Response Variable Transformation for Quantile Regression Model

**Nwakuya Maureen T[(1)], Nwabueze Joy. C.[(2)], Onyegbuchulem, Besta.Okey[(3)]
Imoh, Johndamascene C.[(4)]**

**[1]Department of Maths/Statistics, University of Port Harcourt, River State, Nigeria
[2]Department of Statistics, Federal University of Agriculture Umudike, Nigeria.
[3 & 4] Department of Maths/Statistics, Imo State Polytechnic Umuagwo, Nigeria.**

Corresponding author: bokey@imopoly.net

Abstract — An alternative to ordinary least squares (OLS) regression model based on analytical solution is found to be quantile regression (QR) model. The procedure is well presented in this paper. Data Transformation and Square Root Transformation Simulation by Monte Carlo was carried out. Quantreg package in R software was used to illustrate the various model fitness for quantile regression model. The analysis shows that the best result was obtained from the square root of y transformation with an average error term (ϵ_i) of 0.9539, -0.0494, 0.0238, -0.5309 and -0.7544 for 10th, 25th, 50th, 75th and 90th quantile respectively. From the results obtained, it shows that model transformation can greatly improve the result of quantile regression model.

Key words: Quantile Regression, Joint Test of Equality of Slops, Mean Residual, Power Transformation, Log Transformation, Square root Transformation, Inverse of Square root transformation,

1 Introduction

Quantile regression model is naturally an extension of the linear-regression model. While the linear-regression model specifies the change in the conditional mean of the dependent variable associated with a change in the covariates, the quantile- regression model specifies changes in the conditional quantile. Since any quantile can be used, it is possible to model any predetermined position of the distribution. Thus, researchers can choose positions that are tailored to their specific inquiries

However, if multiple quantiles can be modeled, it is possible to achieve a more complete understanding of how the dependent distribution is affected by covariates, including information about shape change. According to [1] expected error term of multiple quantile regression can be improved by transforming the response variable. Also [2] uses the relationship between variances and means over several groups to find the appropriate transformation for the study data which makes the variance independent of the mean. [2] procedure for determining the appropriate transformation is to determine the coefficient (β) of regression of natural logarithm of group standard deviation ($\hat{\sigma}_i$) on the natural logarithm of group average $\left(\bar{x}_{i,i=1,2,---m}\right)$. He explained that the most popular and common transformations are the power of transformation such as: $\sqrt{x_t}$, $log_e X_t$, ${}^{1}\!/_{X_t}$, ${}^{1}\!/_{\sqrt{X_t}}$, ${}^{1}\!/_{X^2{}_t}$, $X^2{}_t$. Iwueze et'al (2011) stated that selecting the best transformation can be a complex issue and the usual statistical technique used is to estimate both the transformation and the required model for the transformed x_t at the same time. Arshad et'al (2016) empirically analyzed the monthly earning distribution of Pakistan using logarithm transformation, Therefore, aim of this study, is to investigate appropriate power transformation methods for quantile regression model, the study will specifically the study will assess the best transformation fit of the model based on some selected power transformations, assess the impact of selected operational covariates at different locations of the distribution on the response variable and Conduct diagnostic tests on the suggested model.

This study will apply the five powers of transformation stated by [2] on the response variable to ascertain which the power of transformations that will produce the least expected error term.

1.1 Statement of hypothesis

H_0: $\beta_i(\tau) = 0$, there is no significant difference in the slop patterns of the different quantiles

H_1 :$\beta_i(\tau) \neq 0$, there is a significant difference in the slop patterns of the different quantiles. .

2 Methodology

This paper investigates the best power transformation for multiple quantile regression model. The data were generated using Monte Carlo Simulation technique from the data of Annual salaries, income and wages of Health workers in Nigeria. The generated data shall be analyzed using transformed multiple quantile regression Model. The statistical software to be used in the analysis will be quantreg package in R Software.

