On the Misconception of R^2 for $(r)^2$ in a Regression Model

Ijomah, Maxwell Azubuike

Dept. of Maths/Statistics, University of Port Harcourt, Nigeria

***Abstract*:-The coefficient of determination (R^2) is perhaps the single most extensively used measure of goodness of fit for regression models, and measures the proportion of variation in the dependent variable explained by the predictors included in the model. It is however, widely misused as the square of correlation coefficient and this has led to poor interpretation of research reports in regression model. In this paper, we investigate the controversy regarding use of coefficient of determination as the square of correlation coefficient in statistical analysis. Difference between the two statistics are illustrated using examples from simple and multiple regression models.**

***Keywords:* linear regression; coefficient of determination; correlation coefficient; multiple correlation; regression coefficient.**

I. INTRODUCTION

The classical linear regression model is the standard procedure for extracting the statistical information from the data through the determination of relationship between the study and explanatory variables. In the course of model estimation, it is common practice to assess the appropriateness or adequacy of the fitted model in explaining the variations in the data set. A popular tool to determine the adequacy of the fitted model is the coefficient of determination. The coefficient of determination is a measure used in statistical analysis that assesses how well a model explains and predicts future outcomes. It is indicative of the level of explained variability in the data set and considers the variation in the dependent variable explained by the independent variable(s). It provides a summary measure for the goodness of fit of any linear regression model and is based on the proportion of variability of the study variable that can be explained through the knowledge of a given set of explanatory variables. These definitions are found by both econometrics and statistics handbooks and is widely accepted among quantitative scholars. The coefficient of determination R^2 as stated above is generally used by researchers to assess the goodness of fit of the models. In econometrics, Kennedy (2008) argues that "R^2 is measured either as the ratio of the "explained" variation to the "total" variation […] and represents the percentage of variation in the dependent variable "explained" by variation in the independent variables" (Kennedy, 2008). Wooldridge (2009) states that "R^2 is the ratio of the explained variation compared to the total variation; thus, it is interpreted as the fraction of the sample variation in y that is explained by x" (Wooldridge, 2009). According to Moore and McCabe (2009), the square of the correlation, R^2, is the fraction of the variation in one variable that is explained by least-squares regression on the other variable. Kasuya (2018) also pointed out that use of *r* or *R*, not r^2 or R^2, for the coefficient of determination is confusing and inappropriate. The coefficient of determination is the square of the correlation coefficient or multiple correlation coefficients, which are usually denoted by *r* or *R*. The use of *r* or *R* for the coefficient of determination is likely to make the misleading impression that it is on the same scale as the correlation or multiple correlation coefficients despite being the square of them. Similarly, Schroeder et al. (1986) argue that "R^2, the coefficient of multiple determination, measures the percentage of the variation in the dependent variable which is explained by variations in the independent variables taken together" (Schroeder et al., 1996).Anderson-Sprecher (1994) pointed out that "the coefficient of multiple determination, R^2, is a measure many statisticians love to hate. This animosity exists primarily because the widespread use of R^2 inevitably leads to at least occasional misuse" (Anderson-Sprecher, 1994 p. 113). While the controversy over R^2 has its origin in the statistics literature (Kavalseth, 1985; Helland, 1987; Willett and Singer, 1988; Lavergne, 1996; Korn and Simon, 1991; Scott and Wild, 1991; McGuirk and Driscoll, 1995), the R^2 debate is important to all fields of knowledge that employ linear regression models. McGregor (1993) argues that "there is little wonder that the regression model has achieved its preferred status in the social sciences" (McGregor, 1993). In fact, the attractiveness of the regression model can be partially explained by its capacity to summarize the relationship among different variables in a systematic and parsimonious approach. Therefore, since the use of regression models have been increasing in social sciences in general and political science in particular, it is important to understand the controversial role of R^2 and the substantive meaning scholars can draw from it. Some authors largely reject the usage of the coefficient of determination, e.g. Achen (1982): Gary King (1986) argues that R^2 is highly misused as a measure of the influence of X on Y. He states that the more accurate interpretation is that "R^2 is a measure of the spread of points around a regression line, and it is a poor measure of even that" (King, 1986). The most misuse of coefficient of determination which occupies over 80 percent of literature in regression is the misconception of the coefficient of determination (R^2) for the square of correlation coefficient $(r)^2$ in regression analysis by most researchers. Similarly, Achen (1977) states that one of the main limitations of the correlation coefficient is its inability to be compared among samples. He argues that “correlations cannot be compared across samples: two correlations can differ because the variances in the samples differ, not because the underlying relationship has changed” (Achen, 1977 p. 807). To strengthen the case for model fit, consideration must be given to the intercept of the regression. Linear regression finds the best line that predicts dependent variable from independent variable(s). The decision of which variable calls dependent and which calls independent is an important matter in regression, as it'll get a different best-fit line if you swap the two. The line that best predicts independent variable from dependent variable is not the same as the line that predicts dependent variable from independent variable in spite of both those lines have the same value for R^2. A large share of quantitative literature devotes little attention to addressing this misconception, giving to much attention to the “proportion of dependent variable explained by the model”. The purpose is to provide an intuitive understanding regarding the coefficient of determination and point out some not so obvious mistakes that are frequently made when interpreting the coefficient of determination R^2 as the square of correlation coefficient in a linear model. The plan of the paper is as follows: The statistical model for coefficient of determination and its relationship with correlation coefficient are described in Section 2. In Section 3, we demonstrate the inconsistency of the correlation coefficient as the root of coefficient of determination under various scenarios and with simulated data and data from Anscombe`s (1973).The findings of the simulation study are presented in Section4 followed by some concluding remarks in Section 5

