MODEL SELECTION IN BIVARIATE REGRESSION MODELS

BY

Esemokumo Perewarebo Akpos
E-mail Address: contactpere4good@gmail.com
Department of Statistics, School of Applied Science, Federal Polytechnic Ekewe
Yenagoa, Bayelsa State, Nigeria

Bekesuoyeibo Rebecca
E-mail Address: fzimugha@gmail.com
Department of Statistics, School of Applied Science, Federal Polytechnic Ekewe
Yenagoa, Bayelsa State, Nigeria

Nwobi Anderson Chukwukailo
E-mail Address: andersonwbi@yahoo.com
Department of Statistics, Abia State Polytechnic, Aba Nigeria

ABSTRACT

This study is on model selection in bivariate regression models. Data for this study were collected in CBN Annual Report (various issues), CBN Statistical bulletin (various issues) from 1990 to 2019, which consists of international oil prices (response variable) and unemployment rate (independent variable). Eight regression models; Linear Regression, Quadratic Regression, Cubic Regression, Power Regression, ab-Exponential Regression, Logarithmic Regression, Hyperbolic Regression and Exponential Regression were examined in this study. Five model selection techniques known as; coefficient of determination, standard error of Regression, Akaike Information Criterion, Schwarz Information Criterion, and Hannan-Quinn Information Criterion were used to select the best model. From the analysis, in the overall goodness of fit assessment, the study concluded that the ab-Exponential regression model with Exponential regression performs far better than the other six bivariate regression models employed in this study. Therefore, future researchers should look at a similar work by incorporating other nonlinear bivariate regression models like compound, growth and inverse Regression models to compare results.

Key words: Coefficient of Determination, S.E. of Regression, Akaike Information Criterion, Schwarz Information Criterion, Hannan-Quinn Information Criterion, Bivariate Regression models

Background to the Study

Fitting bivariate regression models to data is normally employed within all fields of science; pharmaceutical and biochemical assay quantification, even though fitting a simple linear model to data seldom arises, because most data tend to follow nonlinear models. Nonlinear models exist, and the choice of selecting the right model for the data is a mixture of experience, knowledge about the underlying process and statistical interpretation of the fitting outcome. It is of paramount important in quantifying the validity of a fit by some measure which discriminates a 'good' from a 'bad' fit. Many researchers usually employ a common measure known as the coefficient of determination R^2 used in linear regression when

conducting calibration experiments for samples to be quantified (Montgomery et al, 2006). Hence, in the linear perspective, this measure is very intuitive as values between 0 and 1 produce an easy interpretation of how much of the variance in the data is explained by the fit. Even though for some time, it has been established that R^2 is an inadequate measure for nonlinear regression, many scientists and researchers still make use of it in studies dealing with nonlinear data analysis (Nagelkerke, 1991; Magee, 1990). According to Juliano and Williams (1987), several initial and older descriptions for R^2 being of no avail in nonlinear fitting had pointed out this issue but have probably fallen into oblivion. This observation might be due to differences in the mathematical background of trained statisticians and researchers who often employ statistical methods but lack detailed statistical insight (Spiess and Neumeyer, 2010).

Having stated that researchers indiscriminately employ R^2 as a means of assessing the validity of a particular model when dealing with nonlinear data fit, it is stated that R^2 is not an optimal choice in a nonlinear regime as the total sum-of-squares (TSS) is not equal to the regression sum-of-squares (REGSS) plus the residual sum-of-squares (RSS), as is the case in linear regression, and hence it lacks the appropriate interpretation. The rationale behind a high occurrence in solely using R^2 values in the validity of nonlinear models could be as a result of researchers not being aware of this misconception.

Even though the use of only R^2 to access the performance of nonlinear data analysis has been discouraged, this study will employ it together with other four model selection techniques known as; Akaike Information Criterion, Schwarz Information Criterion, Hannan-Quinn Information Criterion and standard error of regression for proper interpretation and conclusion.

Literature Review

Hamidian et al (2008) researched on comparison of linear and nonlinear models for estimating brain deformation using finite element method. The study presented finite element computation for brain deformation during craniotomy. The results were used to illustrate the comparison between two mechanical models: linear solid-mechanic model, and non linear finite element model. To this end, the study employed a test sphere as a model of the brain, tetrahedral finite element mesh, two models that described the material property of the brain tissue, and function optimization that optimized the model's parameters by minimizing distance between the resulting deformation and the assumed deformation. Linear and nonlinear model assumed finite and large deformation of the brain after opening the skull respectively. By using the accuracy of the optimization process, the study concluded that the accuracy of nonlinear model was higher but its execution time was six time of the linear model.

