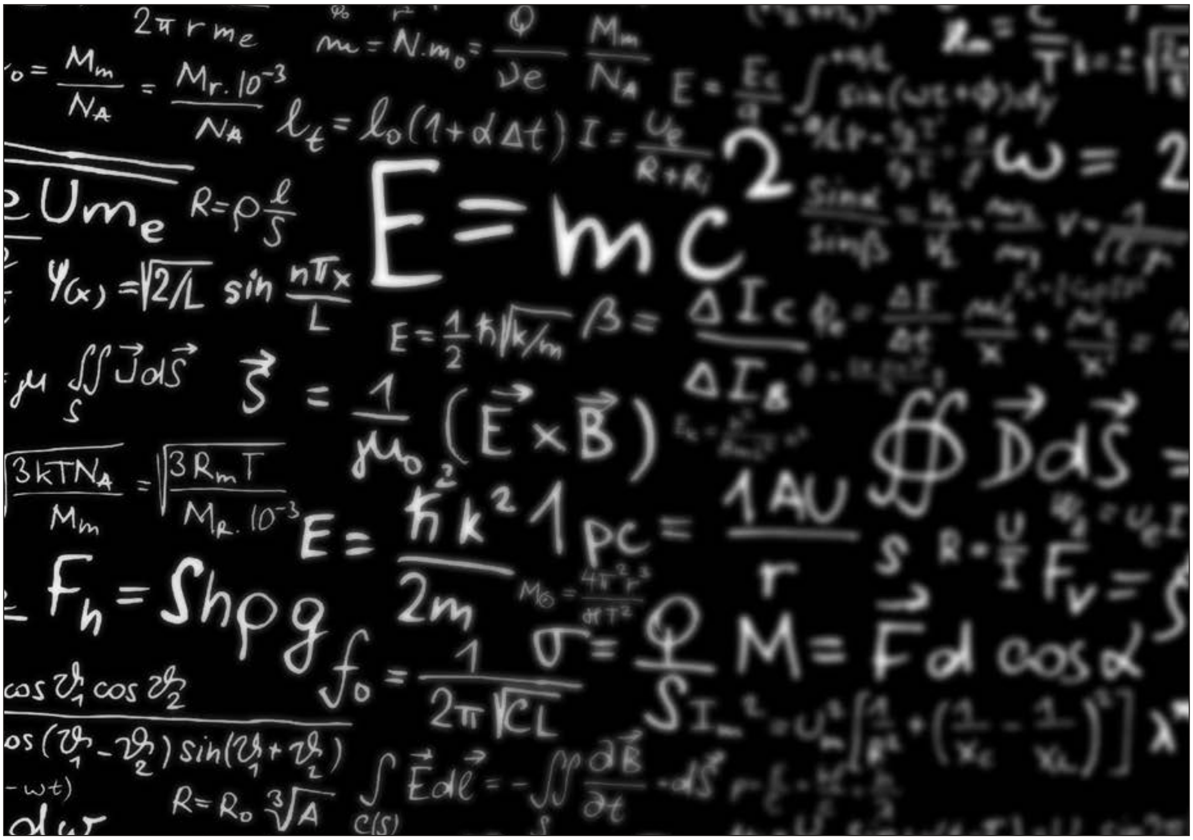

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

Asian Journal of Mathematics and Statistics 4 (4): 174-180, 2011
ISSN 1994-5418 / DOI: 10.3923/ajms.2011.174.180

Effect of Transformation on the Parameter Estimates of a Simple Linear Regression Model: A Case Study of Division of Variables by Constants

O.E. Okereke
Department of Maths, Statistics and Computer Science, Michael Okpara University of Agriculture, Umudike, Nigeria

ABSTRACT

The ability to forecast accurately the future values of a given variable within the minimum possible time gives organizations, governments and business enterprises the opportunity for appropriate decision and policy making. Accurate predictions can be made with a correctly specified model. It is evident that an estimated model comprises parameter estimates. Hence, different sets of estimates may give rise to different forecasts. Again, researchers and experimenters often report large values in some standard forms which include thousands, millions and billions so as to save time required for compilation and computation. In this study, an attempt was made to provide estimates of the parameters of the models involving the transformation obtained by dividing the variables by constants. The effect of the transformation on the parameter estimates was also emphasized. The relationships between the estimates of the parameters of the original model and those involving the transformed variables were derived. It has been shown that the division of the independent variable by a constant did affect the estimate of the slope only. On the other hand, the estimate of the slope of the original model remained unaffected when both variables were divided by the same constant where as other obtained parameter estimates appeared to differ from those of the original model. The theoretically derived estimates were substantiated by empirical data analysis. Moreover, the regression models fitted based on the various transformed variables differed from that of the original model. As a result, transformation by means of dividing the variables by constants affects the parameter estimates as well as the predictability of the model.

