International Journal of Scientific Research in Mathematical and Statistical Sciences **Research Paper**
Volume-4, Issue-6, pp.12-16, December(2017) **E-ISSN:** 2348-4519

Use of R^2 and its Shortcomings

R. K. Borah [1*], K. K. Singh Meitei[2], S. C. Kakaty[3]

[1*]Dept. of Statistics, Manipur University, Imphal-795003, India
[2] Dept. of Statistics, Manipur University, Imphal-795003, India
[3]Dept. of Statistics, Dibrugarh University, Dibrugarh-786004, India

[*]*Corresponding Author: mr_raju06@rediffmail.com, Tel.: 91-9435316601*

Available online at: www.isroset.org

Received: 21/Oct/2017, Revised: 05/Nov/2017, Accepted: 21/Nov/2017, Published: 31/Dec/2017

***Abstract* —** The coefficient of determination R^2 is a general measure of usefulness of the regression model. It shows the percentage of the total variation in the response variable which can be explained by the explanatory variable and is considered as the most commonly used measure of goodness of fit for regression models. It is demonstrated by many statisticians and practitioners that expression for the coefficient of determination is generally not equivalent. However it is widely misused. The primary source of the problem is that except for linear models with an intercept term, the several R^2 statistics are not equivalent.

Keywords— Coefficient of determination, Regression model, Regression analysis, Resistant or robust models

I. INTRODUCTION

For fitting models to some data, a data analyst is likely to use the coefficient of determination R^2 to assess the goodness of fit of the models. If the model is anything but linear with an intercept term, it is not unlikely that, the use of R^2statistic will be inappropriate and ends up with possibly misleading results. We have encountered some of these and identify them in this paper. The underlying problem appears to be essentially two folds. First the data analyst is confused with possible variety of R^2 statistics. For the case of linear least squares regression models with an intercept term, the majority of the R^2 statistics are equivalent. Secondly the other types of models, such as linear no-intercept models or nonlinear (in the parameters) models, the various R^2statistics represent different values. This paper aims at discussing various considerations and potential pitfalls in using the R^2*s.* The work is carried out in the line of Hahn [2,3], Hawkins [4], Kevalseth [6], Willet and Singer [11] and Scott and Wild [9]. A non linear model is one in which at least one of the parameters appears nonlinearly. Sometimes the relationship between response and explanatory variables is not linear. In some cases a nonlinear function can be lineralized by using a suitable transformation. Such non linear models are called intrinsically or transformably linear. In practice, we are not so fortunate about our data. Some theory may suggest that the relationship between certain pairs of variables can adequately represented only by a non linear function. In such situation we are faced with two possibilities (i) apply appropriate non linear relationship directly to data or (ii) transform the nonlinear relation so that standard technique of the linear model may be applied to the transformed data. Logarithmic and reciprocal are the most commonly used transformation.

An attempt is made here to answer the question, why values of R^2 mislead in various regression models? The answer of this question is exemplified here by two numerical examples along with following basic related concepts. These are:

a)Identify the problem generally related to the use of R^2,

b) Compare the various statistics for different types of R^2 models,

c) Find out some possible shortcomings of use of R^2.

Few cautionary comments have been made in the literature (e.g., Hahn [2,3], Marquardt and Snee [8], Montgomery and Peck [7]), but these appear to be confined to linear-no-intercept models or their special case of so-called " mixture model "and cover only a part of the problem. But, Kvalseth [6] addressed the problem in general and made comparison of various R^2 for different types of models. He proposed for an appropriate and generally applicable R^2 statistics. The work in this paper is done in the line of Kvalseth [6].

II. REGRESSION MODELS

In general, scholars use one particular method of obtaining a mathematical relationship between explanatory variable(s) and response variable. In much experimented work we use to

investigate how the changes in explanatory variables affect response variable. The mathematical relationship model between explanatory variable(s) and response variable may be of any degree of polynomial or exponential, or others in the explanatory variable(s).The models considered in this study are.

(a)Simple linear regression model with intercept term i.e

$$E(y) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

where y is the response variable, x_j is the j^{th} explanatory variable and β_0 is the intercept of the regression plane. This model describes a hyper plane in the *k* dimensional space of the explanatory variables x_j,

(b) Simple linear no-intercept regression model, i.e.

