 doi:10.20944/preprints202203.0166.v1

Technical Note

A Note on Novel Normal-Power Non-Linear Function

Matthew Ekum

Department of Pure and Applied Mathematics, Lagos State University of Science and Technology, Ikorodu, Lagos, Nigeria; matekum@yahoo.com

Abstract: Regression models are mostly used in all fields of sciences for modelling the relationship between a dependent variable and independent variable(s). The least square method is often used to estimate the parameters in a linear model because it is the best linear unbiased estimator. These estimates can only be reliable if the assumption of normality is satisfied. In some cases the dependent variable might be bimodal and shows a non-linear relationship with the independent variable(s). In this case, a non-linear model should be used. In non-linear model, the standard errors are often obtained by linearizing the nonlinear function around the parameter, assuming central limit theorem. After the linearization, the least square parameter estimates are obtained. It should be noted that the error of the non-linear model is different from that of the transformed linear model. Thus, there is a need to transform back to the original non-linear model. In this note, a novel non-linear function was developed into a non-linear regression model, called Normal-Power model. The least square method was used to estimate the parameter of the transformed model. Its usefulness in regression model was demonstrated using real data of Nigeria Economy-Tourism model.

Keywords: Bimodal Dependent Variable; Normal-Power; Non-Linear; Least Square Estimation, Economy-Tourism Model

0. How to use this Note

This note gives introduction on why a dependent random variable in a regression model needs to be transformed. What happens to the error of the transformed and non-transformed model. The transformed model should be transformed back to the original data. The note also give real application on Nigeria economic and tourism data.

1. Introduction

In Statistics, Linear and nonlinear stochastic models are widely used in many applications to describe the relationship between a dependent variable Y and independent variable(s) X. Such model is linear if the mean of Y is a linear function of the unknown parameters, otherwise it is a nonlinear model [1]. One of the purposes of fitting regression models is to draw inferences on unknown parameters, or their functions, which have some physical interpretation [2]. A key step in the statistical inference on unknown parameters of a model is to compute the standard errors of various estimates. If the statistical model is either nonlinear or the parameter of interest in a linear model is a nonlinear function of the regression parameters, then the approximate standard errors are usually derived by using the first order term in a suitable Taylor's series expansion. Once the approximate standard errors are obtained the Wald type confidence intervals are derived. Such confidence intervals are used very extensively in applications [1].

Under some conditions on the linear model the Wald confidence intervals are accurate when the parameter of interest is a linear function of the regression parameters. However, if the parameter is either a non-linear function of the regression parameters or if the model is a nonlinear model, then they are not necessarily accurate, unless the sample sizes are "very large". Basically the large sample theory confidence intervals are derived by "linearizing" the nonlinear function. This is accomplished by approximating the function by the first

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

order derivative term in the Taylor series expansion of the nonlinear function. In some cases, it is important that the second and higher order terms in the Taylor series expansion are negligible in comparison to the first order term to be accurate [1]. The effect of the second order term is known as the "curvature effect."

Transformations are used to present data on a different scale. The nature of a transformation determines how the scale of the untransformed variable will be affected [3]. In modeling and statistical applications, transformations are often used to improve the compatibility of the data with assumptions underlying a modeling process, to linearize the relation between two variables whose relationship is non-linear, or to modify the range of values of a variable. Transformations can be done to dependent variables, independent variables, or both [4]. Taken in the context of modeling the relationship between a dependent variable Y and independent variable X, there are several motivations for transforming a variable(s). It should be noted that many transformations are borne by the need to specify a relation between Y and X as linear, since linear relationships are generally easier to model than non-linear relationships [1].

Thus, transformations done to Y and X in their originally measured units are merely done for convenience of the modeler, and not because of an underlying "problem" or "need" of the data themselves. Thus, transformations done to the dependent variable Y should be transformed back to the original units when a model is compared to other models, or when the model is presented to the professional community or to the general public. The general procedure for doing this is to linearize the relation between Y and X's in the model, estimate model parameters, and then perform algebraic manipulations of the resulting equation to return Y to its original units [1].

