On the Asymptotic Behavior of the Estimates of Quantile Regression for Inflation Rates and Consumer Price Index

Carolina B. Baguio

Abstract

This paper investigates the asymptotic behavior of quantile regression which is a nonparametric procedure used for prediction. This method is more robust compared to the Ordinary Least Squares (OLS) for asymmetric data. This was the procedure used in the analysis of the annual data of Inflation Rates (IR) and Consumer Price Index (CPI) of the Philippines from 1958 to 2007. The package Quantreg of the free Statistical Software R was used in the computation of the estimates at different quantiles. The findings revealed that the data are asymmetric thereby resulting to non significant simple linear regression for both the Median and Least Squares Regression Procedures. However, at some quantiles at the left and right of the distribution, the regression models yield positive and negative significant linear regressions respectively. Simulation results reveal that in general, the inflation rates cannot predict the consumer price index by using quantile linear regressions or the least squares method. It is recommended to explore the nonlinear quantile regression procedures in modelling the inflation rates. Moreover, a study on the similarity and differences of Quantile

The author is a faculty member of the Mathematics & Statistics Department of the College of Science and Mathematics, at MSU-IIT with a rank of Professor V. She obtained her doctorate degree in Statistics/Applied Mathematics at the University of the Philippines in Los Baños.

Regression Procedures and Least Trimmed Squares (LTS) be investigated for Nonlinear Regression.

Keywords: quantile regression , asymptotic behavior, robust, asymmetric distribution, nonparametric procedure, ordinary least squares, least trimmed squares, non linear regression.

Introduction

It has been observed that most of the studies in various fields deal with the investigation of causal relationships of realistic phenomena of response and predictor variables. Most often linear regression analysis using least squares is used by researchers without the knowledge of a robust regression. This method is being driven with several assumptions like normally distributed errors, independence of observations, linearity, and absence of outliers. It was proven by so many studies that ordinary least squares is not robust compared to other methods such as the Median Regression, Least Trimmed Squares, and the Least Median Squares (Yaffee, R.A., 2002). Most of these methods used location estimators resulting to what we call L-estimates like the mean, median, quantile, trimmed and winsorized means.

The Least Absolute Deviation (LAD) was proven to be better than the Ordinary Least Squares (OLS) as revealed in many studies which resulted into the Median Regression for the 50th quantile or in general for any quantile, the so called Quantile Regression. In STATA 9.1, this method is classified under the nonparametric statistics since it is not assumed that data follow a certain kind of distribution. Since this method traces all parts of the distribution of the data for different quantiles, it does not matter whether the data is symmetric, highly skewed, or having long tails. The usual approach in dealing with realistic data is by the use of Exploratory Data Analysis (EDA) in order to detect he presence of outliers or bad data as shown by the graphs of the data.

Quantile Regression was never popular several decades ago and is not being taught in basic Statistics courses due to its inherent difficulty in the computation. However, several researchers found out that the (LAD) function is not differentiable unlike the OLS, so they made use of the Linear programming method. The computation of which is greatly facilitated nowadays since the advent of better solution methods and advances in computing. The curse of multidimensionality or parsimony which arise if there are so many data points already addressed by some advanced algorithms other than the simplex method of Linear programming.

In Statistics, there are two areas for consideration: the estimation and the hypothesis testing. The estimation of the coefficients can be done by using the simplex pivot of the primal problem of Linear programming, while the test on the significance of these estimates can be pursued using the solutions of the dual problem by the duality theory using rank scores wherein the confidence intervals of the estimates can be solved. The computations can be facilitated using the R language of computer programming.

In this study, attempts will be made to investigate the applicability of this method to the annual data on Inflation Rates (IR) and Consumer Price Index (CPI) of the Philippines from 1958 to 2007.

Objectives of the Study

1. To elucidate the concepts and procedures of quantile regression.
2. To compare Quantile Regression with Ordinary Least Squares (OLS) using the Inflation Rates (IR) as the response variable and Consumer Price Index (CPI) as the independent or causal variable.
3. To investigate the asymptotic behaviour of the regression estimates using Monte Carlo Simulation.

Theory and Concepts

3.1 Quantile Regression

Quantile regression is just a type of regression analysis in statistics. The basic difference between this method and that of the least squares is on the resulting estimates of the response variables given certain values of the predictors. Quantile regression method results to the estimates approximating the median or other quantiles of interest of the response variables whereas for the least squares, the resulting estimates approximate the conditional mean of the response variables.

