationship between x and y.

Testing for the equivalence of quantile estimates for heteroscedastic data

We can also formally test the equivalence of the quantile estimates across quantiles, which allows us to estimate the model for each of several quantiles in a single model, allowing for cross-equation hypothesis tests.

Quantile Regression Analysis of Deviance Table

```
Model: y ~ x
Joint Test of Equality of Slopes: tau in {0.1 0.25 0.5 0.75 0.9}
  Df Resid Df  F-value   P-value
  4   9    96  12.0690   0.0000
```

The hypothesis of equality is obviously rejected of the estimated coefficients for the five quartiles in each case. This shows that the slopes in the quantile regression models are not the same, therefore the OLS is not appropriate for heteroscedastic data.

Table 2: Summary of models with various quantiles (tau) in homoskedastic data

tau	parameter	coef.	std. error	p-value	lower bound	upper bound
0.1						
	constant	7.5538	0.1326	0.0000	7.2945	7.7071
	slope	0.0474	0.0020	0.0000	0.0452	0.0518
0.2						
	constant	7.7999	0.0958	0.0000	7.6559	7.8737
	slope	0.0472	0.0016	0.0000	0.0455	0.0499
0.3						
	constant	7.8868	0.0957	0.0000	7.6872	8.0190
	slope	0.0475	0.0017	0.0000	0.0449	0.0505
0.4						
	constant	7.9960	0.0704	0.0000	7.8653	8.0856
	slope	0.0479	0.0014	0.0000	0.0454	0.0494
0.5						
	constant	8.0657	0.0960	0.0000	7.9824	8.2551
	slope	0.0484	0.0016	0.0000	0.0455	0.0500
0.6						
	constant	8.2533	0.0806	0.0000	8.0921	8.3388
	slope	0.0468	0.0017	0.0000	0.0459	0.0505
0.7						
	constant	8.3074	0.0681	0.0000	8.1943	8.4077
	slope	0.0489	0.0015	0.0000	0.0464	0.0510
0.8						
	constant	8.3743	0.0673	0.0000	8.2823	8.4783
	slope	0.0504	0.0017	0.0000	0.0475	0.0527
0.9						
	constant	8.5503	0.1388	0.0000	8.4309	8.7821
	slope	0.0517	0.0028	0.0000	0.0471	0.0552
OLS	R-square	0.9037				
	constant	8.0703	0.0655	0.0000	7.9411	8.1994
	slope	0.0490	0.0011	0.0000	0.0467	0.0512