The motivation of this work arises from the fact that some dependent variables in a regression model are not normally distributed. There are many methods in literature on transforming the dependent variable. There are situations where the dependent variables are bimodal, it is often difficult to transform such variables. This function will help to transform a bimodal or skewed-bimodal random variables to normal or symmetric random variables.

In this note, a novel normal-power function is introduced, in which a simple transformation of the dependent variable from normal-power to normal and back to normal-power. The function is then applied to regression model to show its usefulness.

2. Materials and Methods

The normal-power function was adopted from the work of [5]. Given a random variable Y that follows a normal-power$\{logistic\}$ distribution (NPLD). The probability density function (pdf) of Y is given by

$$f_Y(y) = \frac{k\lambda^k}{x(\lambda^k - x^k)\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}\left[ln\left(\frac{y^k}{\lambda^k - y^k}\right) - \mu\right]^2\right\}; \mu, \sigma, k, \lambda > 0; 0 < y < \lambda. \tag{1}$$

The pdf in (1) is bimodal and can be used to model bimodal data [5]. The NPLD is related to normal distribution using a simple transformation with the data that follows a NPLD to a normal distribution.

Theorem 1. *If* $Y \sim NPLD(\mu, \sigma, k, \lambda)$, *a random variable* $W = ln\left(\frac{Y^k}{\lambda^k - Y^k}\right)$ *follows a normal distribution with parameters* μ *and* σ.

Proof of Theorem 1. The proof follows from a simple change of variable formula.

$$f(w) = f(y)\left|\frac{dy}{dw}\right|.$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

Let $w = ln\left(\frac{y^k}{\lambda^k - y^k}\right)$, so that $\frac{dy}{dw} = \frac{\lambda^k - y^k}{k\lambda^k y^{k-1}}$. Thus, $f(w)$ is derived thus

$$f(w) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2\right\}; -\infty \leq w \leq \infty. \tag{2}$$

Equation (2) completes the proof, where $W \sim Normal(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, Equation (2) reduces to standard normal. □

Transforming the support. Note that $y \in [0, \lambda]$. This implies that

$$0 \leq y \leq \lambda. \tag{3}$$

Taking the kth power of (3) gives

$$0 \leq y^k \leq \lambda^k. \tag{4}$$

Dividing (4) by $\lambda^k - y^k$ gives

$$0 \leq \frac{y^k}{\lambda^k - y^k} \leq \frac{\lambda^k}{\lambda^k - y^k}. \tag{5}$$

Taking the log of (5) gives

$$ln(0) \leq ln\left(\frac{y^k}{\lambda^k - y^k}\right) \leq ln\left(\frac{\lambda^k}{\lambda^k - y^k}\right). \tag{6}$$

Note the max $(y) \approx \lambda$, so that (6) becomes

$$-\infty \leq w \leq ln\left(\frac{\lambda^k}{\lambda^k - y^k}\right), \tag{7}$$

$$-\infty \leq w \leq \infty. \tag{8}$$

Thus, $w \in (-\infty, \infty)$. Since W is normally distributed, inferences can be carried out on W in terms of normal distribution.

3. Results

There are situations where in regression model where the dependent variable data is bimodal or skewed. The estimate of the linear regression model will be unreliable. As such, the model is not linear, as can be depicted from a scatter plot. The non-linear function can be used in this case.'

Consider the novel non-linear function given by

$$f(t) = \lambda\left(\frac{e^T}{1+e^T}\right)^{1/k}, \tag{9}$$

where k and λ are known parameters that can be estimated from the data, such that T is a random variable supported on the open interval $-\infty$ to ∞, but $f(t)$ is bounded by 0 and λ [6].

Let T follow the normal distribution. Thus, a linear regression model can be given as

$$t_i = \sum_{j=0}^{p} \beta_j x_{ij} + u_i, \tag{10}$$

where t_i is the dependent variable of the ith observation, x_{ij} is the jth independent variable of the ith observation, u_i is the error term of the ith observation, β_j is the $(j+1)th$ parameter

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

to be estimated, p is the number of independent variables and $i = 1, ..., n$, where n is the number of observations.