Quantile regression is used when an estimate of the various quantiles (such as the median) of a population is desired. One advantage of using quantile regression to estimate the median, rather than ordinary least squares regression to estimate the mean is that quantile regression will be more robust in response to large outliers.

The topic Quantile regression is classified in the nonparametric statistics area since it is assumed that the distribution of the data is not known as well as other criteria like independence and linearity of the variables. It was shown that this method is robust to some classical assumptions of regression. In fact, using the criteria of efficiency and high breakdown point the Median regression when $\tau = .50$ is more resistant to the presence of outliers than the Ordinary Least Squares (OLS) method.

Prior the introduction of Quantile Regression, let us first understand what is the meaning of quantiles. Quantile refers to the location measure of ordered data. The student scores at the τ th quantile in licensure exam in nursing if he performs better than the proportion τ, and worse than the proportion $(1- \tau)$, of the reference group of students. Thus, in the case if $\tau =.50$ half of the students perform better than the median student, and half perform worse. Similarly, the quartiles divide the population into four segments with equal proportions of the population in each segment, the deciles into 10 equal segments and percentiles into 100 segments.

Let Y be any real valued random variable characterized by its distribution function as,

$$F(y) = Prob(Y \leq y)$$

The τ th quantile of Y for any $0 < \tau < 1$, on the other hand, can be written as

$$Q(\tau) = inf\{y|F(y) \geq \tau\} \qquad (1)$$

The quantile function in (1) provides a complete characterization of the random variable just like the distribution function. Using this function the median can be written as Q(1/2).

The quantiles defined above can be formulated as the solution to a simple optimization problem. For any $0 < \tau < 1$, define the piecewise linear function,

$\rho_\tau(v) = v(\tau\text{-}I(v<0))$. Minimizing the expectation of $\rho_\tau(Y - \theta)$ with respect to θ, yields solutions in which the smallest is the Q(τ) defined in (1).

The sample analog q(τ) of Q(τ), based on a random sample, $(y_1,y_2,\ldots,y_n)$ of Y's, is called the τ th sample quantile, which can be found by solving

$$\min_{\theta \epsilon R} \sum_{i=1}^{n} \rho_\tau (y_i - \theta).$$
(2)

Equation in (2) yields a natural generalization to the regression context. The linear conditional quantile function can be estimated by solving, for X ε R^K and β ε R^K:

$$\min_{\beta \epsilon R^K} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^T \beta)$$
(3)

3.2 Linear Programming Model

3.2.1. Estimation of Regression Quantiles Using the Primal Problem

Let us assume that we have for the conditional quantile function a linear statistical model in the form

$$q(X, \beta) = X^T \beta \qquad (4)$$

In this case, the optimization may be formulated as the usual linear programming problem for the quantile regression problem as follow:

We can proceed by writing Y with only positive terms as:

$$Y_n = X_n^T + u_n = \sum_{k=1}^{K} x_{n\_k} \beta_k + u_n$$

$$= \sum_{k=1}^{K} x_{n,k} (\beta_n^{+} - \beta_n^{-}) + (u_n^{+} - u_{+}^{-}) \qquad (5)$$

The optimization problem in equation (3) can be written in the form

Primal problem:

$$\begin{aligned} &\arg\min \; c^{T}\theta \\ &\text{subject to} \quad A\theta = y \\ &\qquad\qquad\qquad \theta \geq \mathbf{0}. \end{aligned} \qquad (6)$$

where $X = (X_1,\ldots,X_N)^T$

$A = (X, -X, I_N, -I_N)$

$y = (Y_1,\ldots,Y_N)^T$

0 is the zero vector **I** is the identity matrix

$$\theta = \begin{pmatrix} \beta^{+} \\ \beta^{-} \\ u^{+} \\ u^{-} \end{pmatrix} \qquad c = \begin{pmatrix} 0 \\ 0 \\ \tau I \\ (1 - \tau I) \end{pmatrix}$$

The corresponding dual problem to this primal problem is shown below:

Dual Problem :

$$\begin{aligned} &\arg\max \quad y^{T}\,\theta \\ &\text{subject to} \quad A^{T}\,\theta = c, \\ &\qquad\qquad\qquad \theta \geq 0 \end{aligned} \qquad (7)$$

3.2.2 Computation of Quantile Regression Estimates

The package Quantreg procedure using the R computer programming language solves the Linear programs in (6) and (7) using the simplex algorithm of Barrodale and Roberts (1973).The algorithm solves the linear program by two stages. The firs stage picks the X or –X as pivotal columns. The second stage interchanges the columns in I and –I as basis or nonbasis columns. The optimal solution can be obtained by executing the two stages interactively. Only the main data matrix is stored in the current memory because of the special structure of the matrix A. This special version of the simplex algorithm for median regression can be extended to quantile regression for any given quantile, even for the entire quantile process (Koenker and d'Orey 1993). It was found out that this procedure reduces greatly the computing time required by a general simplex algorithm, and suitable for data sets with less than 5,000 observations and 50 variables.

