Probability Statistics and Econometric Journal
Vol.4, No.1, January-February2022;
ISSN (5342 –3371);
p –ISSN (9301–4221)
Impact factor: 0.00

COMPARATIVE STUDY OF MANLY TRANSFORMATION AND YEO-JOHNSON TRANSFORMATION IN QUANTILE REGRESSION ANALYSIS

[1]Nwakuya, Maureen Tobechukwu and [2]Biu Emmanuel O.
[1&2]Department of Mathematics and Statistics, University of Port Harcourt, Rivers State, Nigeria.

Abstract: This paper tries to implement Yeo-Johnson and Manly transformations in a quantile regression analysis while using the Chamberlain and Buchinsky two stage estimator to estimate the transformation parameter for both methods. A comparison was done between the two methods. The comparison was based on the bias of the estimates, AIC and MSE. The results of this analysis revealed that at all quantiles considered (0.05, 0.25, 0.5, 0.75 and 0.95), the Yeo-Johnson transformed quantile regression performed better than the Manly transformed quantile regression based on the AIC and MSE. The bias of the estimates also indicated that the Yeo-Johnson transformed Quantile regression had better results except at the 25th quantile where the Manly method had a better result.

Keywords: Yeo-Johnson transformation, Manly Transformation, Chamberlain and Buchinsky two stage method and quantile Regression.

INTRODUCTION

Regression analysis is a cornerstone of statistical analysis in many fields. Benjamin, L. C. and Willard, G. M. [2] stated that Common regression methods measure differences in outcome variables between populations at the mean (i.e., ordinary least squares regression), or a population average effect (i.e., logistic regression models), after adjustment for other explanatory variables of interest. There are assumptions guarding the regression analysis, which includes that the response variable should be normally distributed, but in most situations this assumption is violated. This leads to researchers been faced with data that deviates from normality and that therefore requires treatment such that the assumption is met or approximated. Normalizing data is particularly relevant when parametric tests (e.g., analysis of variance [ANOVA] and linear mixed models) and, even, non-parametric tests are used. Jorge I. Vélez et al [9] stated that one of the methods to enhance data's normality is via transformations. Li, P. [12] in his work mentioned that Box and Cox [3] proposed a parametric power transformation technique defined by a single parameter λ, aimed at reducing anomalies in the data and ensuring that the usual assumptions for a linear model hold. Sakia RM [16] hinted that this transformation results from modifying the family of power transformations as was defined by Tukey J. W. [18] to account for the discontinuity at $\lambda = 0$. Box-Cox transformation basically applies a deterministic power function to the raw data by using the estimate of the power transformation parameter, λ. The Box-Cox transformation as introduced by Box-Cox [3] is a family of power transformations such that the transformed values are a monotonic function of the observations over some admissible range and is given by;

Probability Statistics and Econometric Journal
Vol.4, No.1, January-February2022;
ISSN (5342 –3371);
p –ISSN (9301–4221)
Impact factor: 0.00

$$y_i^{\lambda} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} \; if \; \lambda \neq 0 \\ lny_i \quad if \lambda = 0 \end{cases} \; for \; y_i > 0 \tag{1}$$

Where λ, is an unknown power transformation parameter. However the Box-Cox transformation is only valid for positive y. To circumvent this restriction, Manly [13], proposed a one parameter exponential transformation as an alternative to Box Cox transformation because it allows negative y values. Yeo, I. and Johnson, R. A. [20] introduced an extension of the Box-Cox transformation. They proposed a new family of distributions that can be used without restrictions and also has many of the good properties of the Box-Cox power family.

Powell [14] introduced the Box–Cox quantile regression model (BCQR) as a flexible and numerically attractive extension of linear quantile regression techniques. He used Chamberlain [5] and Buchinsky [4] two stage estimator to estimate the transformation parameter, unlike the traditional Box-Cox transformation that uses the maximum likelihood estimator. The draw back with this two stage method is that in the second stage the objective function is not well defined. In a discussion paper by Fitzenberger, B. et al [7] they implemented this BCQR using this two stage estimation method by Chamberlain [5] and Buchinsky [4], where they suggested a simple modification of the objective function in order to ensure that it is well defined. In this paper we tried to implement Yeo-Johnson and Manly transformations in a quantile regression analysis while using the Chamberlain and Buchinsky two stage estimator.

