DOI: 10.21276/sjpms.2017.4.1.3

Scholars Journal of Physics, Mathematics and Statistics **ISSN 2393-8056 (Print)**
Sch. J. Phys. Math. Stat. 2017; 4(1):17-22 **ISSN 2393-8064 (Online)**
©Scholars Academic and Scientific Publishers (SAS Publishers)
(An International Publisher for Academic and Scientific Resources)

Multiple Linear Regression using Centre Mean and Actual Value – A Comparative Approach

Iwok IA, Akpan NP
Department of Mathematics/Statistics, Faculty of Science, University of Port Harcourt, Nigeria

***Corresponding Author:**
Iwok IA
Email: ibywok@yahoo.com

Abstract: This work compared the performance of a multiple linear regression using the actual and centre mean values. The comparison was based on error analysis on both methods using different comparative Statistics. Nigerian gross domestic product (GDP) data was used as a case study. The analysis employed the ordinary least squares method. The result showed that the centre mean value regression provided better estimates than the actual value regression. Hence, centralized values were recommended in multiple regression problems.

Keywords: Multiple Regression, Ordinary Least Squares, Error Statistics and Coefficient of Determination.

INTRODUCTION

Linear regression is one of the most commonly used techniques for investigating the relationship between two quantitative variables.

In all aspects of life, there is a relationship that exists between two or more sets of variables. This relationship can either be positive or negative and strong or weak. The strength of the relationship can be quantified by correlation analysis while regression analysis expresses the relationship in the form of an equation.

The earliest form of regression was the method of Least squares which was published by Legendre in 1805[11] and by Gauss in 1809[10]. Legendre and Gauss applied the method to the problem of determining from astronomical observations, the orbits of bodies about the sun. Gauss published a further development of the theory of Least squares in 1821 including a version of Gauss-Markov theorem. The term *regression* was coined by Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that heights of tall descendants tend to regress toward a normal average (a phenomenon also known as regression toward the mean). Galton regression had only biological meaning but his work was later extended by Yule and Pearson to a more general statistical context.

In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. However, this assumption was weakened by Fisher in his works in 1925[8]. Fisher assumed that the conditional distribution of the response variable is Gaussian but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation in 1821[9].

Regression methods continued to be an area of active research. In recent times, new methods have been developed for robust regression. Some of such methods are: regression involving correlated responses such as Time series and Growth curves; regression in which the predictors or response variables are curves, images, graphs or other complex data objects; regression methods that accommodate various types of missing data; regression in which the predictor variables are measured with error; regression with more predictor variables than observations and casual inference with regression etc.

According to John [1], regression analysis helps one to study how the typical value of the dependent variable (criterion variable) changes when any of the independent variables is varied, while other independent variables are held fixed. The analysis is widely used for prediction and forecasting. Many techniques for carrying out regression have been developed. Familiar methods are the parametric linear and ordinary least square regression where the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. The non-parametric regression refers to the technique that allows the regression function to lie in a specified set of function. In a narrower sense, regression may refer specifically to the estimation of continuous response variables as opposed to the discrete response variables used in classification.

Williams [2] used multiple linear regression as a tool for linking course work and teacher certification outcomes, and the study was examined in two separate cases. The first case examined data from a smaller private university and the second case examined data from a larger public university. Confounding variables, bivariate correlations, beta weights and statistical assumptions were the statistical considerations. The purpose of her study was to test the value of multiple linear regression as a quantitative method for connecting pre service teacher characteristics to subsequent outcomes. She noted that multiple linear regression is capable of providing data-driven conclusions regarding the relationships between individual components of teacher preparation program and initial certification. In conclusion, she indicated that multiple linear regression can provide meaningful program evaluation information when examining teacher preparation programs where fewer sections of courses are offered.

Rowan *et al* [3] used prospect data to carry out the congressionally mandated study of Educational Growth Opportunity to test and compare models estimating teacher impact on k-12 outcomes using regression analysis. A 3 level nested ANOVA model, adjusted covariate model, a gain score model and a cross-classified model were employed. They also examined the magnitude and stability of teacher impact on k-12 outcomes by exploring variables and determining which variables accounted for classroom–to-classroom differences. They further discussed ways in which results from k-12 outcomes could be used to improve teaching methodology. Rowan *et al* [3] reported 60-61% of reliable variance in reading and 52-72% of the variance using cross-classified model.