So, substituting (10) in place of $t \in T$ in (9) gives the novel non-linear function in terms of linear combination of X as

$$f(x) = \lambda \left(\frac{e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i}}{1 + e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i}} \right)^{1/k}. \tag{11}$$

The function in (11) can then be written in terms of a non-linear regression model as

$$y_i = \lambda \left(\frac{e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i}}{1 + e^{\sum_{j=0}^{p} \beta_j x_{ij} + u_i}} \right)^{1/k}, \tag{12}$$

where y_i is the dependent variable of the ith observation, the values k and λ are known constants and are given in [6].

3.1. Linearizing the Non-Linear Model

In order to estimate the parameters of the non-linear model given in (12), a simple mathematics is done to transform (12) to

$$\frac{y_i^k}{\lambda^k - y_i^k} = exp\left(\sum_{j=0}^{p} \beta_j x_{ij} + u_i \right), \tag{13}$$

where

$$k = \frac{\bar{y}}{\bar{\lambda} - \bar{y}}; \lambda > \bar{y} \tag{14}$$

and

$$\lambda = max(y) + Se(y), \tag{15}$$

where $Se(y)$ is the standard error of y estimated from data [5]. The non linear model in (13) is linearized by taking the log. This gives

$$ln\left(\frac{y_i^k}{\lambda^k - y_i^k} \right) = \sum_{j=0}^{p} \beta_j x_{ij} + u_i. \tag{16}$$

Let

$$w_i = ln\left(\frac{y_i^k}{\lambda^k - y_i^k} \right). \tag{17}$$

The random variable W can be easily generated since the values of k and λ are known. So (16) reduces to

$$w_i = \sum_{j=0}^{p} \beta_j x_{ij} + u_i. \tag{18}$$

Now $W \sim N(\mu, \sigma)$, $w \in W$ and $U \sim N(0, \sigma)$, $u \in U$. The model in (18) is a classical linear model (see equation 2) and can be estimated by different method. In this case, the Least Square is the Best Linear Unbiased Estimator (BLUE). So, the least square estimates for simple linear model are given by

$$\beta_0 = \frac{1}{n} \left(\sum_{i}^{n} w_i - \beta_1 \sum_{i}^{n} x_i \right) \tag{19}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

and

$$\beta_1 = \frac{n\sum_i^n w_i x_i - \sum_i^n w_i \sum_i^n x_i}{n\sum_i^n x_i^2 - \left(\sum_i^n x_i\right)^2}. \tag{20}$$

The multiple linear model written in matrix form is given by

$$W = XB + u, \tag{21}$$

where The parameter estimate is given by

$$B = (X'X)^{-1}X'W. \tag{22}$$

where W represents dependent variables of n observations packed in a n-dimensional vector called the response vector. X represents the independent variables packed into a $n \times p+1$ matrix called the design matrix. (Note the initial column of 1's). B represents the unknown parameters to be estimated packed into a $p+1$-dimensional vector called the slope vector. U represents the errors terms packed into a n-dimensional vector called the error vector [7].

Now that the estimates of the parameters are known, the error of the linear model can then be estimated as

$$u_i = w_i - \hat{w}_i. \tag{23}$$

where

$$\hat{w}_i = \sum_{j=0}^{p} \beta_j x_{ij}. \tag{24}$$

where $\hat{w}_i$ are the predicted values of w_i, $\forall\, i$, $i = 1, ..., n$.

3.2. The Error of the Model

Note that u is the error of W, which is normally distributed. It is quite different from the error of Y. Thus, the coefficient of determination, the standard error and other calculations of error for W is quite different from that of Y. After the linear model estimation, the model of W will need to be transformed back to Y.

The values of y_i is retrieved using this equation

$$\hat{y}_i = \frac{\lambda^k e^{\hat{w}_i}}{1 + e^{\hat{w}_i}} \tag{25}$$

where $\hat{y}$ is the predicted values of y and $\hat{w}$ is the predicted value of w.