Key words: Simple linear regression, estimates, parameters, slope, empirical data, analysis, transformed variables

INTRODUCTION

Regression models are considered to be veritable tools for describing the functional form of the relationship between variables (Ding, 2006). They also play a key role in the implementation of multivariate tools like principal component analysis (Igwenagu, 2011) and factor analysis (Abdullah and Asngari, 2011). It is customary to estimate the model. With an estimated model, one can predict the value of the dependent variable corresponding to a given value of the independent variable (Sarkar and Midi, 2010).

Regression models are classified into two broad category namely linear and non-linear models (Rajarathinam and Parmar, 2011; El-Shhawy, 2008). Linear regression models are those ones that are linear in parameters. These include simple linear, multiple linear and polynomial regression models. A simple linear regression model is the one which involves one dependent variable and one independent variable. A simple linear regression is specified as:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \tag{1}$$

where, Y_i, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and e_1 represents ith value of the dependent variable, intercept and slope of the regression line and ith value of the error associated with the prediction of Y_i.

The least squares method of estimating the parameters of the model in Eq. 1 is usually preferred to other methods because it yields unbiased estimators (El-Salam, 2011; Ramirez *et al.*, 2002). Olaomi and Ifederu (2008) pointed out that the assumption of lack of autocorrelation between the error terms is required for parameter estimation and inference in ordinary least squares regression. The least square estimates b_1 and b_0 of $\boldsymbol{\beta}$ and α, respectively are given by:

$$b_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \tag{2}$$

and

$$b_0 = \overline{Y} - b\overline{X} \tag{3}$$

In practice, we often face the difficulty involved in fitting a regression model to data involving large values. Estimation of parameters of the regression model can be tedious and time consuming. Subtraction of constants from the variables is said to facilitates parameter estimation in regression analysis (Obioma, 2005). There are situations where subtraction of constants may not reduce the time required for the necessary computation. This include when the values are multiples of a given constant. In this case division outperforms subtraction. The effect of such transformation on the parameter estimates of the original model is the main focus of this study.

ESTIMATION OF PARAMETERS OF SIMPLE LINEAR REGRESSION MODELS INVOLVING SOME FUNCTIONS OF THE DEPENDENT AND INDEPENDENT VARIABLES

In this section, the estimates of the parameters of regression models obtained when either one or both of the variables are divided by constants are considered. Emphasis is also laid on the relationships between the estimates of the parameters of the original model and their counterparts obtained when the variables are divided by constants.

Estimation of the parameter of the regression model when the independent variable is divided by a constant: Let α be a constant such that $\chi = X/\alpha$. Suppose we wish to regress Y on χ. Then the associated regression model is of the form:

$$Y = \beta_{01} + \beta_{11}\chi + \varepsilon \tag{4}$$

The symbols Y, χ, $\boldsymbol{\beta}_{01}$, $\boldsymbol{\beta}_{11}$ and ε in Eq. 4 stand for the dependent variable, independent, variable, intercept of the line and slope of the line and the associated error term.

The estimates b_{01} and b_{11} of $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{11}$, respectively are obtained using Eq. 2 and 3 as follows:

$$b_{11} = \frac{\sum_{i=1}^{n}\left(\frac{X_i}{\alpha} - \frac{\overline{X}}{\alpha}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(\frac{X}{\alpha} - \frac{\overline{X}}{\alpha}\right)^2} = \frac{\frac{1}{\alpha}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{\alpha^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \alpha b_1 \tag{5}$$

and

$$b_{01} = \overline{Y} - b_{11}(\overline{\chi}) = \overline{Y} - \alpha b_1 \times \frac{\overline{X}}{\alpha} = \overline{Y} - b_1\overline{X} = b_0 \tag{6}$$

Estimates of the model parameters when the dependent variable is divided by a constant: Consider the regression model:

$$\gamma = \beta_{02} + \beta_{12}X + \varepsilon \tag{7}$$

where, the symbols γ, X, $\boldsymbol{\beta}_{02}$, $\boldsymbol{\beta}_{12}$ and $\boldsymbol{\varepsilon}$ in Eq. 4 stand for the dependent variable, independent, variable, intercept of the line and slope of the line and the associated error term.

Also, γ = Y/d and d is a constant.