Aristizábal-Giraldo et al (2016) carried out a study on a comparison of linear and nonlinear model performance of shia_landslide: a forecasting model for rainfall-induced landslides. The study explained that landslides are one of the main causes of global human and economic losses. The study compared the forecasting performance of linear and nonlinear SHIA_Landslide model. The results obtained for the La Arenosa Catchment during the September 21, 1990 rainstorm showed that the nonlinear SHIA_Landslide replicated more accurately landslides triggered by rainfall features.

Hunt and Maurer (2016) did a work on comparison of linear and nonlinear feedback control of heart rate for treadmill running. The purpose of the study was to compare linear (L) and nonlinear (NL) controllers using quantitative performance measures. Sixteen healthy male subjects participated in the experimental L vs. NL comparison. The linear controller was calculated using a direct analytical design that employed an existing approximate plant model. The nonlinear controller had the same linear component, but it was augmented using static plant-nonlinearity compensation. At moderate-to-vigorous intensities, no significant differences were found between the linear and nonlinear controllers in mean RMS tracking error (2.34 vs. 2.25 bpm [L vs. NL], p=0.26) and average control signal power (51.7 vs. 60.8 $\times 10^{-4}$ m^2/s^2, p=0.16), but dispersion of the latter was substantially higher for NL (range 45.2 to 56.8 vs. 30.7 to 108.7 $\times 10^{-4}$ m^2/s^2, L vs. NL). At low speed, RMS tracking errors were similar, but average control signal power was substantially and significantly higher for NL (28.1 vs. 138.7 $\times 10^{-4}$ m^2/s^2 [L vs. NL], p<0.001). The performance outcomes for linear and nonlinear control were not significantly different for moderate-to-vigorous intensities, but NL control was overly sensitive at low running speed. Accurate, stable and robust overall performance was achieved for all 16 subjects with the linear controller.

Spiess and Neumeyer (2010) worked on an evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. The intensive simulation approach undermined previous observations and emphasized the extremely low performance of R^2 as a basis for model validity and performance when applied to pharmacological/biochemical nonlinear data. With the 'true' model having up to 500 times more strength of evidence based on Akaike weights, this was only reflected in the third to fifth decimal place of R^2. In addition, even the bias-corrected R^2_{adj} exhibited an extreme bias to higher parameterized models. The bias-corrected AIC and also BIC performed significantly better in this respect. The study concluded that researchers and reviewers should be aware that R^2 is inappropriate when used for demonstrating the performance or validity of a certain nonlinear model. It should ideally be expunged from scientific literature dealing with nonlinear model fitting or at least be supplemented with other methods such as AIC or BIC or used in context to other models in question.

Scarneciu et al (2017) worked on Comparison of Linear and Non-linear Regression Analysis to determine pulmonary pressure in hyperthyroidism. The study aimed at assessing the incidence of pulmonary hypertension (PH) at newly diagnosed hyperthyroid patients and at finding a simple model showing the complex functional relation between pulmonary hypertension in hyperthyroidism and the factors causing it. The 53 hyperthyroid patients (H-group) were evaluated mainly by using an echocardiographical method and compared with 35 euthyroid (E-group) and 25 healthy people (C-group). In order to identify the factors causing pulmonary hypertension, the statistical method of comparing the values of arithmetical means was employed. By applying the linear regression method described by a first-degree equation the line of regression (linear model) was determined; by applying the non-linear regression method described by a second degree equation, a parabola-type curve of regression (non-linear or polynomial model) was determined. The study made the comparison and the validation of these two models by calculating the determination coefficient (criterion 1), the comparison of residuals (criterion 2), application of AIC criterion (criterion 3) and use of F-test (criterion 4). The result of the study revealed that from the H-group, 47% have pulmonary hypertension completely reversible when obtaining euthyroidism. The factors causing pulmonary hypertension were identified: previously known- level of free thyroxin, pulmonary vascular resistance, cardiac output; new factors identified in the study- pre-treatment period, age, systolic blood pressure. According to the four criteria and to the

clinical judgment, the study considered that the polynomial model (graphically parabola-type) was better than the linear one. The study thereby concluded that the better model showing the functional relation between the pulmonary hypertension in hyperthyroidism and the factors identified in the study was given by a polynomial equation of second degree where the parabola was its graphical representation.