2 QUANTILE REGRESSION

Quantile regression is a tool that can supplement traditional analysis when data fail the normality of errors assumption, which is key to ordinary least square regression. In ordinary least square regression, these outcome variables may fail to satisfy the assumption that the residuals are normal, homoscedastic (have a constant variance), and uncorrelated. Sometimes such variables can be transformed to satisfy the ordinary least square assumptions and then linear regression is utilized. Key difference is that with quantile regression, a specific quantile or multiple quantiles of the outcome may be of interest. Furthermore, the results produced from the quantile regression model are interpreted in the context of the quantile (ie, the percentile) of the distribution of the outcome being modeled, rather than the mean. Quantile regression is vastly more robust to outliers than ordinary least square regression because observations far from the mean may have high leverage and may cause significant bias in estimates of the mean. Steven J. S. et al [17] hinted in the literature that quantiles of the outcome variable possess monotone equi-variance, which means that monotone transformations (such as linear or a log transformation) which keep the data in the same ascending or descending order can be implemented without altering the estimated quantiles. This paper tries to implement the Yeo-Johnson transformation and the Manly transformation in quantile regression using the two step method by Chamberlain [5] and Buchinsky [4] (CBTS) as an estimation method for the transformation parameter and comparing both methods using some model comparison criteria.

3 YEO-JOHSON TRANSFORMATION

Atkinson [1] advised that in choosing a transformation for a given data set, our choice of transformation should provide simple and more revealing analyses that lead to valid inferences. Raymaekers, J. & Rousseeuw, P. J. [15] mentioned that, in order to transform positive variable to give it a more normal distribution one often resorts to power transformation. Box-Cox power transformation family is often chosen from the family of parametric transformations and equivalent to the power

Probability Statistics and Econometric Journal
Vol.4, No.1, January-February2022;
ISSN (5342 –3371);
p –ISSN (9301–4221)
Impact factor: 0.00

transformation for univariate data transformation. This transformation is defined as in equation (1). The transformation parameter is usually estimated using the method of maximum likelihood as proposed by Box and Cox [3]. The Box-Cox transformation possesses special features which makes it perhaps more preferable as the chosen response transformation method and or for transforming set of exogenous variables towards normality. At $\lambda = 1$, the Box-Cox transformation corresponds to no transformation as we can see in the equation (1) above. When $\lambda = 1/2$, this transformation will resolve to the square root transformation. When $\lambda = 0$, the logarithm transformation is chosen and with $\lambda = -1$ the inverse or reciprocal transformation is opted for. However, a major limitation associated with the family of Box-Cox transformations is that it can only be applied to positive response data points or observations. Yeo I. and Johnson, R. A. [20] proposed an alternative family of transformations that addresses this limitation and can deal with both positive and negative response data points. These transformations are described by the function below:

$$G_\lambda(y) = \begin{cases} \frac{((1+y)^\lambda - 1)}{\lambda} & if\ \lambda \neq 0\ and\ y \geq 0 \\ \log(y+1) & if\ \lambda = 0\ and\ y \geq 0 \\ \frac{-[(-y+1)^{2-\lambda} - 1)]}{(2-\lambda)} & if\ \lambda \neq 2\ and\ y < 0 \\ -\log(-y+1) & if\ \lambda = 2\ and\ y < 0 \end{cases} \quad (2)$$

In equation (2), If y is strictly positive, then the Yeo-Johnson transformation is the same as the Box-Cox power transformation of $(y+1)$. If y is strictly negative, then the Yeo-Johnson transformation is the Box-Cox power transformation of $(-y+1)$ with power$2-\lambda$. When y has both negative and positive values, the transformation is a mixture of these two, so different powers is used for positive and negative values of y. Equation (2) shows the Yeo-Johnson transformation for a range of values of λ. In these transformations $\lambda = 1$ yields a linear relation while the right tail of the distribution of the response is compressed when the transformation is done with $\lambda < 1$ but it expands the left tail with $\lambda > 1$ making the transformation suitable for transforming both right-skewed and left-skewed distributions towards symmetry.