2.1 Linear Quantile Regression Model

If we consider the i.i.d sample of $y_1, - - - y_n$, the unconditional sample mean can be defined as the solution to the problem of minimization a sum of squared residual

$$\hat{\mu} = \min_{\mu \in \Re} \sum_{i=1}^{n} \left(y_i - \mu\right)^2 \qquad (1)$$

Hence the sample median $\check{\xi}$ is the minimizer of the sum of absolute error loss or deviations.

$$\xi = \min_{\xi\in\Re}\sum_{i=1}^{n}\left|y_i - \xi\right| \tag{2}$$

To see why median can be define as a minimization problem, it can be written as:

$$E\left|y-\xi\right| = \int_{-\infty}^{\infty}\left|y-\xi\right| f(y)dy$$
$$\int_{y=-\infty}^{\xi}\left|y-\xi\right| f(y)dy + \int_{y=\xi}^{+\infty}\left|y-\xi\right| f(y)dy \tag{3}$$
$$\int_{y=-\infty}^{\xi}(\xi-y) f(y)dy + \int_{y=\xi}^{+\infty}(y-\xi) f(y)dy$$

Differentiating with respect to ξ and setting the partial derivative to zero will lead to the solution for the minimization problem. The partial derivative of the first term is:

$$\frac{\partial}{\partial\xi}\int_{y=-\infty}^{\xi}(\xi-y) f(y)dy = (\xi-y) f(y)\Big|_{y=\xi} + \int_{y=-\infty}^{\xi}\frac{\partial}{\partial\xi}(\xi-y) f(y)dy$$
$$\int_{y=-\infty}^{\xi} f(y)d(y) = F(\xi) \qquad \left[Huo\ L.\ et'al(2013)\right] \tag{4}$$

And any partial derivative of the second term is:

$$\frac{\partial}{\partial\xi}\int_{y=\xi}^{+\infty}(y-\xi) f(y)dy = \int_{y=\xi}^{+\infty} f(y)d(y) = -\left(1-F(\xi)\right) \tag{5}$$

Combining these two partial derivatives lead to:

$$\frac{\partial}{\partial\xi}\int_{-\infty}^{+\infty}(y-\xi) f(y)dy = F(\xi)-\left(1-F(\xi)\right) \tag{6}$$
$$= 2F(\xi)-1 \qquad \left[Huo\ L.\ et'al(2013)\right]$$

By setting $2F(\xi) - 1 = 0$, we solve for the value of $F(\xi) = \frac{1}{2}$, that is, the median, to satisfy the minimization problem. For the general τth sample quantile ξ(τ), which is the analogue of $\rho(\tau)$, may be formulated as the solution of the optimization problem

$$\rho\tau(\xi) = \min_{\xi\in\Re}\sum_{i=1}^{n}\rho\tau(y_i - \xi) \tag{7}$$

Repeating the above argument for quantiles, the partial derivative for quantiles corresponding to equation (7)

$$\frac{\partial}{\partial\xi}E\left[\rho\tau(y,\xi)\right] = (1-\rho)F(\xi) - \rho\left(1-F(\xi)\right)$$
$$= F(\xi) - F(\xi)\rho - \rho + F(\xi)\rho \tag{8}$$
$$= F(\xi) - \rho \qquad \left[Huo\ L.\ et'al(2013)\right]$$

We set the partial derivative $F(\xi) - \rho = 0$ and solve for the value of $F(\xi) - \rho$ that satisfies the minimization problem. (8) is illustrated thus

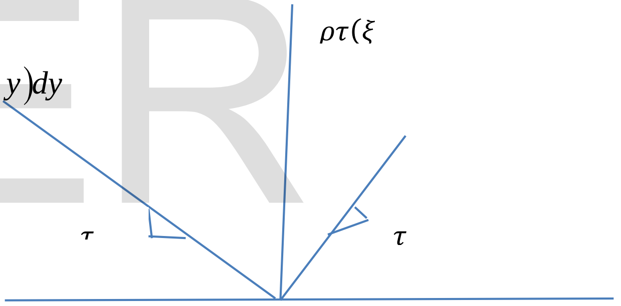


Fig.1 Quantile Regression ρ

Just as the unconditional sample mean in (1) minimizes the sum of square residuals (error lose), the conditional sample mean also minimizes the sum of square residual by replacing the scalar μ by $\mu(x_i, \beta)$, the estimate of the conditional mean function $E[Y/x]$ is obtained

$$E(Y/x) = \min_{\beta\in\Re}\sum_{i=1}^{n}\left(y_i - \mu(x_i,\beta)^2\right) \tag{9}$$

This can be proceeded in the same way in quantile regression. to obtain an estimate of the conditional median function, the scalar ξ in equation (2) is replaced by the parametric function $\xi(x_i, \beta)$:

$$\tilde{\beta}\left(\tau = \tfrac{1}{2}\right) = \min_{\xi\in\Re}\sum_{i=1}^{n}\left|y_i - \xi(x_i,\beta)\right| \tag{10}$$