3.2.3. Hypothesis Testing Using the Dual

As mentioned in the previous section quantile estimation can be represented as a linear program. The confidence intervals of the coefficient estimates can be constructed by the solution of dual problem in quantile regression. It involves the rank statistics due to the dual problem of quantile regression. The advantage of rank test is that the nuisance parameter estimation can be avoided and the result is robust. The testing of the hypothesis for linearity of the quantiles in quantile regression can be done by the Rank–inverse test. Gutenbrunner and Jureckova (1992) showed that the solutions of the dual problem which is formulated for computing regression quantiles generalize the duality of ranks and quantiles to linear models. The dual solution called regression rank –score process establishes the link between the linear rank statistics and regression quantiles. The procedure emanates from the classical theory of rank tests which can be extended to the test of the hypothesis.

By inverting this test, confidence intervals for the regression quantile estimates of β_2 can be computed.

Ho: $\beta_2 = \xi$ in the linear regression model $y = x_1\beta_1 + x_2\beta_2 + u$
where y is the vector of response variable

and x_1 and x_2 are the vectors of the predictor variables, and u the error term.

Here, $X= (x_1,x_2)$ in equation (6). By inverting this test, the Confidence Intervals can be computed.

The rank score function a_T can be solved from the dual

$\max \{ (y-X_2\xi)'a/X_1'a=(1-\tau)X_1'1, a \, \varepsilon \, [0,1]^n\}$ where 1 is a vector of ones.

4. Empirical Results

Using the annual average data for the Philippines Inflation Rates (IR) and Consumer Price Index (CPI) taken from the National Statistics Publication in the website http://www.census.gov.ph/, appropriate regression procedure in the analysis was employed. Using Exploratory Data Analysis (EDA) Procedure the graphs and relevant statistical measures are presented for these data. Table 1 gives the Summary Statistics of the Inflation Rates and Consumer Price Index. It can be observed from the table that the mean and the median are not equal which are 9.446 and 7.20 respectively for the inflation rates which is an indication of asymmetry . Similar situation occurs for the consumer indexes where the mean is very much larger than the median. Figures 1-2 give the graphical displays of the histograms of these two variables. It is apparent from the two histograms that the two variables are not normally distributed. The Ordinary Least Squares Method (OLS) requires the assumption of normality on the data and using this will be detrimental to the desired inference.

Table1. Descriptive Statistics for Inflation Rates and Consumer Price Index (1958-2007)

Descriptive Statistics	Inflation Rates	Consumer Price Index
Minimum	-0.50	1.86
1st Quartile	4	3.27
Median	7.20	35.50
Mean	9.446	40.16
3rd Quartile	11.725	72.05
Maximum	50	141.80
Variance	74.68	1972.89
Standard Deviatio	8.642	44.42