Thus, the error of Y is given by

$$e_i = y_i - \hat{y}_i \tag{26}$$

Note that u is normally distributed but e is normal-power$\{logistic\}$ distributed. Note that this novel non-linear model can be used for both single independent variable and multiple independent variables. The mean square errors of the two models can be compared using u_i for the linear model and e_i for the non-linear model. Also, it should be clearly noted that any further study of the error on the non-linear model is based on e_i only and not on u_i.

3.3. Application to Nigeria Economy-Tourism Model

The economy of Nigeria may be dependent on different factors. Let us assume it depends on the number of international tourist arriving Nigeria, provided other factors remain constant as it is assumed in any regression model.

The data is obtained from World Bank Data base on GDP per Capita (in Current USD) (RGDP) and number of tourist arrivals to Nigeria (TOUR). RGDP is a proxy for Nigeria economy and it is the dependent variable, while TOUR is a proxy for tourism development and it represents the independent variable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

The non linear model is given by

$$RGDP_i = \lambda\left(\frac{e^{\beta_0+\beta_1 TOUR_i+u_i}}{1+e^{\beta_0+\beta_1 TOUR_i+u_i}}\right)^{1/k} \tag{27}$$

where $\lambda > max(RGDP)\,\forall\, i$, and k is a function of the average RGDP and λ only, β_0 and β_1 are unknown parameters to be estimated from the model.

The scatter plot showing the relationship between RGDP and TOUR is depicted in Figure 1. The histogram of the dependent variable (RGDP) is depicted on Figure 2.

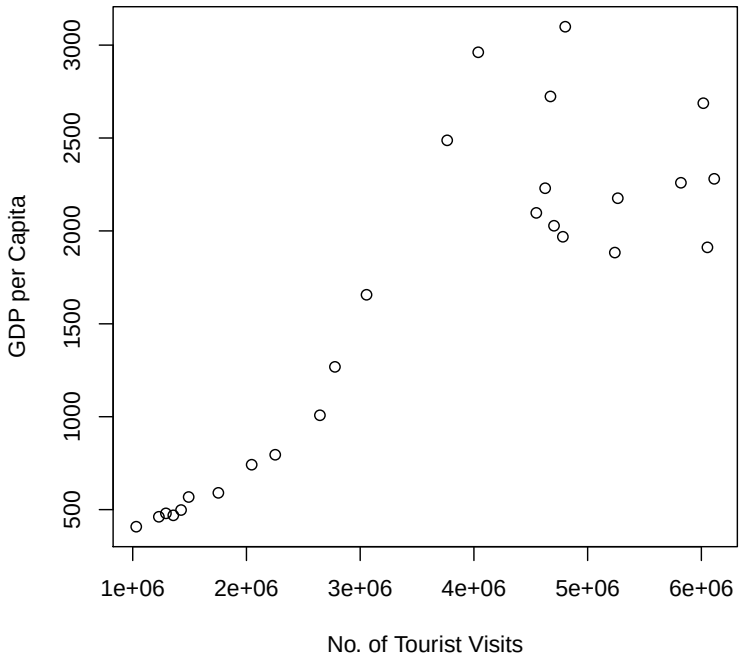

Figure 1. Scatter plot of RGDP and TOUR.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