II. STATISTICAL MODEL

Given paired variables (x_i, y_i), a linear model that explains the relationship between the variables is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1)$$

where e is a mean zero error. The parameters of the linear model can be estimated using the least squares method and denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The parameters are estimated by minimizing the sum of squared residuals between variable y_i and the model

$$\beta_0 + \beta_1 x, \text{that is, } (\beta_0, \beta_1) = \arg\min(y_i - \beta_0 + \beta_1 x_i)^2 \qquad (2)$$

It can be shown that the least square estimates are

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \bar{x}\frac{S_{XY}}{S_{XX}} \text{ and}$$

$$\beta_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where the sample cross-covariance S_{xy} is defined as

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\bar{y}$$

From the above, the sum of squared errors (SSE), or the sum of squared residuals, is given by

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The total sum of squares (SST) is the measure of total variation in the Y variable and is defined as

$$SST = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Since SSE is the minimum of the sum of squared residuals of any linear model, SSE is always smaller than SST. Then the amount of variability explained by the model is SST − SSE, which is denoted as the regression sum of squares (SSR), that is,

$$SSR = SST - SSE$$

The ratio SSR/SST = (SST − SSE)/SST measures the proportion of variability explained by the model with variance estimators given as $\hat{V}(y) = SST/n$, $\hat{V}(\hat{y}) = SSR/n, \hat{V}(\varepsilon) = SSE/n$

The coefficient of determination (R^2) is defined as the ratio

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$= \frac{\hat{V}(\hat{y})}{\hat{V}(y)} = 1 - \frac{\hat{V}(\varepsilon)}{\hat{V}(y)} \qquad (3)$$

R^2therefore measure the explanatory power of the model which in turn reflects the goodness of fit of the model. It reflects the model adequacy in the sense that how much is the explanatory power of explanatory variable.

The interpretation and use of R^2 has been extensively discussed in the literature (some examples are in Crocker, 1972; Barrett, 1974; Draper and Smith, 1981; Ranney and Thigpen, 1981; Healy, 1984; Kvalseth, 1985; Helland, 1987; Willett and Singer, 1988; Nagelkerke, 1991; Scott and Wild, 1991). with increasing n, R^2 tends to

$$R^2 = \frac{\beta' S_x \beta}{\beta' S_x \beta \sigma^2_{y,1.p}} \qquad (4)$$

where S. is the sample covariance matrix for the independent variables (Helland, 1987). Thus, the value of R' depends on the variation among independent variables.