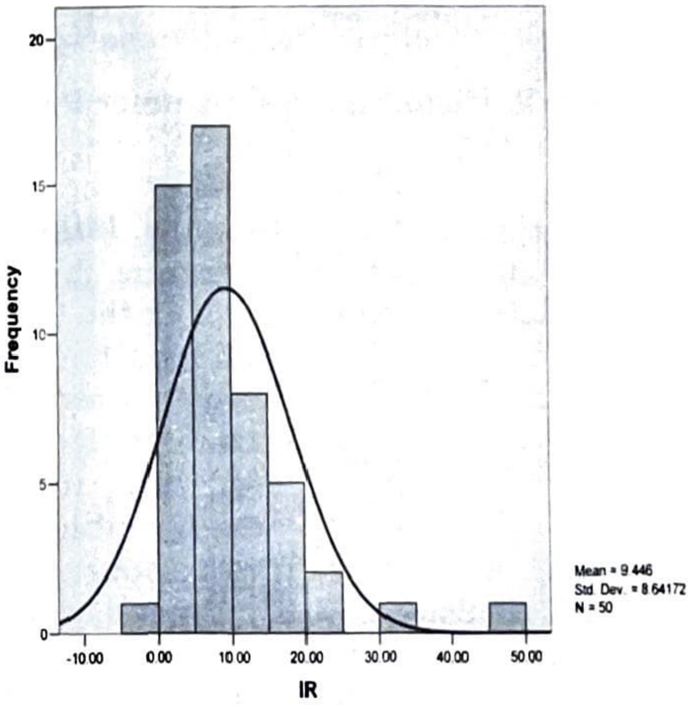

Figure 1. Histogram of Inflation Rates

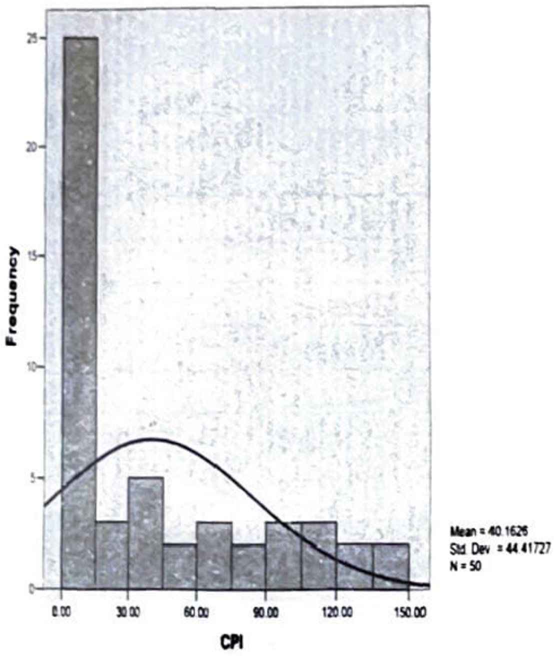

Figure 2. Histogram of Consumer Price Index

Figure 5 gives the scatter plot of the Inflation Rates and the Consumer Price Index. It can be observed from this plot that there are some outliers in the Inflation rates as well as in the Consumer price Index which can adversely affect the regression model. If a regression line is fitted to this plot most likely, the line moves downward as the Consumer Price Index increases. The maximum value of 50 for the inflation rate and the minimum negative value of -.50 can cause a non significant model for both the Median Regression , the Quantile regression when τ = .50 and using the OLS. However, using various quantiles when the values of tau are .05,.10,.25,.75,.90,.95, the relationship of these two variables can give significant regression models with some data to be excluded in the analysis. Table 2 gives the values of the coefficients and intercepts of the simple linear regression as compared to the OLS and Median Regression. The latter two methods both yield a non significant linear regression if all the data are included as indicated by the probability significance greater than .05 under the columns for τ = .50 and the OLS. It can also be observed that the coefficient estimates of the

5%,10% and 25% quantiles are all positive and significant as contrasted to those of 75%, 90% and 95% quantiles which are all negative but significant. This interesting results simply tell us that the left side of the distribution, the Inflation Rates increase as Consumer Price Indices increase whereas the opposite thing happens in the right hand side of the distribution as vividly portrayed in the scatter plot in Figure 5. On the other hand, Figure 6 displays all the regression lines. The three lines below the median regression represent the 5%,10% and 25% quantile regression lines with positive slopes. Whereas, the three black lines above the median and least squares regression lines represent the 75%,90% and 95% quantile regression lines with negative slopes. It can be said that for lower and higher values of the Consumer Price Index, the Inflation Rates increase and decrease respectively. From this result, it can be inferred that for data which violate the assumption of normality, the OLS and Median Regression yield non significant linear Regression models. Hence, by considering the distribution of the data by quantiles the relationship of the response and independent variables can be examined in all parts of the distribution.