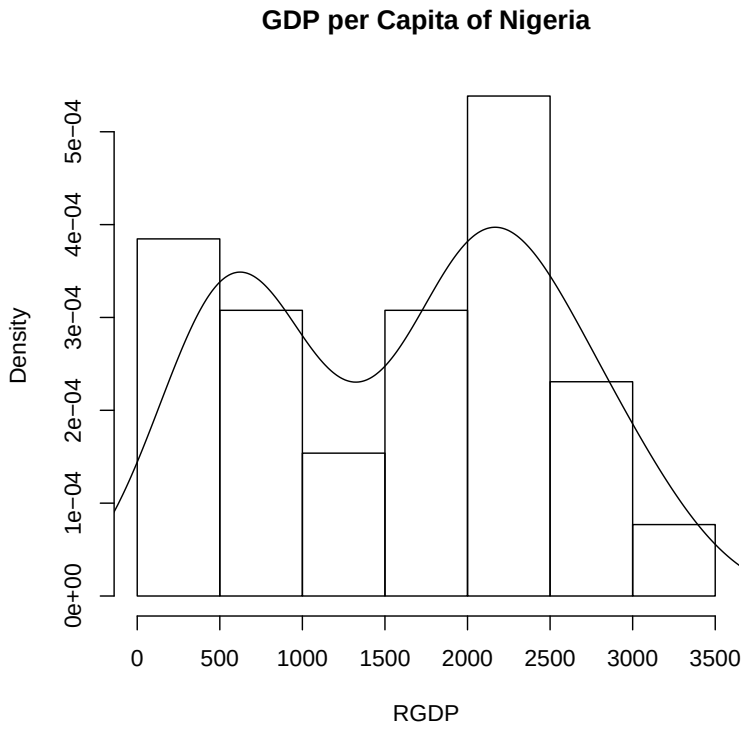

Figure 2. Histogram showing RGDP.

Table 1 shows the normality test of RGDP of Nigeria using Shapiro-Wilk and Kolmogorov-Smirnov tests.

Table 1. Normality test of RGDP.

	Shapiro-Wilk	**Kolmogorov-Smirnov**
Statistic	0.89978	1.00000
P-value	0.01548	0.00000

For the two tests, the null hypothesis states that the data is normally distributed. So, the test is significant if $p < \alpha = 0.05$. If the test is significant, the null hypothesis is rejected and it is concluded that the data is not normally distributed.

3.4. Model Parameter Estimation and Goodness of Fit for the Economy-Tourism Model

The first step in estimating the parameter of the Economy-Tourism model is to first estimate the parameters k and λ from the RGDP data. Recall

$$\hat{k} = \frac{\bar{y}}{\bar{\lambda} - \bar{y}},$$

where

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i \tag{28}$$

and recall

$$\hat{\lambda} = max(y) + Se(y),$$

where

$$Se(y) = \frac{Sd(y)}{\sqrt{n}}, \tag{29}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

where $Sd(Y)$ is the standard deviation of y given by

$$Sd(y) = \sqrt{\frac{1}{n*}\sum_{i}^{n}(y_i - \bar{y})^2}, \tag{30}$$

where $n* = n - 1$ if $n < 30$ and $n* = n$ if $n \geq 30$.
Thus $\hat{\lambda} = 3274.831$ and $\hat{k} = 0.9616$. Since the values of k and λ have been estimated, it is easy to generate the values of w and estimate the linear model using Least Square.

Table 2 shows the parameter estimates and their standard errors in brackets using Least Square and the goodness of fit criteria. The goodness of fit criteria includes log-likelihood, AIC and BIC.

Table 2. Model Parameter Estimates and Goodness of Fit Criteria/Test.

Models	Estimates	P-value	R^2	LogL	AIC	BIC
Linear	β_0 = 18.65 (202.7)	0.9270	0.7589	-195.65	395.3081	397.8243
	β_1 = 0.000445 (0.000051)	0.0000				
Non-Linear	β_0 = -2.281 (0.3480)	0.0000	0.7723	-28.103	64.2063	69.2387
	β_1 = 0.0000006 (0.00000009)	0.0000				

The economy-tourism model is therefore given as

$$\hat{RGDP_i} = 3274.831\left(\frac{e^{-2.281+0.0000006TOUR_i+u_i}}{1+e^{-2.281+0.0000006TOUR_i+u_i}}\right)^{1/0.9616} \tag{31}$$

where $\hat{RGDP}$ is the predicted GDP per Capita for Nigeria using the non-linear Normal-Power model.