Using Eq. 2:

$$b_{12} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\gamma_i - \overline{\gamma}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{\frac{1}{d}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{b_1}{d} \tag{8}$$

Using Eq. 3:

$$b_{02} = \overline{\gamma} - b_{12}\overline{X} = \frac{\overline{Y}}{d} - \frac{b_1}{d}\overline{X} = \frac{\overline{Y} - b_1\overline{X}}{d} = \frac{b_0}{d} \tag{9}$$

Estimates of the parameters of the model when the dependent variable and independent variable are divided by different constants: Let: $U = \frac{X}{e}$ and $V = \frac{Y}{f}$ where, e and f are non-zero constants.

Consider the regression model:

$$V = \beta_{03} + \beta_{13}U + \varepsilon \tag{10}$$

If b_{03} and b_{13} denote the required estimates of the parameters $\boldsymbol{\beta}_{03}$ and $\boldsymbol{\beta}_{13}$, respectively. Using Eq. 2, we obtain:

$$b_{12} = \frac{\sum_{i=1}^{n}\left(U_i - \overline{U}\right)\left(V_i - \overline{V}\right)}{\sum_{i=1}^{n}\left(U_i - \overline{U}\right)^2} = \frac{\frac{1}{ef}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{e^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{b_1 e}{f} \tag{11}$$

Using Eq. 3

$$b_{03} = \overline{V} - b_{13}\overline{U} = \frac{\overline{Y}}{f} - \frac{b_1 e}{ef}\overline{X} = \frac{\overline{Y} - b_1\overline{X}}{f} = \frac{b_0}{f} \tag{12}$$

Estimates of the parameters of the model when the dependent and independent variables are divided by the same constant: Consider the regression model:

$$\nu = \beta_{04} + \beta_{14}\mu + \varepsilon \tag{13}$$

Where:

$\nu = \frac{Y}{g}$, $\mu = \frac{X}{g}$ and g = a non-zero constant.

The least squares estimate of the shape β_{14} is obtained with help of Eq. 2 as follows:

$$b_{14} = \frac{\sum_{i=1}^{n}\left(\mu_i - \overline{\mu}\right)\left(\nu_i - \overline{\nu}\right)}{\sum_{i=1}^{n}\left(\mu_i - \overline{\mu}\right)^2} = \frac{\frac{1}{g^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{g^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = b_1 \tag{14}$$

The corresponding estimate of the intercept β_{04} is obtained from Eq. 3 as:

$$b_{04} = \overline{\nu} - b_{14}\overline{\mu} = \frac{\overline{Y}}{g} - \frac{b_1}{g}\overline{X} = \frac{\overline{Y} - b_1\overline{X}}{g} = \frac{b_0}{g} \tag{15}$$

NUMERICAL ILLUSTRATION

The relationships between the parameter estimates are verified in this section using the data on the export (in billions of dollars) (X) and import (in billions of dollars) (Y) provided by the Bureau of Economic Analysis of the U.S Department of Commerce for the date range 2006-01-01 to 2010-10-01.

The nature of the relationship between X and Y is determined with help of the scatter diagram in Fig. 1.

It can be deduced from Fig. 1 that there is a linear relationship between imports and exports of goods and services in U.S within the given period.

Also, the analysis of variance in Table 1 shows that the simple linear regression of import on export is significant at $\alpha = 0.05$.

Summary of regression analysis of the various functions of the dependent variable on their corresponding independent variables is given in Table 2.

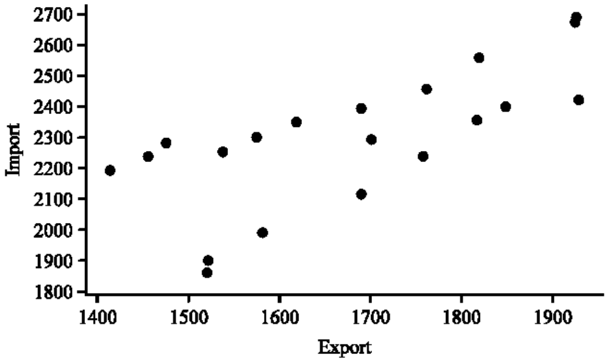

Fig. 1: The scatter diagram for determining the nature of the relationship between X and Y

Table 1: ANOVA table for regression of import on export

Source	df	SS	MS	F	p-value
Regression	1	485813	485813	19.33	0.00
Error	18	452370	25132		
Total	19	938184			

Table 2: Regression analysis of the transformed dependent variables on their associated independent variables

Variables used in regression		
Dependent	Independent	Regression equation
X	Y	Y = 677.781+0.965074X
$\chi = \frac{X}{10}$	Y	Y = 677.781+9.65074χ
X	$\gamma = \frac{Y}{10}$	γ = 6.77781+0.0965074X
$U = \frac{X}{10}$	$V = \frac{Y}{100}$	V = 6.77781+0.0965074U
$\mu = \frac{X}{1000}$	$v = \frac{Y}{1000}$	v = 0.677781+0.965074μ

For simplicity and clarity the constants 10, 100 and 1000 are used in this numerical illustration. As we can see from Table 2, the estimate of the slope is not affected only when the variables involved in regression are divided by the same constant. On the other hand, the estimate of the intercept of the original variable remains unaffected when only the independent variable is divided by a constant. All the estimates in Table 2 agree with results obtained in section 2.