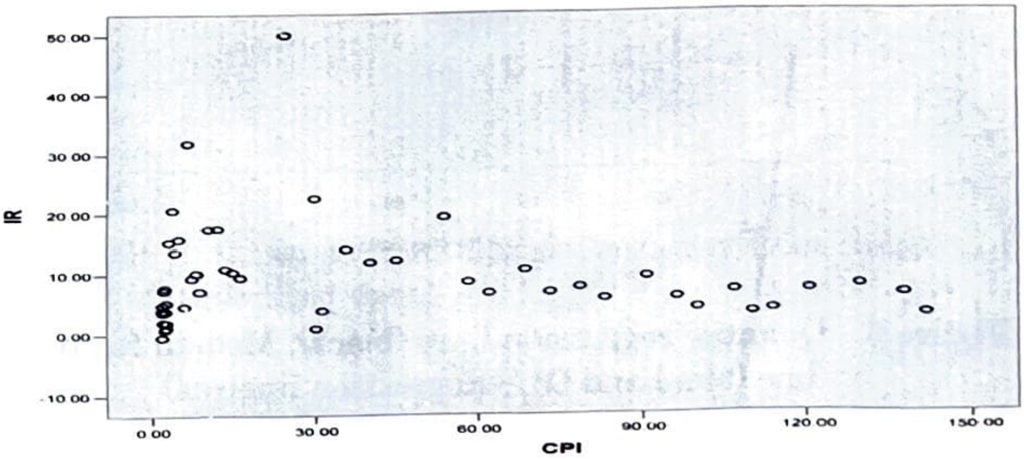

Figure 5. Scatter Plot of Inflation Rate versus Consumer Price Index

Table 2. Quantile Regression Estimates at Various Tau (τ) Values and Ordinary Least Squares Method (OLS)

Estimates	τ = .05	τ = .10	τ = .25	τ = .50	τ = .75	τ = .90	τ = .95	OLS
Intercept	0.77 (ns)	1.05 (ns)	3.90 (s)	8.86 (s)	15.41 (s)	20.99 (s)	33.17 (s)	10.77 (s)
Pro(>ltl)	0.295	.213	< .001	< .001	< .001	< .001	.001	< .001
CPI	0.014 (s)	0.017 (s)	.00102 (s)	-.019(ns)	-0.07 (s)	-.103 (s)	-.20 (s)	-.033 (ns)
Prob(>ltl)	< .001	< .001	< .001	.2798	< .001	< .001	< .001	.238

s- significant at .05 level ns- not significant at .05 level

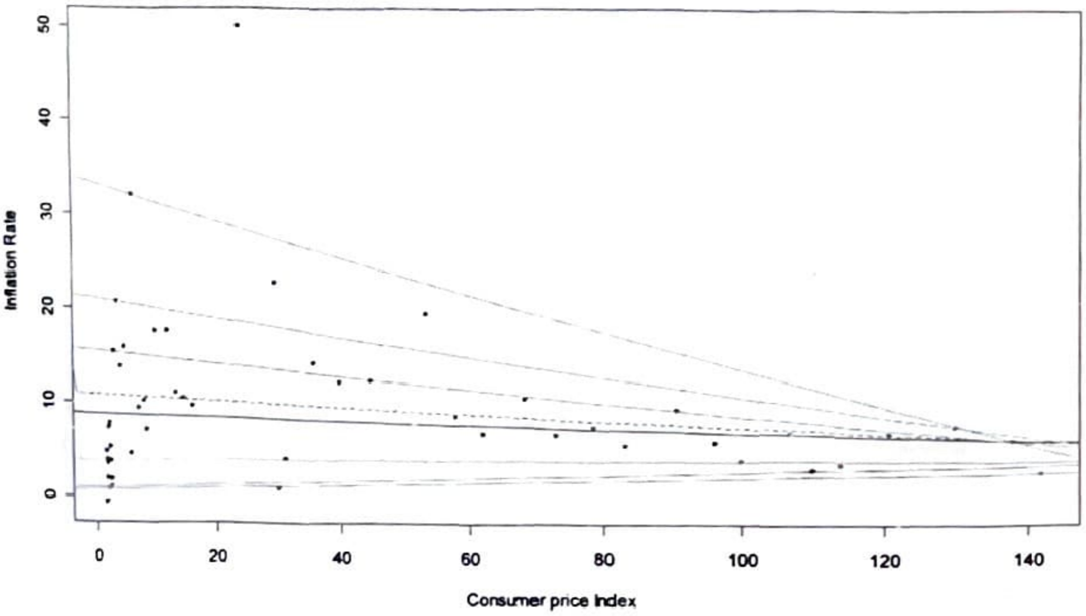

Figure 6. Quantile regression lines (black), Median Regression line (blue) and OLS Regression line(red)

5. Simulation Results

In order to draw relevant inference from the empirical results, a Montecarlo simulation was used using the R language to generate 1000 values for the first trial and 5000 values for the second trial. The normal distribution with parameters based from the estimates, the means and the standard deviations of the 50 annual values of inflation rates and consumer price indices of the Philippines were utilized in producing the simulated data. Repeating the computations done for the empirical results to the generated data, tables 3 to 6 were constructed . On the other hand, Figures 7 to 10 give the histograms of the four sets of generated data corresponding to the inflation rates and consumer price indices for trials 1(1,000 data) and trial 2(5,000 data).