The economy-tourism model is fitted on the scatter plot as shown in Figure 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2022 doi:10.20944/preprints202203.0166.v1

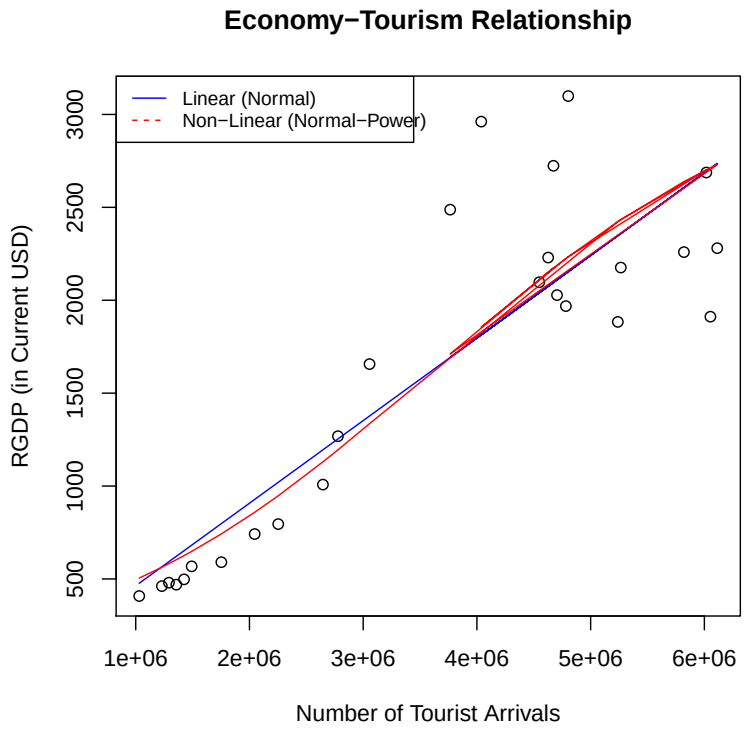

Figure 3. Model fitted on Scatter Plot.

The LRT is carried out to see how better is the Normal-Power model over the normal model. The LRT is given by

$$LRT = -2(\ell_0 - \ell_1), \tag{32}$$

where ℓ_0 is the logL of the linear model and ℓ_1 is the LogL of the non-linear model. The LRT is asymptotically Chi-square distributed with 2 degrees of freedom. The model with ℓ_1 is significantly better than the model with ℓ_0 if LRT is greater than the critical Chi-square value at a certain level of significance, say, $\alpha = 0.05$.
So, substituting the values of LogL of both linear and non-linear models into Equation (32) gives

$$LRT = -2(-195.65 - (-28.103)) = 335.094 \tag{33}$$

where $LRT \sim \chi^2(d)$, where $d = 2$ is the degrees of freedom, which is derived by subtracting the number of parameters of the linear model, 2, from that of the linear model, 4.
The $LRT = 335.094$ is greater than the $\chi^2_{0.05,2} = 5.991$. This shows how much better is the non-linear model than the linear model.

4. Discussion

The study fits a regression model of GDP per capita (RGDP) on the number of tourism arrivals (TOUR) using Nigeria data. Figure 1 shows that the relationship between RGDP and TOUR is not linear. Furthermore, the histogram in Figure 2 also shows that the RGDP is not linear, it has bimodal features. More so, the normality tests in Table 1 confirm that the RGDP is not normally distributed using both Shapiro-Wilk and Kolmogorov-Smirnov tests, because their p-values are less than 0.05.

The estimate of β_0 in the linear model is not significant, which implies that it cannot be relied on, but the estimate of β_1 is significant. On the other hand, the estimates of β_0 and β_1 for the non-linear model are significant ($p < 0.05$). The coefficient of determination (R^2) of the non-linear model (0.7723) is greater than that of the linear model (0.7589) showing

 doi:10.20944/preprints202203.0166.v1

that the non-linear model can predict the RGDP better than the linear model, because the closer R^2 to 1 the better the model. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are goodness of fit criteria for model selection. The closer their values to zero the better the model. Table 2 shows that the non-linear model is a better model with lower AIC and BIC.