Siers [4] studied the relationship between socio-economic status, attendance rates per pupil expenditures, teacher qualifications and 'on-time' educational attainment rates within the state of Virginia using multiple linear regression. He examined the possible predicting abilities of socio-economic status of highly qualified teachers and attendance rates for 'on-time' educational attainment in the state of Virginia. The study also focused on comparing Appalachian school division of Virginia and non-Appalachian school divisions for each of the variables. To address his first purpose, a stepwise multiple linear regression analysis on the variables was conducted. A general linear model repeated measures ANOVA was used to address the second purpose. The result showed that socio-economic status rates are independent and was significantly able to predict educational attainment rates. Socio-economic status was found to be significantly higher in the Appalachian division than in the non-Appalachian large school division. Teacher qualification rates were found to be significantly higher in the Appalachian divisions than the non-Appalachian division of similar size. 'On-time' educational attainment rates were found to be higher in the Appalachian school division than in all three classifications of the non-Appalachian divisions.

The comparative analysis of this work shall be based on the Gross Domestic Product (GDP) data. The GDP of a country is one of the main indicators used to measure the performance of a country's economy. It can be defined as the total value of goods and services produced within the borders of a country during a specific period (usually a year or a quarter).

According to Benedict [5], the Nigerian economy with an enterprising population and a wealth of natural resources, offers tremendous potential for economic growth. However, poor economic policy, political instability and an overreliance on oil exports has created a severe structural problem in the economy. In Nigeria, oil and agriculture hitherto remains the basic economic activity for the majority of Nigerians, employing roughly 70% of labour force and accounting for 36.3 % of GDP in 2004. Crop yields have not kept pace with the average population growth of 2.5 (2001-2005 average) and must import most of its food. In 1992, real GDP grew at only 4.1 while the large government deficits 10% of GDP in 1992 continued to expand.

In terms of Agriculture, it is an important sector of Nigeria's economy engaging about one-third of the labour force. Agriculture holdings are generally small and scattered; farming is often the subsistence variety, characterized by simple tools and shifting cultivation. Despite an abundant water supply, a favorable climate and wide areas of arable land, productivity is restricted, owing to low soil fertility in many areas and inefficient methods of cultivation. Agriculture contributed 26% to GDP in 2003. In 2004, agricultural exports totaled $486.7 million, while agricultural imports exceeded $2.2 billion.

The industry accounted for 30.5% of GDP in 2004, mostly in the oil sector, and experienced 1.8% growth that year. Due to high costs of production that results from inadequate infrastructure, Nigeria's manufacturing capacity utilization remains low. An estimated 10% of labour force is employed in industrial sector. Nigeria is the eleventh-largest producer of oil in the world, and first in Africa. The oil sector supplies 95% of foreign exchange earnings and 90% of total exports.

The distribution of consumer goods is affected largely by a complex network of intermediary traders, who extend the network of distribution and often break down products into very small units for delivery to the illuminate consumer. Domestic commerce is limited by poor infrastructure, widespread fraud and corruption and shortages of fuel that are exacerbated by illegal smuggling of gasoline across Nigeria's border. The economy is still primarily cash based. Nigeria's export have been on a dramatic upswing. Exports in 2006 were 90% dominated by crude oil. Liquified Natural Gas accounted for 8.1% of exports in 2004. Cocoa is the largest agricultural export. Leading imports are machinery, chemicals, transportation equipment, manufactured goods and food. Despite all these resources, the rate of unemployment in Nigeria is still high. Ojimadu [6] reviewed the impact of unemployment on the growth rate of Nigeria's economy from 2000-2008. He highlighted the effect of unemployment on the Gross Domestic Product (GDP) of Nigeria which served as a measure of the Nigerian economy. The ordinary Least Square method was used to regress the Nigerian unemployment rate on the Gross Domestic Product of Nigeria for the period of 2000-2008. The results he obtained showed that the rate of unemployment in Nigeria has been a major problem faced by the Nigerian Government and it contributes negatively to the growth of the Nigerian economy. It also showed that the Government bears the cost of unemployment which is manifest in high cost of security gadgets and cost of drugs (HIV retroviral drugs) etc.

Echambadi and Hess [7] worked on the centre mean covariance structure of variables. They found that the cross product term in moderated regression may be collinear with its constituent parts; making it difficult to detect main, simple and interaction effect. The result showed that mean centering can reduce the covariance between linear and the interaction terms. They analytically proved that mean centering neither changes the computational precision of parameters, the sampling accuracy of main-effects, simple effects, interaction effects nor the coefficient of determination. They also showed that the determinant of cross product matrix is identical for un-centered and mean centered data. It was added that researchers using moderated regression models should not mean center in a specious attempt to migrate collinearity between linear and interaction terms.

This work uses the GDP as the basis for the comparison of the two models: Centre mean and Actual value models.