$E(y)= \sum_{j=1}^{k} \beta_j x_j$,where the data lie in a region of *x* space remote from origin, which is a particular case of simple linear regression model with intercept term, assuming the intercept term β_0 is equal to 0,

(c) Power model i.e.

$E(y)= \beta_0 x^{\beta_1}$,where β_0 and β_1 are parameters,

(d) Exponential model i.e.

$$E(y) = \beta_0 e^{\beta_1 x}$$

,where β_0 and β_1 are parameters and

(e) Reciprocal model i.e.

$$E(y) = \beta_0 + \beta_1 \frac{1}{x}$$

Where β_0 and β_1 are parameters. These models are first lineralized by using appropriate standard technique to the transformed data. Logarithmic and reciprocal are the most commonly used transformation and then fitted to data by using ordinary least squares method.

III. SOME COMMON MISTAKES GENERALLY RELATED TO THE USE OF R^2

$y_i, \hat{y}_i y_i, \hat{y}_i \log_e y_i$, $\log_e \hat{y}_i \hat{y} y_i, \hat{y}_i y_i, \hat{y}_i$ One of the most frequent mistakes occurs when the fits of a linear and nonlinear model are compared by using the same R^2 expression but different variables and *y* and fitted for the linear model and transformed variables for the non linear models. Thus, for example, a power model $E(y) = \beta_0 x^{\beta}_1$ or or an exponential model $E(y) = \beta_0 e^{\beta}_1{}^{x}$ may first be linearized by using logarithmic transformation and then fitted to data by using ordinary least squares method. The R^2 value is often calculated using the data points ( ) and interpreted as a measure of the goodness of fit of the non linear model and compared with the fit of linear model that R^2 is determined by the same R^2 expression but using the data points ( ). The R^2 value based on the transformed data points, however, provides a measure of fit for the linearized model and not for the nonlinear model. To make sensible comparison between the fits of a linear and nonlinear model to the same set of data, comparable data points ( ) and R^2 expressions have to be used otherwise misleading results may be obtained.

IV. SALIENT POINTS OF R^2 STATISTIC

$y_i - \hat{y}_i$To make a correct choice of R^2 statistic, it is essential to impose some basic priory requirements relating to the properties of a good statistic. Such as follows

(i) R^2 must possess utility as a measure of goodness of fit and have perceived by the mind reasonable interpretation.
(ii) R^2 should be independent of the units of measurement of the model variables.
(iii) The coefficient of determination ranges from 0 to 1, i.e.0 $\leq R^2 \leq 1$, should be well defined with end points. An R^2 of zero means that the predictor accounts for none of the variability of the response variable and there is no regression prediction of *y* by *x*. An R^2 of 1 means perfect prediction of *y* by *x* and 100% of the variability of *y* is accounted for by *x* and R^2 of greater than 0 is for any reasonable model specification.
(iv) R^2should not be confined to any specific model fitting technique.
(v) R^2 should be such that its values for different models fitted to the same data set are directly comparable.
(vi) Relative values of R^2 ought to be generally compatible with those derived from other acceptable measures of fit (e.g. standard error prediction and root mean squared residual)
(vii) Positive and negative residuals ( ) should be weighted equally by R^2.

V. ALTERNATIVE STATISTICS

Goodness of fit of a model is generally assessed by the coefficient of determination, R^2. However, as pointed out by Kvalseth (1985), eight different expressions for R^2 appear in the literature,

$$\sum e_i^2 \sum (y_i - \hat{y}_i)^2 R_1{}^2 = 1 - \sum (y - \hat{y})^2 / \sum (y - \bar{y})^2$$

$$= 1 - \frac{\text{error variation}}{\text{total variation}} \quad (1) \qquad R_2{}^2 = \sum (\hat{y} - \bar{y})^2 / \sum (y - \bar{y})^2 = \frac{\text{explained variation}}{\text{total variation}}$$

$$(2)\ R_3{}^2 = \sum \left(\hat{y} - \bar{\hat{y}}\right)^2 / \sum (y - \bar{y})^2$$

$$(3) \qquad R_4{}^2 = 1 -$$

$\sum(e-\bar{e})^2/\sum(y-\bar{y})^2$, $e = y-\hat{y}$, and (4)