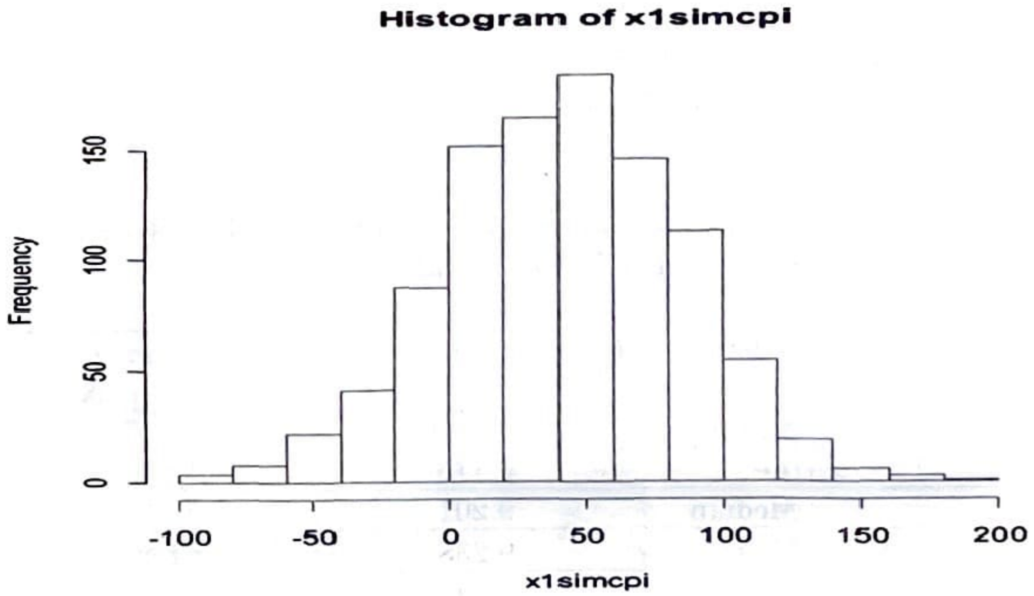

Figure7. Histogram of Simulated Consumer Price Indices(1,000 generated data)

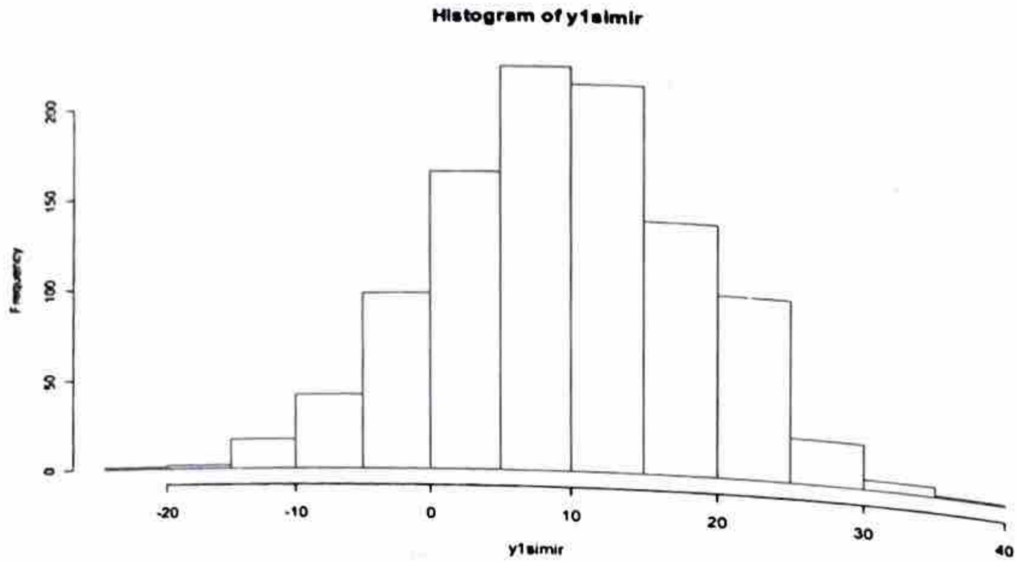

Figure 8. Histogram of Simulated inflation Rates (1000 generated data)