4 MANLY TRANSFORMATION

Manly [13], proposed a one parameter exponential transformation as an alternative to Box Cox transformation because it allows negative y values. The transformation by Manly is given as:

$$Y^* = \begin{cases} \frac{y^\lambda - 1}{\lambda} & ,\lambda \neq 0 \\ lny & ,\lambda = 0 \end{cases} \quad for\ y > 0 \quad (3)$$

Where λ is the transformation parameter, Y* is the transformed response variable and y is the observed response variable and it is restricted to be positive. Watthanacheewakul L. [19] stated that Manly transformation is quite effective in turning skewed unimodal distribution into nearly symmetric distributions but is not quite useful for bimodal or U-shaped distribution.

4 METHODOLOGY:

The Traditional mean regression model is given by;

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},\ i = 1, \ldots, n \quad (4)$$

Where *p* is the number of predictor variable in the equation and *n* is the number of data points. Dye, S. [6] stated that the best linear regression line is found by minimizing the mean square error, which is given by;

Probability Statistics and Econometric Journal
Vol.4, No.1, January-February2022;
ISSN (5342 –3371);
p –ISSN (9301–4221)
Impact factor: 0.00

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})\right)^2 \quad (5)$$

Just as we can define the sample mean as the solution to the problem of minimizing a sum of squared residuals, we can define the median as the solution to the problem of minimizing a sum of absolute residuals. Koenker, R. and Hallock K. F. [11] hinted that the symmetry of the piecewise linear absolute value function implies that the minimization of the sum of absolute residuals must equate the number of positive and negative residuals, thus assuring that there are the same number of observations above and below the median. Hence in other to obtain other quantiles a sum of asymmetrically weighted absolute residuals is minimized by assigning weights to positive and negative residuals since the symmetry of the absolute value yields the median. Hence the unconditional minimization problem is thus:

$$\min_{\zeta\epsilon\Re}\sum\rho_\tau(y_i-\zeta) \quad (6)$$

Where $\rho_\tau\ \rho_\tau(u)=\tau|u|I_{u\geq0}+(1-\tau)|u|I_{u<0}$.To obtain an estimate of the conditional median function, the scalar ζ in the equation is replaced by the parametric function $\zeta(x_i\,,\beta)$ and τ is set to half. To estimate other conditional quantile functions, the absolute value that yields the median is replaced by $\rho_\tau(.)$ and the minimization problem below in equation () is solved using a linear programming method.

$$\min_{\zeta\epsilon\Re}\sum\rho_\tau(y_i-\zeta(x_i\,,\beta)) \quad (7)$$

The conditional quantile regression model equation for the τth quantile is given by Koenker R. [10] as; $Q_\tau(y_i)=\beta_0(\tau)+\beta_1(\tau)x_{i1}+\cdots+\beta_p(\tau)x_{ip},\ i=1,\ldots,n$ (8)

Where the quantile $\tau\in0,1$ and for each quantile level τ, the solution to the minimization problem yields distinct sets of regression coefficients at different levels of the quantile.

4.1 Chamberlain and Buchinsky Two Stage method

In other to implement Yeo-Johnson and Manly transformations, the transformation parameters were obtained using the two step method (CBTS) by Chamberlain, [5] and Buchinsky, [4]. The procedure for estimating the transformation parameter is as follows:

First estimate $\beta_\tau(\lambda)$ conditional on λ by solving the minimization problem;

$$\hat{\beta}_\tau(\lambda)=arg\min_{\beta}n^{-1}\sum_{i=1}^{n}\rho_\tau(y_{\lambda i}-x'\beta) \quad (9)$$

Secondly estimate λ_τ by solving the minimization problem;

$$\hat{\lambda}_\tau=\min_{\lambda}n^{-1}\sum_{i=1}^{n}\rho_\tau\left(y_i-(\lambda x'\hat{\beta}_\tau(\lambda)+1)^{-1/\lambda}\right) \quad (10)$$