Methodology

Regression Analysis

Regression analysis is a statistical technique that express mathematically the relationship between two or more quantitative variables such that one variable (the dependent variable) can be predicted from the other or others (independent variables). It is very useful in predicting or forecasting. It can also be used to examine the effects that some variables exert on others. It may be simple linear, multiple linear or non linear. The study is limited to bivariate regression models.

Regression Models

Fitted Linear Regression Equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ … (1)

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad \ldots \quad (2)$$

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad \ldots \quad (3)$$

Linear correlation coefficient is given by;

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}} \quad \ldots \quad (4)$$

Coefficient of determination is given by;

$$R^2 = r_{xy}^2 \quad \ldots \quad (5)$$

Fitted Quadratic Regression Equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ … (6)

System of equations to find $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ is shown in Equation (7)

$$\left.\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i + \hat{\beta}_2 \sum x_i^2 &= \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 + \hat{\beta}_2 \sum x_i^3 &= \sum x_i y_i \\ \hat{\beta}_0 \sum x_i^2 + \hat{\beta}_1 \sum x_i^3 + \hat{\beta}_2 \sum x_i^4 &= \sum x_i^2 y_i \end{aligned}\right\} \quad \ldots \quad (7)$$

Correlation coefficient is given by;

$$R = \sqrt{1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \bar{y}\right)^2}} \quad \ldots \quad (8)$$

Fitted Cubic Regression Equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3$ … (9)

System of equations to find $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ is shown in Equation (10)

$$\left.\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i + \hat{\beta}_2 \sum x_i^2 + \hat{\beta}_3 \sum x_i^3 &= \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 + \hat{\beta}_2 \sum x_i^3 + \hat{\beta}_3 \sum x_i^4 &= \sum x_i y_i \\ \hat{\beta}_0 \sum x_i^2 + \hat{\beta}_1 \sum x_i^3 + \hat{\beta}_2 \sum x_i^4 + \hat{\beta}_3 \sum x_i^5 &= \sum x_i^2 y_i \\ \hat{\beta}_0 \sum x_i^3 + \hat{\beta}_1 \sum x_i^4 + \hat{\beta}_2 \sum x_i^5 + \hat{\beta}_3 \sum x_i^6 &= \sum x_i^3 y_i \end{aligned}\right\} \quad \dots \quad (10)$$

Correlation coefficient for cubic regression is the same with Equation (8)

Fitted Power Regression Equation: $\hat{y} = \hat{\beta}_0 x^{\hat{\beta}_1}$ … (11)

$$\hat{\beta}_1 = \frac{n\sum(\ln x_i \ln y_i) - \sum \ln x_i \sum \ln y_i}{n\sum \ln^2 x_i - (\sum \ln x_i)^2} \quad \dots \quad (12)$$

$$\hat{\beta}_0 = \exp\left(\frac{1}{n}\sum \ln y_i - \frac{b}{n}\sum \ln x_i\right) \quad \dots \quad (13)$$

Correlation coefficient is the same with Equation (8)

Fitted $\hat{\beta}_0\hat{\beta}_1$ - Exponential Regression Equation: $\hat{y} = \hat{\beta}_0\hat{\beta}_1^x$ … (14)

$$\hat{\beta}_1 = \exp\frac{n\sum x_i \ln y_i - \sum x_i \sum \ln y_i}{n\sum x_i^2 - (\sum x_i)^2} \quad \dots \quad (15)$$

$$\hat{\beta}_0 = \exp\left(\frac{1}{n}\sum \ln y_i - \frac{\ln b}{n}\sum x_i\right) \quad \dots \quad (16)$$

Correlation coefficient is the same with Equation (8)

Fitted Hyperbolic Regression Equation: $\hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{x}$ … (17)

$$\hat{\beta}_1 = \frac{n\sum \frac{y_i}{x_i} - \sum \frac{1}{x_i}\sum y_i}{n\sum \frac{1}{x_i^2} - \left(\sum \frac{1}{x_i}\right)^2} \quad \dots \quad (18)$$

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \frac{b}{n}\sum \frac{1}{x_i} \quad \dots \quad (19)$$

Correlation coefficient is the same with Equation (8)