=

R_5^2 = squared multiple correlation coefficient between the response and the explanatory variables, i.e.

$$\frac{\{\sum(\hat{y}_i-\bar{\hat{y}}_i)(y_i-\bar{y})\}^2}{\{\sum(y_i-\bar{\hat{y}})\}^2\{\sum(y_i-\bar{y})\}^2}\hat{y} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

(5) R_6^2 = squared correlation coefficient between the response y and $\hat{y}\hat{y}$. i.e,

= (6)

$$R_7^2 = 1-\frac{\sum(y-\hat{y})^2}{\sum y^2} \quad (7)$$

$$R_8^2 = \frac{\sum\hat{y}^2}{y^2} \quad (8)$$

Numerical Example 1.

To exemplify the differences between the various R^2 statistics and some of the problems outlined earlier, two sets of data are used.

Considering the height in inches (y) and base diameter in inches (x) of 8 teak wood trees of a certain variety respectively, produced the first data set as follows.

x, *diameter in(Inches)*	1	2	3	4	5	6	7	8
y, *height in(Inches*)	37	72	97	105	147	155	191	206

(Data source: Student research project Department of Botany, Kaliabor college, Kuwaritol, Nagaon, Assam)

Table 1. Comparing the various Statistics for different types of models (*Numerical Example 1*)

Parameters/ Statistics	$\hat{y}$ =b_0+ b_1x	$\hat{y}$ = b_1x	$\hat{y}$ = $b_0x^{b_1}$	$\hat{y}$ = $b_0e^{b_1x}$	$\hat{y}$ = b_0 +$b_1\frac{1}{x}$
b_0	19.429	––	38.34	41.17	184.60
b_1	23.738	27.16	0.807	0.221	–172.0
R_1^2	0.9850	0.959	0.985	0.922	0.733
R_2^2	0.9850	1.295	0.981	1.308	0.744
R_3^2	0.9850	1.295	0.981	1.308	0.744
R_4^2	0.9850	0.959	0.985	0.898	0.7338
R_5^2	0.9850	0.980	0.980	0.985	0.985
R_6^2	0.9850	1.264	0.967	1.217	0.548
R_7^2	0.9976	0.993	0.997	0.983	0.9577
R_8^2	0.9976	0.993	0.995	1.055	0.958
RMSE	6.695	11.06	6.536	17.49	28.273
MAE	6.434	9.75	5.521	13.45	24.531
MSE	59.770	140.0	56.95	407.9	1065.8

Numerical Example 2.

The second data set is of one of the algal growth studies. The dry weights of the alga were measured in relation to varied doses of a pesticide. 7 observations are recoded as per different doses of pesticide concentration in (ppm) . Considering the pesticide conc. in (ppm) as x and variation in algal dry weight (mg/flask) as y respectively, the second data set is as follows.

x, *Pesticide Conc. In (ppm)*	0	10	20	30	40	50	60
y, *Algal dry wt. in (mg/flask)*	700	500	340	195	178	165	90

(Data source: "Biostatistics Theory and Applications" by G. B. N Chainy, G. Mishra, P. K. Mohanty, second edition, 2008, Example 11.4, Page no, 287.)

Table 2. Comparing the various statistics for different types of models (*Numerical Example 2.*)

Parameters/ Statistics	$\hat{y} = b_0+b_1x$	$\hat{y} = b_1x$	$\hat{y} = b_0e^{b_1x}$
b_0	594.929	------	657.207
b_1	-9.507	4.2219	-.0322
R_1^2	0.875	-1.7610	0.977
R_2^2	0.875	0.9837	0.856

R_3^2	0.875	0.1725	0.856
R_4^2	0.875	-0.9499	0.977
R_5^2	0.875	0.875	0.875
R_6^2	0.875	0.083	0.083
R_7^2	0.962	0.1688	0.993
R_8^2	0.962	0.1688	0.936
RMSE	71.806	337.728	30.774
MAE	61.754	242.888	24.130
MSE	7218.657	133070.89	1325.871

$$\left[\frac{\sum(y-\hat{y})^2}{n}\right]^{1/2}$$

$\frac{\sum|y-\hat{y}|}{n}$ Root Mean Squared Error (*RMSE*) = .