To obtain the estimates of the other conditional quantile function the conditional quantile is considered and the absolute values is replace by $\rho\tau(.)$:

$$\rho\tau\left(\tilde{\beta}\right) = \min_{\xi \in \Re} \sum_{i=1}^{n} \rho\tau\left(y_i - \xi\left(x_i, \beta\right)\right) \qquad (11)$$

Minimizing (11) results in a quantile regression model. The resulting minimization problem of (11), when $\xi(x_i, \beta(\tau))$ is formulated as a linear function of the parameters can be solved very efficiently by linear programming method. The progression of ideas that led to (11) motivated the original quantile regression model presented in [6].

2.2 Model Specification

Following [6] and [7], our proposed model will take the form:

$$Z^{\tau} = \beta_0^{(\tau)} + \beta_i^{(\tau)} x_i \ \varepsilon_i^{(\tau)} \qquad (12)$$

Where
k = number of covariates
Z = a transformed vector containing n observations of Health Workers Allowances, ie the transformed response variable (transformed Y_t)
β = a vector containing 16 coefficients to be estimated
ε= a classical error terms
τ = Specified quantiles of Health Workers Annual Allowances. This research examines the following quantiles: 0.1, 0.25, 0.5, 0.75, 0.9
X = an (3000) x 15 matrix of the covariates
And we used the sample sizes: $n = 3000$

2.3 Coefficient of determination

The goodness of fit according to [4] will be measured in a manner that is consistent with this criterion. But [8] suggested measuring goodness of fit by comparing the sum of weighted distances for the model of interest with the sum in which only the intercept parameter appears.

Let $V^1(p)$ be the sum of weighted distance for the full p^{th} quantile regression model,

Let $V^0(p)$ be the sum of weighted distance for the model that includes only a constant term. Therefore, using the one covariate model

$$V^1(p) = \sum_{i=1}^{n} d_p\left(y_i, \hat{y}_i\right)$$

$$= p \sum_{y_i \geq \beta_0^{(p)} + \beta_i^{(p)} x_i} \left|y_i - \beta_0^{(p)} - \beta_i^{(p)} x_i\right| \quad + \quad 1 - p \sum_{y_i < \beta_i^{(p)} + \beta_i^{(p)} x_i} \left|y_i -\right.$$

$$V^0(p) = \sum_{i=1}^{n} d_p\left(y_i, \hat{y}_0\right)$$

$$= \sum_{y_i \geq \beta_0^{(p)}} p\left|y_i - \beta_0^{(p)}\right| + \sum_{y_i < \bar{y}} (1-p)\left|y_i - \beta_0^{(p)}\right| \qquad (13)$$

For the model that only includes a constant term, the fitted constant is the sample p^{th} quantile $\hat{Q}^{(p)}$ for the sample $y_1, \ldots\ldots\ldots. y_n$ the goodness of fit is then defined as

$$R(p) = 1 - \frac{V^1(p)}{V^0(p)} \qquad \left[Huo\ L.et'al(2013)\right] \qquad (14)$$

Since $V^0(p)$ and $V^1(p)$ are nonnegative, R(p) is at most 1. Also, because the sum of weighted distance is minimized for the full-fitted model, $V^1(p)$ is never greater than $V^0(p)$, so R(p) is greater than or equal to zero. Thus, R(p) is within range of [0,1], a larger R(p) indicates a better model fit. The R(p) defined above allows for comparison of a fitted model with any number of covariates beyond the intercept term to model in which only the intercept term is present. This is the restricted form of a goodness-of-fit introduced by [8] for nested models.

3 Simulated Data set for Quantile Regression Model:

From the results of 'Individual Distribution Identification tool' of the Minitab software, Normal distributions was found to fit the variables under study. To simulate the data, quantile function of normal distribution function (probit function) was derive by equating the CDF to p and theoretically solve for x. Probability Density Function of a normal distribution function is given as

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{(2\sigma^2)}} \qquad (15)$$

Let the Cumulative Density Function (CDF) be denoted by F

$$\therefore \quad F \quad = \int_{-\infty}^{x} p(x)dx \qquad (16)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{2}\left[1 + erf\left(\frac{x - \mu}{\sigma\sqrt{\sqrt{2}}}\right)\right] \qquad (17)$$