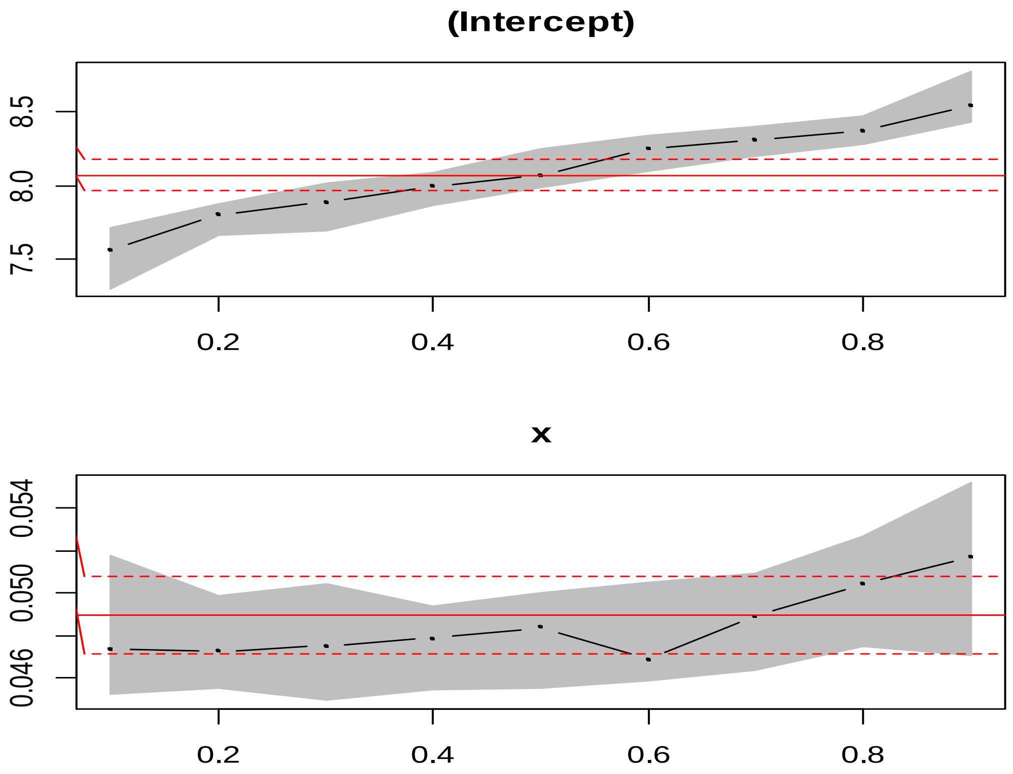

Fig.5: (a) OLS and quantile estimated intercepts for homoskedastic data. (b) OLS and quantile estimated slopes for homoskedastic data

The graph in fig. 5 illustrates how the effect of the predictor x, varies over quantiles, and how the magnitude of the effects at various quantiles differ considerably from the OLS coefficient, even in terms of the confidence intervals around each coefficient.

It can be deduced from table 2 that OLS is better for an homoskedastic data, most of the slope estimates fall within the OLS confidence interval bounds, only few quartiles are outside the upper bound of the OLS estimates.

Testing for the equivalence of quantile estimates for homoscedastic data

Formal test for the equality of the quantile estimates across quantiles for the several models arising from the tau vector is also performed for homoscedastic data.

Quantile Regression Analysis of Deviance Table

```
Model: y ~ x
Joint Test of Equality of Slopes: tau in {0.1 0.25 0.5 0.75 0.9}
 Df  Resid  Df  F-value  Pr(>F)
 4    9    96  0.7824   0.5367
```

The estimates clearly do not reject the hypothesis of equality of the estimated coefficients for the five quartiles in each case. Since there is no significant difference in the quantile slopes of the several models, it can be concluded that the OLS can be used for the homoscedastic data.

Detecting the best predicting quantile model in an heteroscedastic data

It is of necessity to detect the best predicting model out of the numerous quantile models. Using the simulated data, only five models were formulated, that is when tau = 0.1, 0.25, 0.50, 0.75, and 0.90, the last one is the normal OLS model. The graph below displays the various predicted values as generated from the models and predicted values were obtained from different quantile levels.