Conditioned on the assumption that, $(\lambda x'\hat{\beta}_\tau(\lambda)+1>0$, this procedure was implemented in R by Geraci [8]. In this research work, two quantile regression analysis were performed; one implementing the Yeo-Johnson transformation and the other implementing the Manly transformation. The analysis considered 0.1, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80 and 0.90 quantiles. Results of both analyses were investigated and the resulting models were compared based on bias of the estimates, their mean square error (MSE) and Akaike Information criteria (AIC). The data for this work is Iris data obtained from R

Probability Statistics and Econometric Journal
Vol.4, No.1, January-February2022;
ISSN (5342 –3371);
p –ISSN (9301–4221)
Impact factor: 0.00

data base. The analysis was done in R using QTOOLS package.

5 RESULTS

Table 1: Quantile Regression Estimates, Bias and P-values from Yeo-Johnson and Manly Transformed Data

		Yeo_Johnson results			Manly Results		
Quantiles	Coefficients	Estimates	Bias	Pvalue	Estimates	Bias	Pvalue
0.05	Petal.Width	2.4502715	**-0.007468433**	0.0000e+00**	2984.529	441.21359	0**
	Sepal.Width	-0.9046627	**-0.068308360**	1.181745e-06**	1187.332	89.93136	0**
	Sepal.length	1.2819336	**-0.035456325**	0.0000e+00**	981.328	16.70265	0**
0.25	Petal.Width	2.4337489	-0.29183400	0.0000e+00**	5855.390	**-2039.03767**	0**
	Sepal.Width	-0.8588496	0.05039043	0.0000e+00**	2474.371	**-734.32431**	0**
	Sepal.length	0.9656594	-0.10265957	6.128431e-14**	1180.378	**-53.45443**	0**
0.50	Petal.Width	1.8545153	**0.17758666**	0.000000**	14.816876	234.1537	0**
	Sepal.Width	-0.7522313	**0.02391979**	0.000000**	1.505880	148.5627	0**
	Sepal.length	0.7657556	**0.05763558**	0.000000**	4.517948	108.6005	0**
0.75	Petal.Width	1.9755970	**0.21425554**	0.00000e+0**	3.842851	0.8484161	0**
	Sepal.Width	-0.7099961	**-0.02024253**	3.504062e-08**	-1.111153	0.3029901	0**
	Sepal.length	1.2407516	**-0.13686256**	0.00000e+00**	3.447569	0.1977202	0**
	Petal.Wid	3.0501713	-	0.00000e+00	2.50945	0.032818	0.0000e+00*

Probability Statistics and Econometric Journal
Vol.4, No.1, January-February2022;
ISSN (5342 –3371);
p –ISSN (9301–4221)
Impact factor: 0.00

0.95	th		**0.16731426**	**	1	53	*
	Sepal.Width	-0.9970387	**-0.05528129**	1.924838e-06**	-1.001720	0.08847003	9.959878e-09**
	Sepal.length	1.8868346	**-0.13174537**	0.00000e+00**	2.005500	0.25300723	0.0000e+00**

**significant at 0.05 level significant

The table 1 above shows results of the estimates, bias and P-value of the quantile regression using both Yeo-Johnson and Manly transformed data. The coefficients were all significant at all quantiles for both methods. All the bias of the estimates from Yeo-Johnson transformed data was smaller than the bias from the estimates of Manly transformed data except at the 25th quantile.

Table 2: AIC and MSE results from the Yeo-Johnson Transformation and Manly Transformation Method

Tau	Comparison Criteria	Manly Transformation	Yeo-Johnson Transformed
0.05	AIC	1255.83	**152.7394**
	MSE	4100.521	**0.3424615**
0.25	AIC	1354.821	**-286.0995**
	MSE	7933.147	**0.140765**
0.50	AIC	1458.643	**-338.4462**
	MSE	15850.43	**0.09929684**
0.75	AIC	1596.748	**-267.824**
	MSE	39801.12	**0.1590039**
0.95	AIC	1701.812	**-159.3893**
	MSE	80184.08	**0.327611**

Table 2 above shows us that the Yeo-Johnson transformed data produced better results based on the AIC and MSE at all quantiles compared to the Manly transformation data. Hence it is considered as a better model.