METHODOLOGY

The method employed in the analysis of the data set is the ordinary least square technique (OLS) which is the most widely used method for multiple linear regression. The two (2) different models considered are the Centre mean and Actual value models.

The Actual Value Multiple Regression Model

The actual value multiple regression model of p variables for a sample of size n is given as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \quad ; \quad i = 1,2,\ldots,n.$$

In matrix form, the n observations can be represented as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{pmatrix}$$

which can be compressed in matrix form as

$$\boldsymbol{y} = \boldsymbol{x}\,\boldsymbol{\beta} + \boldsymbol{e}$$

with the following assumptions:

1. $\boldsymbol{E}[\boldsymbol{e}] = \boldsymbol{0} \Rightarrow \boldsymbol{E}[\boldsymbol{y}] = \boldsymbol{x}\,\boldsymbol{\beta}$
2. $\boldsymbol{cov}[\boldsymbol{e}] = \boldsymbol{\sigma^2 I} \Rightarrow \boldsymbol{cov}[\boldsymbol{y}] = \boldsymbol{\sigma^2 I}$

where

y_i's are the dependent variables
x_i's are the independent variables
β_i's are the model parameters
e_i's are the error terms

The Centre Mean Regression Model

If the observations are centralized, the model becomes

$$y_i = \beta_1(x_{i1} - \overline{x_1}) + \beta_2(x_{i2} - \overline{x_2}) + \cdots + \beta_p(x_{ip} - \overline{x_p}) + e_i \quad ; \quad i = 1,2,\ldots,n.$$

This can be represented in matrix form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11}-\overline{x_1} & x_{12}-\overline{x_2} & \cdots & x_{1p}-\overline{x_p} \\ 1 & x_{21}-\overline{x_1} & x_{22}-\overline{x_2} & \cdots & x_{2p}-\overline{x_p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1}-\overline{x_1} & x_{n2}-\overline{x_2} & \cdots & x_{np}-\overline{x_p} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{pmatrix}$$

which can be expressed as:

$$\boldsymbol{y}^* = \boldsymbol{x}^*\, \boldsymbol{\beta}^* + \boldsymbol{e}^*$$

Parameter Estimation

Let the parameter estimates be denoted by $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}$ and the estimated y_i be denoted by $\widehat{y_i}$; where

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}x_{i1} + \widehat{\beta_2}x_{i2} + \cdots + \widehat{\beta_p}x_{ip}$$

Thus, by least squares principle, we seek for $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}$ that minimizes the sum of squares error (SSE):

$$SSE = \sum_{i=1}^{n} \widehat{\varepsilon_i^2} = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

$$= \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_{i1} - \widehat{\beta_2}x_{i2} - \cdots - \widehat{\beta_p}x_{ip}\right)^2$$

The value of $\widehat{\boldsymbol{\beta}} = \left(\, \widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}\, \right)'$ that minimizes the SSE is given (in matrix form) by:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$$

where $\boldsymbol{x}'\boldsymbol{x}$ is assumed to be non-singular.

Similarly, for the centralized data, we have

$$\widehat{\boldsymbol{\beta}^*} = (\boldsymbol{x}^{*\prime}\boldsymbol{x}^*)^{-1}\boldsymbol{x}^{*\prime}\boldsymbol{y}^*$$

Coefficient of Determination

This is the proportion of the total variation in the y's that can be attributed to the regression on the x's. It is given by

$$R^2 = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{x}'\boldsymbol{y} - \boldsymbol{n}\bar{\boldsymbol{y}}^2}{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{n}\bar{\boldsymbol{y}}^2}$$

The positive square root R of R^2 is the multiple correlation.

Test for Overall Regression

The $F-$ test for the overall regression is given by

$$F = \left(\frac{n-p-1}{p}\right)\left(\frac{R^2}{1-R^2}\right)$$

Basis of Comparison

The two competing models (Actual Value and Centre Mean Model) shall be compared based on the following statistics:

(1) The Mean Square Error (MSE) given as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{Y}_i - \widehat{Y}_i\right)^2$$

(2) The mean absolute error (MAE), given as

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left| \mathrm{Y}_i - \widehat{Y}_i \right|$$

(3) The mean absolute percentage error (MAPE) given as

$$MAPE = \left[\frac{1}{N}\sum_{i-1}^{N}\left|\frac{\mathrm{Y}_i - \widehat{Y}_i}{\mathrm{Y}_i}\right|\right] \times 100$$

where Y_i 's are the observed values and $\widehat{Y}_i$'s are the estimated values.

The smaller the value of these statistics, the more efficient the model is.