Furthermore, the log-likelihood (LogL) is another important criteria for determining a better model. The closer it is to zero (positive or negative), the better the model. Table 2 shows that the LogL of the non-linear model is far better than that of the linear model. To show how much better the non-linear model than the linear model, the Likelihood Ratio Test (LRT) was carried out. The result of the test, which is Chi-square distributed shows that the non-linear model significantly out performed the linear model.

5. Conclusions

This technical note is design to take care of regression model where the dependent variable is bimodal or asymmetric. The novel Normal-Power is a non-linear model proposed to handle such situation. The normal-power model is easily transformed to the linear model to estimate its parameters. The study of the error of the proposed model non-linear model is achieved by transforming back the linear model to its original form. The necessary steps required to accomplish the task was well articulated in this note using real data of Nigeria economy-tourism model.

Economy-Tourism was chosen as an example to validate the novel non-linear model. The linear model assumes that the dependent variable should be normally distributed, but in this case of Economy-Tourism model, it is not normally distributed. The dependent variable has bimodal features. So, the non-linear model was adopted. The Normal-Power Non-linear model fits the data and outperformed the linear model both in predictive power and goodness of fitness. Thus, it is recommended that in regression model, if the dependent variable is bimodal or asymmetric, the normal-power non-linear model should be adopted to fit such model. It can also be used in any situation as substitute in cases where normal model fails.

Funding: This research received no external funding.

Data Availability Statement: The data was harvested from World Bank Database on GDP per Capita (in Current USD) (RGDP) of Nigeria and number of tourist arrivals to Nigeria (TOUR). The data was subsequently deposited on the Research Gate page of the author. The following link is used to publicly archived the dataset analyzed in this note at https://www.researchgate.net/publication/359011132_Economy-Tourism_Model_Data.

Acknowledgments: The author acknowledge the Department of Pure and Applied Mathematics, Lagos State University of Science and Technology, and Department of Mathematics, University of Lagos for allowing me have access to their facilities in their respective statistical laboratory. The author also acknowledge his PhD supervisors, Prof. Adamu, M.O. and Dr. Akarawak, E.E.E. for their academic and moral support during the course of the PhD.

Conflicts of Interest: The author declare no conflict of interest".

Abbreviations

The following abbreviations are used in this manuscript:

NPLD	Normal-Power Logistic Distribution
RGDP	GDP per Capita
TOUR	Number of Tourist Arrivals
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
LRT	Likelihood Ratio Test

 doi:10.20944/preprints202203.0166.v1

References

1. Peddada, S.D. and Haseman, J.K. Analysis of Nonlinear Regression Models: A Cautionary Note. *Formerly Nonlinearity in Biology, Toxicology, and Medicine* **2005**, *3*, 342–352.
2. Chu, J. A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain. *PLOS ONE* **2021**, *16(3)*: e0249037. https://doi.org/10.1371/journal.pone.0249037.
3. Lee, D.K. Data transformation: a focus on the interpretation. *Korean J Anesthesiol* **2020**, *6*, 503–508: https://doi.org/10.4097/kja.20137
4. Wiedermann, W., Li, X. Direction dependence analysis: A framework to test the direction of effects in linear models with an implementation in SPSS. *Behav Res* **2018**, *50*, 1581—1601: https://doi.org/10.3758/s13428-018-1031-x
5. Ekum, M.I., Job, O., Taylor, J.I., Amalare, A.A., Khaleel, M.A. and Ogunsanya, A. S. Normal-Power Function Distribution with Logistic Quantile Function: Properties and Application. *American Journal of Applied Mathematics and Statistics* **2021**, *9(3)*, 90—101.
6. Ekum, M.I., Adeleke, I.A. and Akarawak, E.E. Lambda Upper Bound Distribution: Some Properties and Applications. *Benin Journal of Statistics* **2020**, *3*, 12—40.
7. Ekum, M.I.,Akinmoladun, O.M., Aderele, O.R. and Esan, O.A. Application of Multivariate Analysis on the effects of World Development Indicators on GDP per capita of Nigeria (1981-2013). *International Journal of Science and Technology (IJST)* **2015**, *4(12)*, 254–534.