Conclusions:

This paper investigated the implementation of Yeo-Johnson transformation and Manly transformation in quantile regression while using the two-step estimator proposed by Chamberlain, (1994) and Buchinsky, (1995) to estimate the transformation parameter. The Yeo-Johnson transformed quantile regression method was compared to the Manly transformed quantile regression method. The comparison was done using the bias of the estimates, AIC and MSE. The results of this analysis revealed that at all quantiles considered (0.05, 0.25, 0.5, 0.75 and 0.95), the Yeo-Johnson transformed quantile regression performed better than the Manly transformed quantile regression based on the AIC and MSE. The bias of the estimates also indicated that the Yeo-Johnson transformed Quantile regression had better results except at the 25th quantile. The analysis was done in R using Qtools package.

Reference:

Atkinon, A. C. (2020). The Box-Cox Transformation: Review and Extensions. The London school of Economics.
http://eprints.lse.ac.uk/103537/1/StatSciV4.pdf

Benjamin, L. C. and Willard G. M. (2013). Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. Shanghai Arch Psychiatry; 25(1): 55–59. doi: 10.3969/j.issn.1002-0829.2013.01.011

Box, G. and Cox, D. (1964). An Analysis of Transformation. Journal of the Royal Statistical Society B, 26(2): 211–252.

Buchinsky, M. (1995). Changes in the US wage structure 1963-1987: Application of quantile regression. Econometrica: Journal of the Econometric Society, 62(2): 405-458.

Chamberlain, G. (1994). Quantile Regression, Censoring, and the Structure of Wages. In: Sims, C. (ed.), Advances in Econometrics: Sixth World Congress, Volume 1, Econometric Society Monograph.

Dye, S. (2020). Quantile Regression; www.https//towardsdatascience.com/quantile-regression-ff2343c4a03. Accessed 23rd Nov. 2021.

Fitzenberger, Bernd and Wilke, Ralf A. and Zhang, Xuan (2005). A Note on Implementing Box-Cox Quantile Regression (December 2005). ZEW - Centre for European Economic Research Discussion Paper No. 04-061 Available at SSRN: http://dx.doi.org/10.2139/ssrn.604441

Gercia, M. (2016). Qtools: A Collection of Models and Tools for Quantile Inference. The R Journal, 8(2):117-138

Jorge I. Vélez, Juan C. Correa and, Fernando Marmolejo-Ramos(2015). A new approach to the Box–Cox transformation. Appl. Math. Stat., 30 October 2015 | https://doi.org/10.3389/fams.2015.00012

Koenker, R. (2005). Quantile Regression, Cambridge University Press.

Koenker, R. and Hallock, K. F. (2001). Quantile Regression. Journal of Economic Perspective: 15(4), 143–156.

Li P. (2005). Box-Cox Transformations: An Overview. Department of Statistics, University of Connecticut. Available online at: http://tinyurl.com/pli2005

Manly, B. F. (1976). Exponential data transformations. Journal of the Royal Statistical Society: Series D, The Statistician: 25(1), 37-42.

Powell, J. (1991). Estimation of monotonic regression models under quantile restrictions. In: W.Barnett, J.Powell, and G.Tauchen, eds., Nonparametric and semiparametric methods in Econometrics, (Cambridge University Press, New York, NY), 357–384.

Raymaekers, J., Rousseeuw, P.J. (2021). Transforming variables to central normality. Machine Learning https://doi.org/10.1007/s10994-021-05960-5

Sakia R. M. (1992) The Box-Cox transformation technique: a review. Journal of Royal Statistical Society series D: 41:169–78. doi: 10.2307/2348250

Steven J. Staffa, MS, Daniel S. Kohane, MD, and David Zurakowski, MS,(2019). Anesthesia and Analgesia. 128:820–30. DOI: 10.1213/ANE.0000000000004017.

Tukey, J. W. (1957). The comparative anatomy of transformations. Ann Math Stat. 28:602–32. doi: 10.1214/aoms/1177706875

Watthanacheewakul L.(2014). A New Family of Transformations for Lifetime Data. Proceedings of the World Congress on Engineering, Vol1 WCE 2014, July 2-4, London UK.

Yeo, I. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika; 87(4).