(4) Coefficient of Determination

DATA ANALYSIS

The data used in this study was obtained from the Nigeria statistical bulletin with the following variables: Nigeria's Gross Domestic Product (Y), Agriculture (X_1), Industry(X_2), Building and Construction (X_3), and Wholesale and Retail (X_4). The data span between 1984-2016. The software used for the analysis is Minitab 16.

Actual Value Regression

Table 1: Minitab Software Output of the Actual Value Regression

Predictor	Coef.	S.E. Coef.	T	P
Constant	-29.13	21.94	-1.33	0.005
Agriculture	0.704	0.061	11.54	0.031
Industry	0.073	0.015	4.87	0.000
Building & Construction	0.633	1.459	0.43	0.000
Wholesale & Retail Trade	0.256	0.119	2.15	0.000

R-Sq = 73.8%   R-Sq(adj) = 73.6%

Table 2: Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	5440413893	1360103473	195860.94	0.000
Residual Error	28	194438	6944		
Total	32	5440608331			

Using the output from the Minitab, the regression model for the actual value

$$y_i = -29.13 + 0.704x_{i1} + 0.073x_{i2} + 0.633x_{i3} + 0.256x_{i4} + e_i \quad ; \quad i = 1,2,\dots,n.$$

Centre Mean Value Regression

Table 3: Minitab Software Output for Centre Mean value regression

Predictor	Coef	S.E.Coef.	T	P
Constant	9865.19	14.51	680.06	0.542
Centre Agric	0.023	0.011	2.091	0.000
Centre Industry	0.005	0.012	0.417	0.000
Centre Building	0.551	0.323	1.706	0.003
Centre Wholesale	0.222	0.211	1.052	0.000

R-Sq = 97.7%   R-Sq(adj) = 97.8%

Table 4: Analysis of Variance

Source	DF	SS	MS	F	p
Regression	4	5440413893	1360103473	195860.94	0.000
Residual Error	28	194438	6944		
Total	32	5440608331			

The regression model for the centre mean is

$$y_i = 9865.19 + 0.023x_{i1} + 0.005x_{i2} + 0.551x_{i3} + 0.222x_{i4} + e_i$$

Table 5: Comparison Table between the Centre Mean Value and the Actual Value Regression

Regression Type	MSE	MAE	MAPE	R^2
Actual Value Regression	7.34	4.15	21.61%	73.8%
Centre Mean Value Regression	3.13	2.61	9.32%	97.7%

RESULT AND CONCLUSION

From table 5 above, the different error statistics (MSE, MAE and MAPE) used show that the errors incurred by using centre mean value regression are less than that incurred by the actual value regression. This simply indicates that the centre mean value regression provides better estimates than the actual value regression. The coefficients of determination R^2 calculated for both regression type support this efficiency of centre mean value regression over the actual value regression. The R^2 measures the percentage of the variation in Y explained by the entire regression equation. Indeed, the centre mean value regression stands tall comparatively. It is therefore pertinent that in multiple regression analysis, there is greater need for the values involved to be centralized before any analysis is carried out. This would help to minimize the errors that would be incurred by using the raw values; thereby improving the estimates of these values.

REFERENCES

1. John SH. Illusions in Regression Analysis. International Journal of Forecasting (forthcoming). 2013;28(3):689.
2. Williams C. Journal of case studies in Accreditation and Assessment. Worldmark Encyclopedia of Nations. 2012;64-78.
3. Rowan K, Correnti E, Miller J. CPRE Research Report Series RR-05 consortium for policy Research in Education, 2002.
4. Siers MN. A Descriptive Statististical Analysis of the Relationship between Socioeconomic Status, Attendance Rates, Per pupil Expenditures, Teacher Qualifications and On-time Educational, Attainment Rates within the State of Virginia. Journal of Educational Studies. 2010;3:23-32.
5. Benedict HM. Causes of Economic down trend in Nigeria. Journal of Economics. 2015;5(3):36-42.
6. Ojimadu K. The Impact of unemployment on the economic growth of Nigeria. Journal of Economics and Statistics. 2011;4(3):56-68.
7. Echambadi P, Hess U. Marketing Science. 2007;26(3):438-445.
8. Fisher RA. The goodness of fit of regression formulae, and the distribution of regression coefficients. Journal of the Royal Statistical Society. Blackwell Publishing. 1925; 85 (4): 597–612.
9. Gauss HK. RegressionAnalysis —Theory, Methods and Applications, Springer-Verlag, Berlin, (4th printing). 1821.
10. Gauss HK. Introduction to Regression Analysis. Encyclopedia of Mathematics, Springer, 1809.
11. Legendre KS. Regression by example. Encyclopedia of Mathematics, Springer, 1805.