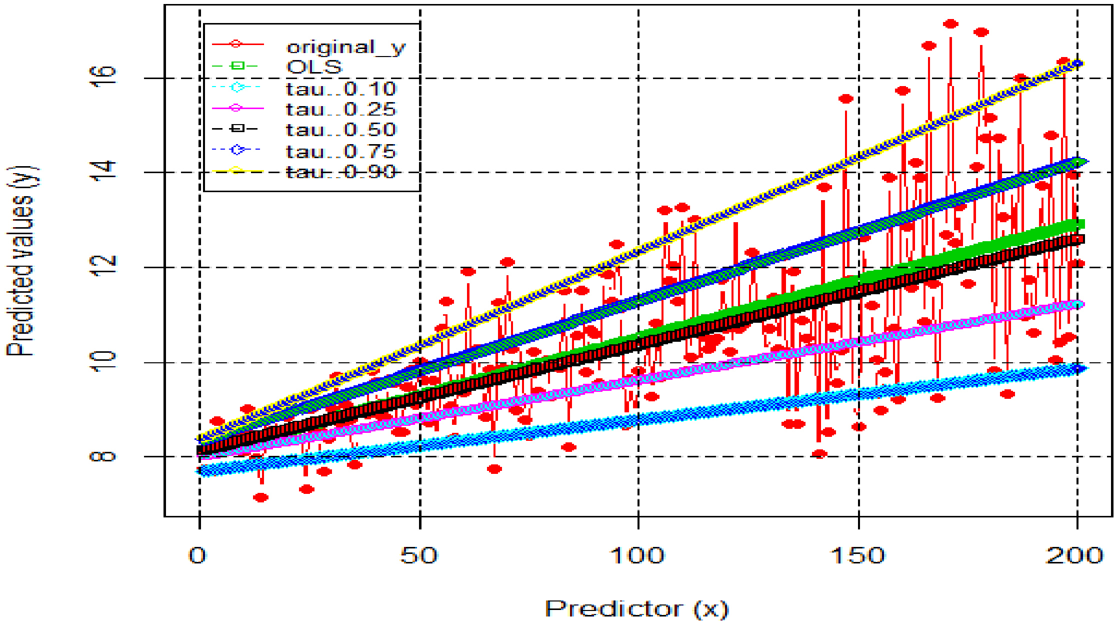

Fig. 6: Regression Models for Quantile Levels with heteroscedastic data

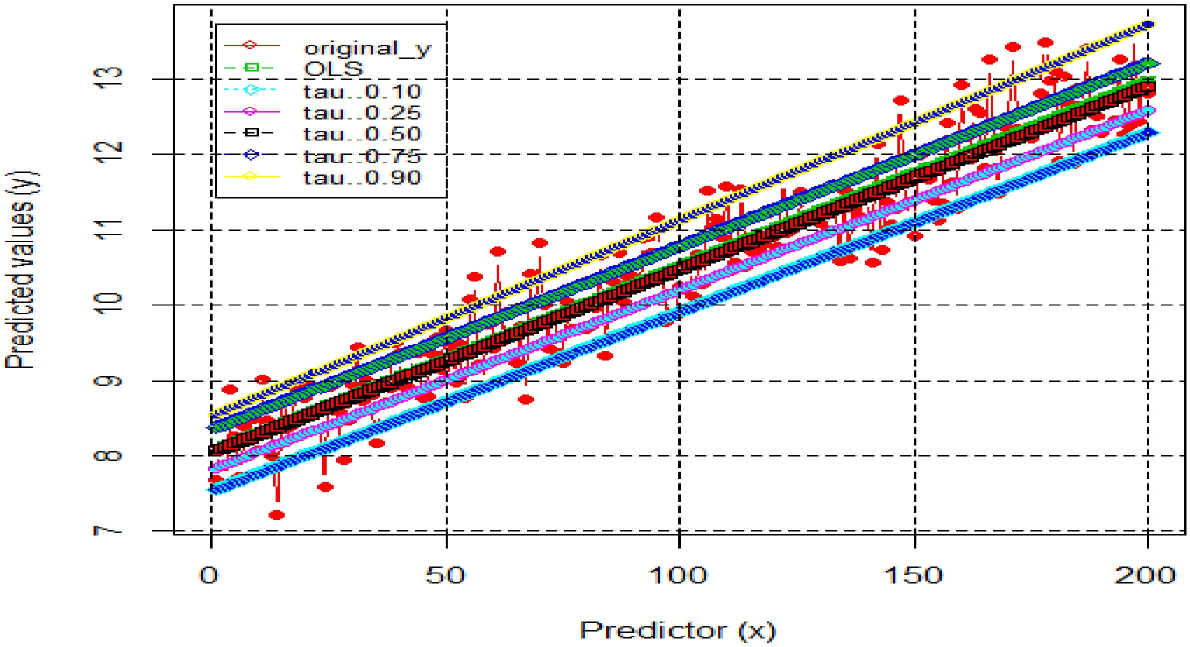

Fig. 7: Regression Models for Quantile Levels with homoscedastic data

In fig. 7, the median regression model (tau = 0.5) coincides with the OLS model. This means that either of them has the best fit for the homoscedastic data.

Using the measure of accuracy to determine the best model in a heteroscedastic data

The models obtained from the various quantile levels were used to make predictions. Metrics of Accuracy measures employed in this paper are the mean absolute percentage error (MAPE), the median absolute deviation (MAD) and the mean squared deviation (MSD) were then carried out on each predicted values. The model with the least accuracy measure values is the best.