Fitted Logarithmic Regression Equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x$ … (20)

$$\hat{\beta}_1 = \frac{n\sum(y_i \ln x_i) - \sum \ln x_i \sum y_i}{n\sum \ln^2 x_i - (\sum \ln x_i)^2} \quad \dots \quad (21)$$

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \frac{b}{n}\sum \ln x_i \quad \ldots \quad (22)$$

Correlation coefficient is the same with Equation (8)

Fitted Exponential Regression Equation: $\hat{y} = e^{\hat{\beta}_0 + \hat{\beta}_1 x}$ … (23)

$$\hat{\beta}_1 = \frac{n\sum x_i \ln y_i - \sum x_i \sum \ln y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad \ldots \quad (24)$$

$$\hat{\beta}_0 = \frac{1}{n}\sum \ln y_i - \frac{b}{n}\sum x_i \quad \ldots \quad (25)$$

Correlation coefficient is the same with Equation (8)

Akaike Information Criterion (AIC)

The Akaike's information criterion AIC (Akaike, 1974) is a measure of the goodness of fit of an estimated statistical model and can also be used for model selection. Thus, the AIC is defined as;

$$\mathrm{AIC} = \mathrm{e}^{\frac{2k}{n}} \frac{\sum \hat{u}_i^2}{n} = \mathrm{e}^{\frac{2k}{n}} \frac{\mathrm{RSS}}{n} \quad \ldots \quad (26)$$

where k is the number of regressors (including the intercept) and n is the number of observations. For mathematical convenience, Equation (26) is written as;

$$\ln(\mathrm{AIC}) = \left(\frac{2k}{n}\right) + \ln\left(\frac{\mathrm{RSS}}{n}\right) \quad \ldots \quad (27)$$

where ln (AIC) = natural log of AIC and $\frac{2k}{n}$ = penalty factor.

Schwarz Information Criterion (SIC)

Schwarz Information Criterion SIC (Schwarz, 1978) is a measure of the goodness of fit of an estimated statistical model and can also be used for model selection. It is defined as

$$SIC = n^{\frac{k}{n}} \frac{\sum \hat{u}_i^2}{n} = n^{\frac{k}{n}} \frac{RSS}{n} \quad \ldots \quad (28)$$

Transforming Equation (28) in natural logarithm form, it becomes (See Equation (29));

$$\ln(SIC) = \frac{k}{n}\ln(n) + \ln\left(\frac{RSS}{n}\right) \quad \ldots \quad (29)$$

where $\frac{k}{n}\ln(n)$ is the penalty factor.

Hannan-Quinn Information Criterion (HQIC)

The Hannan-Quinn Information Criterion HQIC (Hannan and Quinn, 1979) is a measure of the goodness of fit of an estimated statistical model and is often employed as a criterion for model selection. It is defined as

$$HQIC = n\ln\frac{RSS}{n} + 2k\ln(\ln n) \quad \ldots \quad (30)$$

Where n is the number of observations, k is the number of model parameters. RSS is the residual sum of squares that result from the statistical model.

For model comparison, the model with the lowest AIC, SIC, HQIC score is preferred.

Data Analysis

Data used for this study is secondary obtained from CBN Annual Report (various issues), CBN Statistical bulletin (various issues). The data for the period of 30 years on international oil prices and unemployment rate are shown in Table 1.

Table 1: Data on International Oil Prices (IOP) (y_i) in $ and Unemployment Rate (UNR) (x_i)

Year	y_i	x_i	Year	y_i	x_i	Year	y_i	x_i
1990	22.26	3.50	2000	27.60	4.20	2010	77.38	23.9
1991	18.62	3.10	2001	23.12	3.00	2011	107.46	24.0
1992	18.44	3.40	2002	24.36	14.8	2012	109.45	23.0
1993	16.33	2.70	2003	28.10	13.4	2013	105.87	23.5
1994	15.53	2.00	2004	36.05	11.9	2014	96.29	22.0
1995	16.86	1.80	2005	50.59	14.6	2015	37.48	20.0
1996	20.29	3.40	2006	61.00	12.7	2016	38.37	14.2
1997	18.86	3.20	2007	69.04	14.9	2017	47.95	18.8
1998	12.28	3.10	2008	94.10	19.7	2018	64.90	22.6
1999	17.44	4.70	2009	60.86	21.4	2019	57.05	23.1