Relation to Correlation Coefficient

With the previous Equations 1 and 2, R^2 can also be written as a function of the sample cross covariance

$$R_{Y,X}{}^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \frac{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

This implies that

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \quad (5)$$

This shows that the coefficient of determination of a simple linear regression is the square of the sample correlation coefficient of $\left(x_i, y_i\right)$. Equation 4 perhaps formed the basis for the above misconception. Note however that this definition does not refer to one variable as dependent and the other as independent. Rather, it simply refers to two random variables.

Again, consider the case of one of the regressors is constant, the empirical correlation between y and $\hat{y}$ is non negative and equals

$$\hat{\rho}(y,\hat{y}) = \frac{\hat{C}(y,\hat{y})}{\left[\hat{V}(y)\hat{V}(\hat{y})\right]^{\frac{1}{2}}} \quad (6)$$

where

$$\hat{C}(y,\hat{y}) = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \bar{y}\right)\left(\hat{y}_t - \bar{y}\right) = \frac{1}{n}\left(Ay\right)'\left(A\hat{y}\right)$$

and $A = \mathrm{I}_n = \frac{1}{n} ii'$

since one of the regressors is a constant,

$$A\hat{\varepsilon} = \hat{\varepsilon},\ Ay = A\hat{y} + \hat{\varepsilon},\ \hat{\varepsilon}(A\hat{y}) = \hat{\varepsilon}'\,\hat{y} = 0$$

and $\hat{C}(y,\hat{y}) = \frac{1}{n}\sum_{t=1}^{n}\left(A\hat{y} + \hat{\varepsilon}\right)'\left(A\hat{y}\right)$

$$= \frac{1}{n}\sum_{t=1}^{n}\left(A\hat{y}\right)'\left(A\hat{y}\right) = \hat{V}(\hat{y})$$

therefore

$$\hat{\rho}(y,\hat{y}) = \frac{\hat{V}(\hat{y})}{\left[\hat{V}(y)\hat{V}(\hat{y})\right]^{\frac{1}{2}}} = \left[\frac{V'(y}{V(y)}\right]^{\frac{1}{2}} = \sqrt{R^2} \geq 0 \quad (7)$$

The above equation clearly shows that the root of the coefficient of determination is always positive. This again disagrees with correlation coefficient which ranges from -1 to 1. Squaring this renders all negative values of the correlation coefficient meaningless.

III. NUMERICAL APPLICATION

Both simulated data and data extracted from Anscombe`s (1973) were used for the analysis. Two models were considered for the illustration, the first was a simple regression model (i.e. Y was a function of X and vice versa). Then, we ran another model in this case a multiple regression by including the new independent variable (X_2) with each of the variable as a dependent variable. The result summary is as show below:

Table 1: Simple and Multiple Regression with varying standard errors

Statistics	**Model 1**		**Model 2**		
	$y = f(x)$	$x = f(y)$	$y = f(x_1, x_2)$	$x_1 = f(y, x_2)$	$x_2 = f(y, x_1)$
	$\beta_0 = 9.8405$ $\beta_1 = -0.8355$	$\beta_0 = 8.2925$ $\beta_1 = -0.6959$	$\beta_0 = 1.2919$ $\beta_1 = 0.8586$ $\beta_2 = 1.0930$	$\beta_0 = -0.3686$ $\beta_1 = 0.1722$ $\beta_2 = 0.3984$	$\beta_0 = -0.3511$ $\beta_1 = 0.4097$ $\beta_2 = 0.1391$
r	-0.7625	-0.7625	-	-	-
$(r)^2$	0.5814	0.5814	-	-	-
R	0.7625	0.7625	0.9661	0.9392	0.9492
R^2	0.5814	0.5814	0.9333	0.8821	0.9011
R_{adj}^2	0.5727	0.5727	0.9305	0.8771	0.8969
$RMSE$	2.2598	2.0623	0.9116	0.6209	0.5581
CV	17.4479	21.4262	13.1032	21.3920	24.2261