Table 3: Summary of models with various quantiles (tau) in homoskedastic data

	OLS	tau = 0.10	tau = 0.25	tau = 0.50	tau = 0.75	tau = 0.90
MAPE	0.2340	0.1241	0.1687	0.2181	0.2659	0.3319
MAD	0.0242	0.0109	0.0161	0.0223	0.0295	0.0397
MSD	0.0006	0.0001	0.0003	0.0005	0.0009	0.0016

It is obvious from table 3 that the best model that establishes relationship between the response variable y and the predictor x and for a reliable prediction is when tau = 0.10, having the least values in the accuracy measures.

SUMMARY OF RESULTS

It can be seen from fig. 3 and table 1 and table 2 that in heteroscedastic data the intercept estimate does not change much but the slopes increase consistently in the models as the quantiles increase. On the other hand using homoscedastic data, results show that most of the slope estimates fall within the confidence interval bounds of the OLS, few quantiles are outside the upper bound of the estimates. The hypothesis that the quantile estimates are equal is rejected at 95% confidence level, which shows that the OLS is not appropriate for heteroscedastic data, however, the assumption is not rejected in the case of homoscedastic data at 95% confidence level, which shows that the quantile regression is not necessary in a constant variance data. The best model for the heteroscedastic data is obtained at the first quantile level (tau = 0.10), this is obtained using the accuracy measures: mean absolute percentage error (MAPE), the median absolute deviation (MAD) and the mean squared deviation (MSD).

CONCLUSION AND RECOMMENDATION

Quantile regression differs from conventional linear regression in its emphasis on issue related to the distribution of a dependent variable. The Monte Carlo study is a good representative of a situation in which OLS estimation can give misleading results. Since varying values of quantiles determine to a large extent the outcome of estimates, quantile regression is therefore recommended whenever non-constant variance is detected in data for reliable model and predictions, it is also robust to data having outliers. Although quantile regression methods are usually applied to continuous-response data, it is possible to utilize them in the context of count data, such as would appear in a Poisson or negative binomial model.

CONFLICT OF INTEREST STATEMENT

We declare that there are no conflicts of interest regarding the publication of this manuscript. The research was conducted independently and no funding was received for this study. None of the authors has any financial or personal relationships that could influence or appear to influence the content of the manuscript

REFERENCES

1. Albrecht J, Bjorklund A, Vroman S (2003). Is there a glass ceiling in Sweden? *J Labor Econ 21:145–177*
2. Buchinsky M (1994). Changes in U.S. Wage Structure 1963–1987: an application of quantile regression. *Econometrica 62:405–458*
3. Buchinsky M (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *J Human Resour 33:88–126*
4. Chen C-L, Kuan C-M, Lin C-C (2007). Saving and housing of Taiwanese households: new evidence from quantile regression analyses. *J Hous Econ 16:102–126*
5. Cobb-Clark D.A., Sinning M.G. (2011). Neighborhood diversity and the appreciation of native- and immigrant-owned homes. *Reg Sci Urban Econ 41:214–226*
6. Deng Y, McMillen D.P, Sing TF (2012). Private residential prices indices in Singapore: a matching approach. *Reg Sci Urban Econ 42:485–494*
7. Eide E. R., Showalter M. E. (1999). Factors affecting the transmission of earnings across generations: a quantile regression approach. *J Human Res 34:253–267*
8. Gyourko J, Joseph T (1999). A Look at Real Housing Prices and Incomes: some Implications for Housing Affordability and Quality. *Federal Res Bank of N Y Policy Rev 63–77*
9. Hartog J, Pereira P.T, Vieira J. A. C (2001). Changing returns to education in Portugal during the early 1990s: OLS and quantile regression estimators. *Appl Econ 33:1021–1037*
10. Koenker R, Hallock K. F (2001). Quantile regression. *J Econ Perspect 15:143–156*
11. Koenker, R. W., Bassett, G. W. (1982). Robust Tests for Heteroscedasticity based on Regression Quantiles, *Econometrica, 50, 43–61.*
12. Koenker, R. W. (2005). Quantile Regression, Cambridge U. Press.

13. Machado J. A. F., Mata J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *J Appl Econ 20:445–465*
14. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
15. Xiao, Z., Guo, H., and Lam, M. S. (2015). “Quantile Regression and Value at Risk.” In Handbook of Financial Econometrics and Statistics, edited by C.-F. Lee and J. Lee, 1143–1167. New York: Springer.