DISCUSSION

A statistical technique for technique for reducing the time required for estimation of parameters in simple linear regression has been examined in this study. Estimates of parameters of four regression models resulting from the division of the variables by constants were compared with those of the original model. On dividing the independent variable by a constant, it was observed that the estimate of the slope of the resulting model was equal to that of the original model multiplied by the given constant. The estimated intercept remained unaffected by the transformation. Dividing the dependent variable by a constant and regressing the quotient on the independent variable yielded parameter estimates equal to their corresponding estimates in the original model divided by the constant.

The regression model involving the variables divided by different constants was also fitted. This gave rise to the estimate of the slope which could be obtained by multiplying that of the original model by the constant by which the independent has been divided and dividing the product by the constant by which the dependent was divided. The associated estimate of the intercept would be obtained by dividing that of the original model by the constant by which the dependent has been divided.

Furthermore, the division of the variables by the same constant appeared not to affect the estimate of the slope of the original variable. This agrees with the results in the literature with regard to addition and subtraction of constants from the variables (Steel and Torrie, 1981; Spiegel *et al.*, 2000). The result obtained based on this transformation suggested that the estimate of the intercept be found by dividing that of the original model by the chosen constant. Since estimated regression model is a function of the concerned parameter estimates, prediction in simple linear regression can be affected by division of variables by constants.

CONCLUSION

Based on the results obtained in this research, the division of either or both of the variables by different constants in simple linear regression could affect the estimates of the parameters of the original model. Again, division of the variables by the same constant could also affect the estimate of the intercept of the original model while the estimate of the intercept would be unaffected by division of the independent variable by a constant. Hence, it is obvious that division of any or both of the variables in simple linear regression affects the predictability of the fitted model. It is now recommended that the derived relationships be considered by analysts using the proposed transformations in regression analysis so as to ensure proper prediction.

REFERENCES

Abdullah, L. and H. Asngari, 2011. Factor analysis evidence in describing consumer preferences for a soft drink product in Malaysia. J. Applied Sci., 11: 139-144.

Ding, C.S., 2006. Using regression mixture analysis in educational research, practical assessment. Res. Eval., 11: 1-11.

El-Salam, M.E.F.A., 2011. An efficient estimation procedure for determining ridge regression parameter. Asian J. Math. Stat., 4: 90-97.

El-Shhawy, S.A., 2008. Selection of a NRL-model by re-sampling technique. Asian J. Math. Stat., 1: 109-117.

Igwenagu, C.M., 2011. Principal component analysis of global warming with respect to CO_2 emission in Nigeria: An exploratory study. Asian J. Math. Stat., 4: 71-80.

Obioma, N.V., 2005. Principles of Statistical Inferences. Peace Publishers Ltd., Owerri, Nigeria, pp: 179.

Olaomi, J.O. and A. Ifederu, 2008. Understanding estimators of linear regression model with AR(1) error which are correlated with exponential regressor. Asian J. Math. Stat., 1: 14-23.

Rajarathinam, A. and R.S. Parmar, 2011. Application of parametric and nonparametric regression models for area, production and productivity trends of castor (*Ricinus communis* L.) crop. Asian J. Applied Sci., 4: 42-52.

Ramirez, O.A., S.K. Misra and J. Nelson, 2002. Estimation of efficient regression models for applied agricultural economics research. Proceedings of the Annual Meeting of Agricultural Economics Association, July 28-31, Long Beach, CA., USA., pp: 1-32.

Sarkar, S.K. and H. Midi, 2010. Importance of assessing the model adequacy of binary logistic regression. J. Applied Sci., 10: 479-486.

Spiegel, M.R., J.J. Schiller and R.A. Srinivasan, 2000. Schaum's Outline of Theory and Problems of Probability and Statistics. 2nd Edn., Tata McGraw-Hill Publishing Company Limited, New Delhi, India, ISBN-13: 9780071350044, pp: 408.

Steel, R.G.D. and H.J. Torrie, 1981. Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill International Book Company, London.