For model 1, we considered when y is a function of x and vice versa since in both cases we should have the same correlation coefficient as shown in the above table 1. It was observed that the correlation between the two variables is negative (i.e. x is inversely related with y when y is depending on x and when x depends on y).Now the question is how do we reconcile these two interpretations of negative relationship and a coefficient of determination of 0.5814? Again, we noticed that there is clear difference in both root mean squared error and coefficient of variation (i.e. when y depends on x, RMSE of 2.2598 with CV 17.4479 was obtained as against RMSE of 2.0623 and CV 21.4262 for x depending on y) even though the square of correlation coefficient (r)^{2}is equal to coefficient of determination (R^2).In order to prove our point, since the correlation between x and y cannot change irrespective of which variable comes first, it is expected that all the available statistics should be the same. But the result shows that the standard deviation of the residual in case 2 of model 1 is higher than that in case 1 as can be seen in the above table. To make our case clearer, in model 2 we considered a multiple regression were each of the variables are dependent on others. Though in each case of model 2, the coefficient of determination was equal to the square of multiple correlation (R), but there was variation in the spread of points about the fitted regression lines. (i.e. there is variation in noise of the system). Recall by definition, the coefficient of determination, measures the percentage of the variation in the dependent variable which is explained by variations in the independent variables taken together" (Schroeder et al., 1996). This goes to prove that for three different scenarios, the coefficient of determination varied with each dependent variable, also their standard error (RMSE) and coefficient of variation were not the same. The principal problem here is that the variance in the population of the explanatory variables studied can strongly influence R^2 magnitude unlike that of correlation coefficient which is influence by not only explanatory variable(s) but also dependent of other variable. Therefore, there is no guarantee that a small R^2 indicates a weak relationship, given that the statistic is largely influenced by variation in the independent variable Filho et. al (2011).It is worth noting that the coefficient of multiple correlation R which is the square root of coefficient of determination, is only positively skewed and as such could not account for the inverse correlation coefficient.

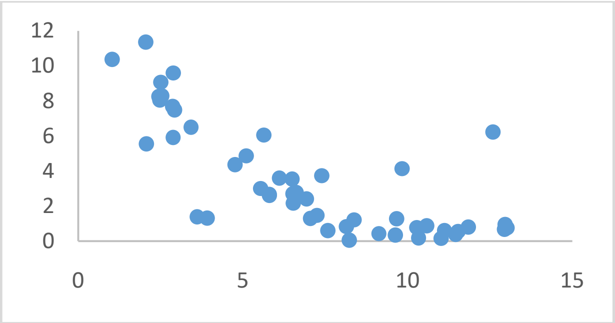

Fig. 1a: Regressing Y vs X

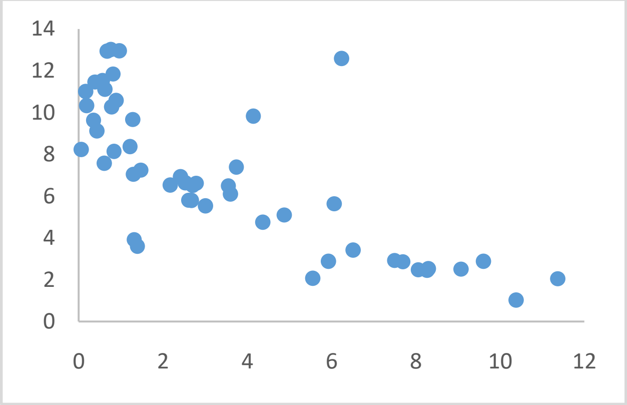

Fig.1b: Regressing X vs Y

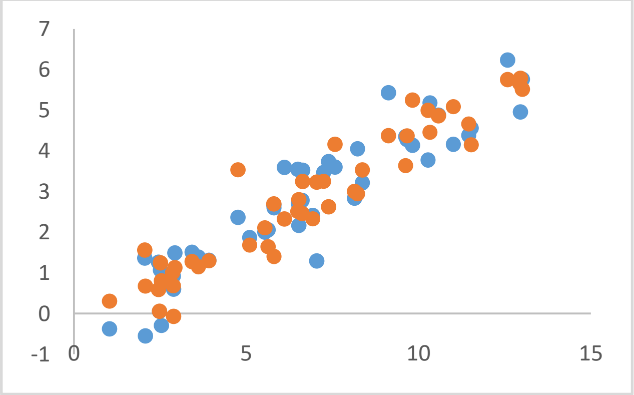

Fig. 2a: Regressing Y vs (X_1,X_2)