This is proceeded by deriving the probit function theoretically as:

$$F = \frac{1}{2}\left[1 + erf\left(\frac{x - \mu}{\sigma\sqrt{\sqrt{2}}}\right)\right]$$

$$1 + erf\left(\frac{x - \mu}{\sigma\sqrt{\sqrt{2}}}\right) = 2F$$

$$erf\left(\frac{x - \mu}{\sigma\sqrt{\sqrt{2}}}\right) = 2F - 1$$

$$\frac{x - \mu}{\sigma\sqrt{\sqrt{2}}} = erf^{-1}(2F - 1)$$

$$x - \mu = \sigma\sqrt{\sqrt{2}}\, erf^{-1}(2F - 1)$$

$$x = \mu + \sigma\sqrt{\sqrt{2}}\, erf^{-1}(2F - 1)$$

With $\mu = 0$ and $\sigma = 1$

$$x = \sqrt{\sqrt{2}}\, erf^{-1}(2F - 1), \; p \in (0,1) \qquad (18)$$

Where

$x = probit\ function$

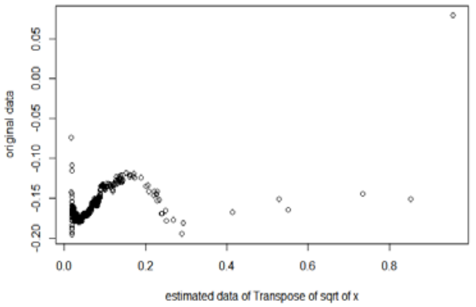
erf = the error function

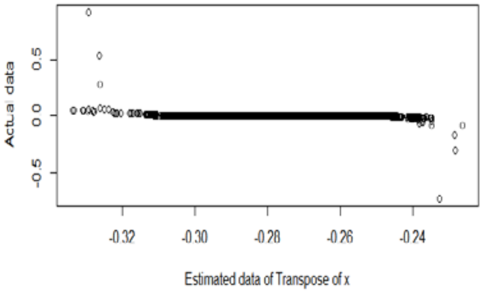
F = Cumulative Density Function (CDF)

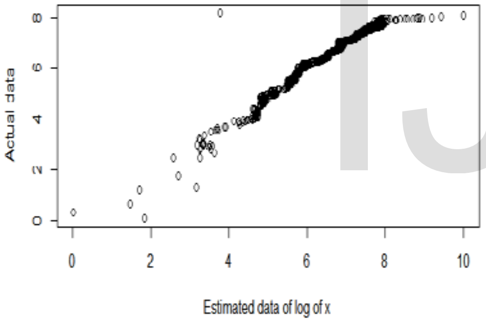
Following the derivation of probity function, Monte Carlo simulation will then be applied on the derived function to generate sample size of 3000 on each variable x_i

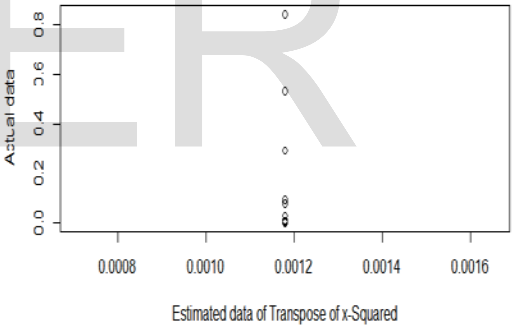
3.1 Results and Discussion

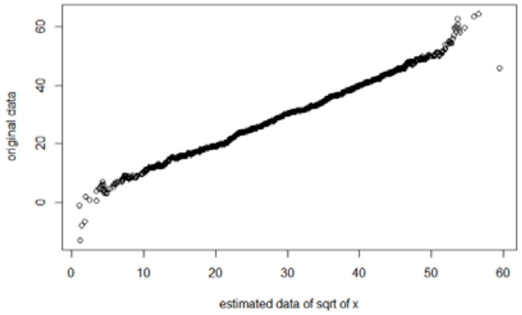
The result of the model transformation was presented in table 1, expected error term of different models of the median regression where compared together with the other quantiles,