$\frac{\sum(y-\hat{y})^2}{n-p}$ Mean Absolute Error (*MAE*) = .

Mean Squared Error (*MSE*) = .

Where, *n* is the total number of observed values and *p* denotes the number of model parameters.

VI. ANALYSIS (FROM TABLE. 1 AND TABLE. 2)

(i) For linear intercept model in the Table 1 & Table 2, it is seen that

$$R_1^2 = R_2^2 = R_3^2 = R_4^2 = R_5^2 = R_6^2$$

(ii) For linear no-intercept model and non linear model in the Table 1, it is seen that R_2^2 and R_3^2 exceed 1. A similar finding was proposed by Kvalseth (1985) for power model.

iii) For linear no-intercept model, in the Table 2, it is seen that R_1^2 and R_4^2 become negative in some situations.

$\hat{y}\,\hat{y}\,\hat{y}\,\hat{y}\,\hat{y}$ (iv) For all 5 models viz., $=19.429 + 23.738x$, $= 27.1667x$, $\hat{y}\,\hat{y}\,\hat{y}$ $=38.340x^{0.8070}$, $= 41.1706e^{0.2218x}$, $= 184.6090 + (-172.049)1/x$ and all 3 models viz., $= 594.929+(-9.507)x$, $= 4.2219x$, $= 657.207e^{(-0.0322)x}$, the values of R_5^2 are same in the Table 1 and in the Table 2 respectively. Because R_5^2 measures the strength of the association between the response and explanatory variables when non linear model is transformed in to a linear model. Further the variables *x* & *y* of the *Numerical Example 1* are closely related and same thing happens in the *Numerical Example 2*.

(v) For no-intercept and exponential models in the Table 1, the value R_6^2 exceeds 1.

(vi) In the Table 1, for all 4 models except exponential model the value of R_7^2 and R_8^2 are different from their common value. Further in case of an exponential model, the value of R_8^2 exceeds one in the Table 1. The findings drawn here corroborate what Kvalseth (1985) drew for power model in his studies.

Table.3 Findings of Fitted models

(For** Numerical Example* 1)***

Fitted Models $\hat{y}$	$\bar{y}$	$\bar{\hat{y}}$	$\bar{e}$	Remarks
$= 19.429+ 3.738x$	126.25	126.25	0.0	$\bar{\hat{y}} = \bar{y}$ & $\bar{e} = 0$
$\hat{y} = 27.1667x$	126.25	122.25	3.99	$\bar{\hat{y}} \neq \bar{y}$ & $\bar{e} \neq 0$
$\hat{y} = 38.340x^{0.8070}$	126.25	126.25	0.149	$\bar{\hat{y}} = \bar{y}$ & $\bar{e} \neq 0$
$\hat{y} = 41.1706e^{0.2218x}$	126.25	1263	-0.496	$\bar{\hat{y}} = \bar{y}$ & $\bar{e} \neq 0$
$\hat{y} =184.6090+ (-172.049)1/x$	126.25	126.16	0.09	$\bar{\hat{y}} = \bar{y}$ & $\bar{e} \cong 0$

Table.4 Findings of Fitted models

(For Numerical Example 2 *)*

Fitted Models	$\bar{y}$	$\bar{\hat{y}}$	$\bar{e}$	Remarks
$\hat{y} = 594.92 + -9.507x$	309.71	309.72	0.0	$\bar{\hat{y}} = \bar{y}$ & $\bar{e} = 0$

$\hat{y} = 4.2219x$	309.71	126.65	183.05	$\bar{\hat{y}} \neq \bar{y}$ & $\bar{e} \neq 0$
$\hat{y} = 657.207e^{-0.032x}$	309.71	309.70	4.483	$\bar{\hat{y}} = \bar{y}$ & $\bar{e} \neq 0$

VII. SOME POSSIBLE SHORTCOMINGS OF USE OF R^2

(i) For linear no-intercept model and non linear model, it is seen that R_2^2 and R_3^2 exceed 1 (exemplified in the Table 1).