Table 3. Descriptive Statistics for Simulated Inflation Rates and Consumer Price Index Trial 1 (1,000 data)

Descriptive Statistics	Inflation Rates	Consumer Price Index
Minimum	-22.10	-92.67
1st Quartile	3.118	12.04
Median	9.201	42.16
Mean	9.238	41.43
3rd Quartile	15.39	69.99
Maximum	37.02	182.10
Variance	80.16	1865.024
Standard Deviation	8.95	43.19

It is apparent from figures 7 and 8 that the histograms approximate normal distribution with the means also approximating the medians. However, it can be observed that the median regression and the ordinary least squares consistently show a non significant linear relationship of the consumer price index and inflation rates for trial 1. The 5% and 10% quantile regression show a significant linear relationship of the consumer price index to the inflation rates. All the other quantile regressions indicate non significant relationship for the two variables. Increasing the number of generated samples to 5,000, it seems very apparent from Table 6 that all the regressions are consistently showing that the consumer price index is not significantly linearly related to the inflation rates. This can be attributed to the fact that these two variables are very much dependent on some seasonal variations factors and world conditions in general.

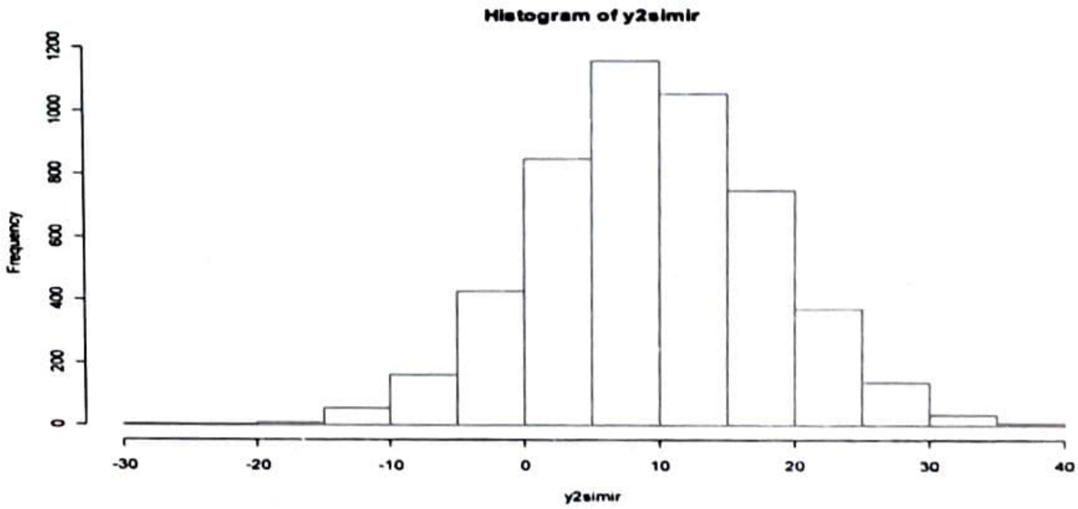

Figure 9. Histogram of Simulated Inflation Rates (5000 generated data)

Table 4. Descriptive Statistics for Simulated Inflation Rates and Consumer Price Index Trial 2 (5,000 data)

Descriptive Statistics	Inflation Rates	Consumer Price Inde
Minimum	-25.23	-114.50
1st Quartile	3.761	10.61
Median	9.368	39.92
Mean	9.407	40.15
3rd Quartile	15.27	71.41
Maximum Variance Standard Deviation	38.47 73.53 8.58	201.30 1969.048 44.374