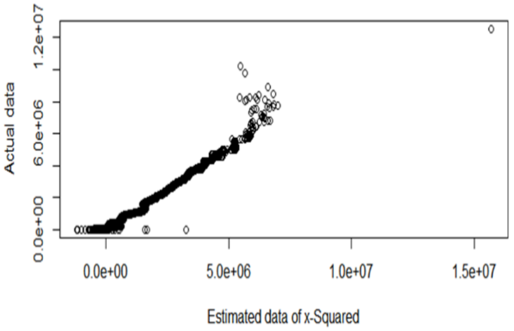












from the results of the expected error term, log transformation seems to has significantly improved the result of the model having the expected error term of -0.0196 for the median regression, this is followed by the result of square root of y transformation with the expected error term of 0.0238 for its median regression. y- squared transformation shows a significant retardation in all the quantiles including the median regression that is, its expected error term is 6907.125; the result of the inverse of y-squared transformation shows a more improved results of expected error terms in all the quantile than the five other model transformation with the result of the expected error term of the median regression as -0.0005 but surprisingly the expected error term of the 75th quantile seems to have more improved result than the expected error term of the median regression which make the result spurious. Appendix 1 shows the graphs of the expected value plotted against the actual value.

Table 1:Mean Residual of the Transformed Quantile Regression Model

	Quantile				
Transf. Model	.10	.25	.50	.75	.90
Log of y.	0.1596	0.0319	-0.0196	-0.4734	-0.0973
Sqrt of y T	0.9539	-0.0494	0.0238	-0.5309	-0.7544
Inverse of Sqrt of y T	0.2646	-0.0581	0.2096	0.1705	0.3199
Inverse of y	0.3820	0.5598	0.2773	0.0063	0.0123
Inverse of y sqrd	0.0005	0.0003	-0.0005	-0.0027	-0.0067
y sqrd	210104.	-2302894	6907.12	-118156.	-217223
Psuado R	0.9987	0.9997	0.9999	0.9998	0.9985

Graph of Log transformation shows that the estimated data may be partially correlated with the actual data, the graph of the square root of y transformation shows a more perfect correlation between the estimated value and the actual value. Both the graph of the estimated data and the actual data for the inverse of square root of y transformation and y-squared transformation show an imperfect correlation. While the graphs of the inverse of y and the inverse of the square root of y show no correlation between the estimated data and the actual data. From the results of the graphs, it can be suggested that the model of the results of log of y transformation and the inverse of y-squared transformation are more of spurious results suggesting that the best model transformation may be the square root of y transformation.

Table 2: Coefficient and p-value of Square Root Transformed Model

	0.10		0.25		0.50		0.75		0.90	
	Co Value	P(>|t|	Coe v	P(>|t|	Coe v	P(>|t|	Coe v	P(>|t|	Coe v	P(>|t|)
β_0	-180.9	0.00	-161.89	0.00	-122.8	0.00	-13.20	0.30	-15.60	0.02
β_1	-0.043	0.00	-0.0274	0.00	-0.019	0.00	0.001	0.81	-0.023	0.00
β_2	-0.018	0.00	-0.0083	0.00	-0.001	0.55	0.004	0.08	0.009	0.00
β_3	-0.047	0.00	-0.0373	0.00	-0.029	0.00	-0.014	0.00	-0.024	0.00
β_4	0.0350	0.00	0.0268	0.00	0.0217	0.00	0.029	0.00	0.026	0.00
β_5	0.0209	0.00	0.0226	0.00	0.0234	0.00	0.018	0.00	0.008	0.00
β_6	-0.012	0.00	-0.0138	0.00	-0.011	0.00	-0.004	0.00	0.001	0.27
β_7	-0.004	0.00	-0.0067	0.00	-0.008	0.00	-0.006	0.00	0.000	0.77
β_8	0.0049	0.00	0.0053	0.00	0.0049	0.00	0.003	0.00	0.004	0.00
β_9	0.0036	0.00	0.0033	0.00	0.0020	0.00	-0.000	0.44	-0.000	0.44
β_{11}	-0.004	0.00	-0.0045	0.00	-0.004	0.00	-0.004	0.00	-0.000	0.58
β_{12}	0.0117	0.00	0.0065	0.00	0.0040	0.00	0.000	0.88	0.002	0.00
β_{13}	-0.009	0.00	-0.0071	0.00	-0.008	0.00	-0.008	0.00	-0.004	0.00
β_{14}	0.0042	0.00	0.0051	0.00	0.0060	0.00	0.004	0.00	0.004	0.00
β_{15}	0.0062	0.00	0.0052	0.00	0.0043	0.00	0.002	0.00	-0.003	0.00
β_{16}	0.0002	0.51	0.0009	0.00	0.0004	0.01	0.000	0.05	0.000	0.63