CBN Annual Report (various issues), CBN Statistical bulletin (various issues)

Table 2: Summary Result of Bivariate Regression Models

Model Form	AIC	SIC	HQIC	R^2	S.E. of Regression
Linear Regression	8.5095	8.6029	8.5393	0.7338	16.5044
Quadratic Regression	8.4597	8.5998	8.5045	0.7631	15.8568
Cubic Regression	8.4937	8.6805	8.5534	0.7707	15.8967
Power Regression	0.6153	0.7087	0.6452	0.7926	0.3187
ab-Exponential Regression	0.3967	0.4901	0.4266	0.8334	0.2857
Logarithmic Regression	8.8063	8.8997	8.8362	0.6418	19.1448
Hyperbolic Regression	9.1302	9.2236	9.1600	0.5048	22.5102
Exponential Regression	0.3967	0.4901	0.4266	0.8334	0.2857

Source*: *E-view software

Looking at the summarized results in Table 2, it can be observed that the ab-Exponential regression model with Exponential regression has the highest coefficient of determination (0.8334) with the lowest AIC (0.3967), SIC (0.4901), HQIC (0.4266) and standard error of regression (0.2857), which makes it the best model with respect to the data used in this study. The next to ab-Exponential Regression model and Exponential Regression model is power

regression model which has a coefficient of determination of 0.7926 with the AIC (0.6153), SIC (0.7087), HQIC (0.6452) and standard error of regression (0.3187). It is clear from the result that the hyperbolic regression model is the least performed model.

Conclusion

From the analysis, in the overall goodness of fit assessment, the study concluded that the ab-Exponential regression model with Exponential regression performs far better than the other six bivariate regression models employed in this study. Therefore, future researchers should look at a similar work by incorporating other nonlinear bivariate regression models like compound, growth and inverse Regression models to compare results.

REFERENCES

Akaike, H. (1974), *"A new look at the statistical model identification"* (PDF), IEEE Transactions on Automatic Control 19 (6): 716–723, *doi*:*10.1109/TAC.1974.1100705*, *MR 042371*

Aristizábal-Giraldo, E. V., Vélez-Upegui, J. I., and Martínez-Carvaja, H. E. (2016). A comparison of linear and nonlinear model performance of shia_landslide: a forecasting model for rainfall-induced landslides. Revista Facultad de Ingeniería, No. 80, pp. 74-88, 2016

Hamidian, H., Soltanian-Zadeh, H., Akhondi-Asl, A., Faraji-Dana, R. (2008). Comparison of Linear and Nonlinear Models for Estimating Brain Deformation Using Finite Element Method. In: Sarbazi-Azad H., Parhami B., Miremadi SG., Hessabi S. (eds) Advances in Computer Science and Engineering. CSICC 2008. Communications in Computer and Information Science, vol 6. Springer, Berlin, Heidelberg

Hannan, E. J., and Quinn, B. G. (1979). "The Determination of the order of an autoregression", *Journal of the Royal Statistical Society, Series B*, 41: 190–195.

Hunt, K. J. and Maurer, R.R. (2016). Comparison of linear and nonlinear feedback control of heart rate for treadmill running. Systems Science & Control Engineering, 4:1, 87-98, DOI: 10.1080/21642583.2016.1179139

Juliano, S. A., Williams, F.M. (1987). A comparison of methods for estimating the functional response parameters of the random predator equation. J Anim Ecol. 1987;56:641–653. doi: 10.2307/5074.

Magee L. (1990). R^2 measures based on Wald and likelihood ratio joint significance tests. Amer Stat. 1990; 44:250–253. doi: 10.2307/2685352.

Montgomery, D.C., Peck, E.A., Vining, G.G. (2006). Introduction to Linear Regression Analysis. Wiley & Sons, Hoboken; 2006.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. Biometrika. 1991;78:691–692. doi: 10.1093/biomet/78.3.691.

Scarneciu, C.C., Sangeorzan, L., Rus, H., Scarneciu, V.D., Varciu, M.S. Andreescu, O. and Scarneciu, I. (2017). Comparison of Linear and Non-linear Regression Analysis to Determine Pulmonary Pressure in Hyperthyroidism. Pak J Med Sci. 2017 Jan-Feb; 33(1): 111–120.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6, 461-464

Spiess, A. and Neumeyer, N. (2010). An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC Pharmacol. V.10; 2010