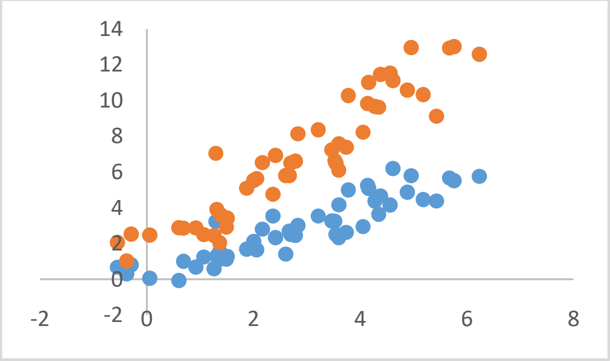

Fig.2b: Regressing X_1 vs (Y,X_2)

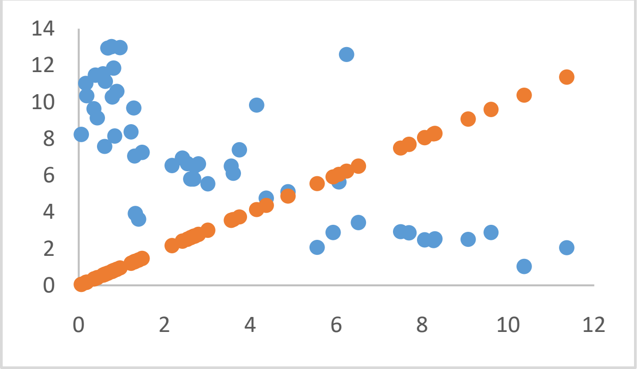

Fig.2c: Regressing X_2 vs (Y,X_1)

The above graph clearly shows that fig.1a and fig.1b are not the same even though the have the same R^2 value. This goes to prove that even though the coefficient of determination and the square of multiple correlation are mathematically the same, their interpretation is not the same from the graph. Similarly, Fig.2a in the above graph appears more linear than fig. 2b. The point we are trying to make here is that for coefficient of determination, it is the explanatory variable(s) explaining variation in the dependent variable while for correlation, we are interested in the strength of relationship. Therefore, interpreting R^2as the square of r can be misleading.

Empirical example 2: replicating Anscombe`s (1973) data

Consider again the result extracted from Filho et.al (2011) below, in model 2 surprisingly the square of multiple correlation (R) is not equal to coefficient of determination. Model 2 shows both a higher correlation coefficient (0.889) and a higher coefficient of determination (0.809) when compared with model 1 (0.816 and 0.667, respectively). The standard error of the estimate of the model 2 is smaller (0.993) than the error of model 1 (1.237). Finally, model 2 reached a significance level (0.001) more reliable than model 1 (0.002).

Table 2: Result Extracted from Figueredo et.al (2011)

Statistics	Model 1	Model 2
R	0.816	0.889
R^2	0.667	0.809
R^2_{adj}	0.629	0.761
$Std \quad Error$	1.237	0.993
F	17.990	16.916
$sig.$	0.002	0.001
df	10	8

***Note**: Result was extracted from: FIGUEIREDO FILHO Dalson. B.; SILVA, José. A. and; ROCHA, Enivaldo (2011). What is R² all about?*