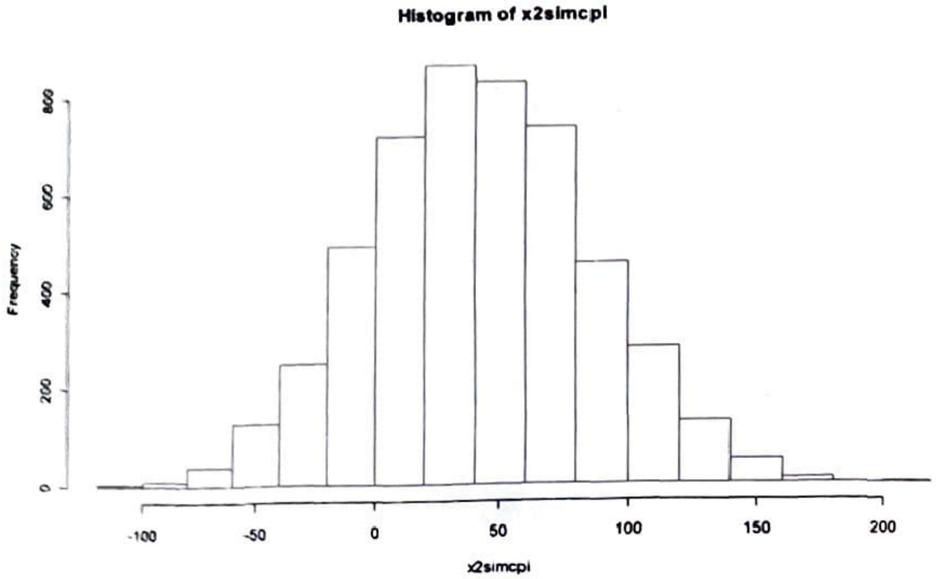

Figure 10. Histogram of the Simulated Consumer Price Indices(5000 generated data)

Table 5. Quantile Regression Estimates at Various Tau (τ) Values and Ordinary Least Squares Method (OLS) for trial 1~ 1000 generated data

Estimates	τ = .05	τ = .10	τ = .25	τ = .50	τ = .75	τ = .90	τ = .95	OLS
Intercept	-5.135(s)	-1.912(s)	3.702(s)	9.121(s)	15.2(s)	21.24(s)	23.84(s)	9.447(s)
Pro(>ltl)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
CPI	-.016(s)	-0.009(s)	-0.015(ns)	.00103(ns)	.004(ns)	-.003(ns)	-.0001(ns)	-0.005(ns)
Prob(>ltl)	<.001	<.001	.14	.914	.636	0.78	.996	0.442

s- significant at .05 level ns- not significant at .05 level

Note: Joint test on the Equality of Slopes: F value=.8381 (not significant)

The joint test on the equality of the slopes of the quantile linear regressions indicate the values of F are not significant for trials 1 and 2.

This implies that the slopes for the different quantile regressions are not significantly different.

Table 6. Quantile Regression Estimates at Various Tau (τ) Values and Ordinary Least Squares Method (OLS) for trial 2~ 5000 generated data

Estimates	τ = .05	τ = .10	τ = .25	τ = .50	τ = .75	τ = .90	τ = .95	OLS
Intercept	-4.593	-1.45(s)	3.839(s)	9.301(s)	14.98(s)	20.30(s)	23.68	9.32(s)
Pro(>ltl)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
CPI	0.002(ns)	0.0011(ns)	-.003(ns)	.002(ns)	.006(ns)	.0014(ns)	.001(ns)	.002(ns)
Prob(>ltl)	.748	.82	.455	0.619	.098	0.75	.87	.432

s- significant at .05 level ns- not significant at .05 level

Note: Joint test on the Equality of Slopes: F value=.9469 (not significant)

Conclusion

It can be inferred from the empirical results of the study that for data in the response and independent variables which violate the assumption of normality, both the Median and the OLS regression cannot give a significant linear regression models which is confirmed by the simulated data. Hence, the inflation rates cannot be predicted by the consumer price index using linear regression. At different quantiles, it was shown that the estimates of the regression models can be significantly linear for small sample sizes but not as the number of sample size increases. Maybe, Least Trimmed Squares (LTS) is appropriate which eliminates the outliers which might be the cause of the nonlinear significance.

Future Direction

It is recommended that a nonlinear quantile regression procedure be explored on how to model the relationship of the consumer price index and the inflation rates for accurate prediction. In addition, a study on the similarity and differences of Quantile Regression Procedures and Least Trimmed Squares (LTS) should be investigated.

References

Barrodale,I. And Roberts,F.D.K. (1973). "An Improved Algorithm for Discrete l_1 Linear Approximation," *SIAM J. Nmer. Anal.*, 10, 839-848.

Gutenbrunner, C. and Jureckova, J. (1992)."Regression rank scores and regression quantiles," *Annals Of Statistics* 20(1), 305-330.

Koenker, R and d'Orey,V. (1993), "Computing Regression Quantiles,"*Applied Statistics*,43, 410-414.

Koenker, R. And Bassett, G.W. (1978). "Regression Quantiles,",*Econometrica,* 46, 33-50.

Yaffee,R.A. (2002). " Robust Regression Analysis: Some Popular Statistical package Options", Statistics, Social Science and mapping Group, Academic Computing Services, Information Technology Services.