(ii) For linear no-intercept model, it is seen that R_1^2 and R_4^2 become negative in some situations (exemplified in the Table 2).

(iii) For the linear no - intercept model and for non linear models in the Table 1. & Table 2, it is recommended that R_4^2 is rejected, sinc a nonzero mean residual $\bar{e}$ should be regarded as contributing to a reduction in the model fit rather than to an increase as implied by

$\sum(y_i - \hat{y}_i)^2 \sum e_i^2 R_4^2 = 1 -$
$\sum(e - \bar{e})^2 / \sum(y - \bar{y})^2$, $e =$ $y - \hat{y}$, and = .

$\sum(y_i - \hat{y}_i)^2$(iv) $R_1^2 = 1 - \sum(y - \hat{y})^2 / \sum(y - \bar{y})^2$,

for no-intercept linear models may possibly be negative in some different situations where is large. (Montgomery and Peck 2003, pp, 47-48) i.e. $R_1^2 < 0$, but in linear case $0 \leq R_1^2 \leq 1$, in general.

(v) For no-intercept and exponential models, it is seen that R_6^2 exceeds 1 (exemplified in the Table 1).

(vi) Except exponential model the value of R_7^2 and R_8^2 are equivalent from their common value (exemplified in the Table 1).

(vi) $R_8^2 \geq 1$, for non linear models, (exemplified in the Table1).

VIII. CONCLUSION

The study of the present paper reveals that R^2 statistics are distinctly more sensitive to extreme values, as they contain terms involving squared residuals or squared values of the dependent variable or both. That is, these statistics have relatively low degree of resistance towards outliers in the data set.

R^2 values are found sensitive to nonlinear model and no-intercept linear model. This is reflected as the evidence in the example.

In the same manner the method of least square itself is not resistant one as it relies on equal weights to all observations. That is why R^2 values are less resistant. It is also clear from the above two examples that when resistant or robust model fitting technique is used for data containing no outliers or when clearly appropriate least squares regression model is used, R_1^2 is generally preferable than alternative R^2, which will lead to potentially expected results.

Our empirical findings are in conformity with those revealed by statisticians like Kvalseth [6], Hahn [2,3], Marquardt and Snee [8], Montgomery and Peck [7],Theil [10] etc.

REFERENCES

[1] Draper, N. R., and Smith. H. (1981), *Applied Regression Analysis (2nd ed.).*New York: John Wiley.

[2] Hahn, G. J. (1973), "The Coefficient of Determination Exposed," *Chemtech*, 3, 609-612.

[3] Hahn, G. J. (1977), "Fitting Regression Models With No Intercept Term." *Journal of Quality Technology*, 9, 56-61.

[4] Hawkins, D. M. (1980), " A Note on Fitting a Regression Without an Intercept Term," *The American Statistician*, 34, 233.

[5] Healy, M. J. R. (1984), "The Use of R2 as a Measure of Goodness of Fit," *Journal of Royal Statistical Society* ,Ser. A ,147 , 608-609.

[6] Kvalseth, T. O. (1985). "Cautionary Note about R2,"*The American Statistician*, Vol. 39, No.4, Part 1(Nov., 1985), pp. 279-285.

[7] Montgomery, D. C. and Peck, E. A. (2003), *Introduction to Linear Regression Analysis(3rd ed.).* New York :John Wiley.

[8] Marquardt, D. W., and Snee, R. D. (1974), " Test Statistics for Mixture Models," *Technometrics*, 16, 533-537.

[9] Scott, A. and Wild, C. (1991), "Transformations and R^2" *The American Statistician*, Vol. 45, No. 2 (1991), pp. 127-129.

[10] Theil, H. (1971), *Principles of Econometrics*, New York: John Willy.

[11] Willet, J. B., and Singer, J. D. (1998), "Another Cautionary Note about R^2: Its Use in Weighted Least-Squares Regression analysis," *The American Statistician*, Vol. 42, No. 3 (1991), pp. 236-238.