IV. DISCUSSION

It is often said that multiple correlation can be used to identify good predictors. This is not the case. Multiple correlation does not identify predictors of a criterion as shown in the table1 above. It identifies variables that add to prediction. The correlation coefficient is the regression coefficient in standard score form while the regression coefficient in multiple regression is a measure of the extent to which a variable adds to the prediction of a criterion, given the other variables in the equation. It is not a correlation coefficient. If strength and direction of a linear relationship should be presented, then r is the correct statistic. If the proportion of explained variance should be presented, then R^2 is the correct statistic. These are just different things. The coefficient of determination just like the regression line provides no information about how strongly the variables are related. In contrast, a correlation does not fit such a line and does not allow such estimations, but it describes the strength of the relationship. A low R^2 does not necessary mean that there is no relationship but that the explanatory variable alone cannot capture enough variation in the dependent variable. If R^2 should tell something about the virtues of a model for some given population, the variation among the values of the independent variables should be representative of that population. Another misconception that should be taken note of is that multiple correlation (R) considers only positive correlation coefficient (r). R^2 also does not deal with signs (- or +) in expressing the explanatory power of the independent variable but the degree. It should therefore be noted that the value of R^2 does not depend only on the distances between predicted and observed values but also on the variation of the outcome variable. So anything that influences this variation also influences the value of R^2.In fact, the attractiveness of the regression model can be partially explained by its capacity to summarize the relationship among different variables in a systematic and parsimonious approach. Little wonder while some researchers have criticized use of $(r)^2$ as a measure of goodness of fit. Again, we often denote coefficient of determination by R^2 while correlation coefficient is denoted by r. Statistically speaking r is always classified as a subset of R and so R^2 cannot be equal to $(r)^2$.

V. CONCLUSION

Due tothe fact that coefficient of determination (R^2) depends on the correlation coefficient (r), researchers often assume coefficient of determination as the square of correlation coefficient. Such an issue has continued to reoccur in the literature, to the best of our knowledge. Both techniques have a close mathematical relationship, but distinct purposes and assumptions. With the above discussion, it is evident, that there is a big difference between these two mathematical concepts, although these two are studied together. Correlation is used when the researcher wants to know that whether the variables under study are correlated or not, if yes then what is the strength of their association. Correlation is a very useful research statistic but do not address the predictive power of variables. This task is left to coefficient of determination. Coefficient of determination is based on the idea that the researcher must first have some valid reasons for believing that there is a causal relationship between two or more variables.

REFERENCES

[1]. Achen, C. H. (1977). Measuring Representation: Perils of the Correlation Coefficient. American Journal of Political Science 21: 805-815.

[2]. Anderson-Sprecher, R. (1994). Model Comparisons and R^2. The American Statistician 48: 113-117.

[3]. Anscombe, F. (1973). Graphs in Statistical Analysis. American Statistician 27: 17-21.

[4]. Filho DBF, Silva JA, Rocha E.(2011). What is R2 all about? Leviathan – Cadernos de Pesquisa Política;3:60–68.

[5]. Helland, I. S. (1987). On the Interpretation and Use of R^2 in Regression. Biometrics 43(1): 61-69.

[6]. Kasuya E. (2019). On the use of r and r squared in correlation and regression. Ecol. Res. 34:235–236.

[7]. Kennedy, P. (2008). A Guide to Econometrics. San Francisco, CA: Wiley-Blackwell.

[8]. King, G. (1986). How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. American Journal of Political Science 30:666-687.

[9]. Korn, E. L., and Simon. R. (1991). Explained Residual Variation, Explained Risk, and Goodness of Fit. The American Statistician 45(3): 201-206.

[10]. Kvalseth,T. 0. (1985). Cautionary Note About R^2. The American Statistician 39: 279-285. Lavergne, P. (1996). The Hot Air in R^2: Comment. American Journal of Agricultural Economics 78(3): 712-714.

[11]. McGregor, J. P. (1993). Procrustus and the regression model: On the misuse of the regression model. PS: Political Science & Politics 26: 801-804.

[12]. McGuirk, A. and Driscoll P. (1995). The Hot Air in R^2 and Consistent Measures of Explained Variation. American Journal of Agricultural Economics 77: 319-328.

[13]. Moore, D. S., and McCabe, G.P(2009). Introduction to the Practice of Statistics. West Lafayette, IN: W.H. Freeman Press.

[14]. Schroeder, L. D., David L. S., and Paula E. S. (1986). Understanding Regression Analysis: An Introductory Guide. Beverly Hills, CA: Sage Publications.

[15]. Scott, A. and Wild. C.(1991). Transformations and R^2. The American Statistician 45(2): 127-129.

[16]. Willett, J. B., and SingerJ.D.(1988). Another Cautionary Note about R2: Its Use in Weighted Least-Squares Regression Analysis. The American Statistician 42(3): 236-238.

[17]. Wooldridge, J. M. (2009). Introductory Econometrics: A Modern Approach. Boston, MA: South-Western College Publishing.