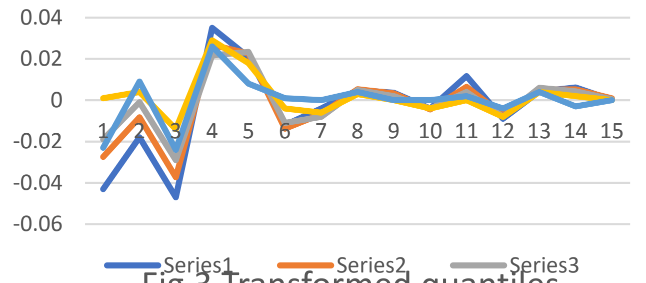


Fig.3 Transformed quantiles Model Coefficient Plot

Table 3: Joint Test of Equality of Slops

Joint Test	Df R	Df	F Value	$Pr(> F)$
0.10 & 0.90	15	5985	318.23	2.2e-16***
0.25 & 0.75	15	5985	75.869	2.2e-16***

The results of the p-values in Table 2, show that all the coefficient values are all significant expect for x_{15} *in the* $10th$ quantile and x_2 in the 50th quantile also x_1, x_2, x_9, x_{11} *and* x_{15} in the $75th$ *quantil* and x_6, x_7, x_9, x_{10} and x_{15} in the 90th quantile. The result of the Joint test for equality of slopes in Table 3 shows a significant difference in the slop of the 10th and 90th quantiles and the 25th and 75th quantiles also a significantly difference in their slop and in fact the 10th, 25th, 50th, 75th and 90th quantile, significantly follow different slope pattern. Also, as shown in Table.2, while the covariates x_1, x_2, x_3, x_6, x_7, x_{10} and x_{12} of the 10th, 25th and 50th quantile show that they negatively impact on the response variable, the covariates of x_4, x_5, x_7, x_8, x_9, x_{11}, x_{13}, x_{14}, and x_{15} have positive impact on the response variable. Judging from the 75th and 90th quantile, while the covariates x_3, x_6, x_7, x_9, x_{10}, and x_{11} show that they negatively impact on the response variable. The covariates x_1, x_2, x_4, x_5, x_8, x_{13} and x_{14} have positive impact on the response variable. x_{11}, and x_{15} of the 75th and 90th quantile did not show any significant impact on the response variable.

References

[1] Arshad, I. A., Younas, U., Shaikh,A.W & Chandio,M.S (2016). Quantile Regression Analysis of Monthly Earnings in Pakistan; *Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (4) 919-924 (2016)*

[2] Bartlett, M.S (1974). The use of Transformation, Biometrica 3,39 - 52

[3] Frost, J (2012) How to Identify the Distribution of Your Data using Minitab, http://www.scribd.com/doc/84506538/Body-Fat-Data-for-Identifying-Distribution-in-Minitab

[4] Hao L. &Naiman,D.Q., (2007). Quantile Regression*; 01-Hao.qxd. 3/13/2007.3.28*

[5] Iwueze, S.I., Nwogu, E.C., Ohakwe, J. & Ajaraogu, J.C. (2011) Uses of the Buys-Ballot Table in Time Series Analysis, Applied Mathematics *Journal. (2) 633-645*

[6] Koenker,R & Bassett, G. (1978); Regression Quantiles, Econometrica, Vol. 46, No. 1, pp. 33-50. http://links.jstor.org/sici?sici=0012-9682%28197801%2946%3A1%3C33%3ARQ%3E2.0.CO%3B2-J

[7] Koenker, R. (2011). *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press. (6)6.

[8] Koenker, R. & Machado J.A (1999) Goodness of fit and related inference processes for quantile regression. *Journal of Econometrics*, 93, 327-344

[9] Lee, B.-J. & Lee, M. J. (2006). Quantile regression analysis of wage determinants in the Korean labor market, *The Journal of the Korean Economy*, 7, 1–31.

[10] Meinshausen, N. (2006); Quantile Regression Forests, *Journal of Machine Learning Research*, (7) 983–99

[11] Young, T.M., Shaffer, L.B., Guess, F. M., Bensmail, H. &Leon, R.V (2008), A comparison of multiple linear regression and quantile regression for modeling the internal bond of medium density fiberboard; *Forest Products Journal*, 58(4).