International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013 1715
ISSN 2229-5518

Robust Coefficients of Determination: A Measure of Goodness of Fit

Md.Siraj-Ud-Doulah

Abstract— Various statistical methods, the R-square, modified R-square and adjusted R-square are the classical estimators for a wide range of commonly using measures of goodness of fit. They are however unreliable in presence of outliers. This paper proposes a new approach to robust modification of the coefficients of determination (RR-square). In this paper, I briefly review some of the more fundamental advantages and disadvantages with conventional as well as propose measure by utilizing a real data sets as well as Monte Carlo simulation. The proposed statistic is relatively good power than the classical measures for different sample sizes.

Index Terms— adjusted R-square, MR-square, R-square, RR-square, Simulation

---

1 INTRODUCTION

THE coefficient of determination (R^2) is one of the most popular goodness-of-fit tests employed in regression, economics, econometric and etc. The R^2 measures the information of the proportion or percentage of the total variation in Y explained by the regression model. Two properties of R^2 may be noted: (i). It is a nonnegative quantity and (ii) its limits are $0 \le R^2 \le 1$ [7 and 9] to name but a few. The main drawback of R^2: if we add a regressor variable to the model, R^2 increases [13]. But this does not mean the new model is superior to the old one. The theoretical and practical consequences of R^2, modified-R^2 (MR^2) and adjusted-R^2 ($\overline{R}^2$) have been documented in several books [5, 10, 11, 15 and 17] and journal articles [1, 2 and 16]. According to model selection criteria, we use all measures of goodness of fit as well as AIC and SIC, have already been studied extensively in the literature [3, 4, 8, 12 and 14]. All of these measures are based on R^2. For this reason, I make a new and simple measure of goodness of fit. I label it the robust coefficients of determination (RR^2), which is introduced in section 2. The properties of these classical and new measures are illustrated in section 3 with real life data sets. The performance of the classical and proposed RR^2 is investigated in section 4 through a Monte Carlo simulation experiment.

2. PROPOSE ROBUST COEFFICIENTS OF DETERMINATION

Let us now consider the regression line as a whole and examine its goodness of fit: $Y = \alpha + \beta X + e$. Suppose a sample regression line has been obtained by the method of least squares. $\sum y_i^2 = \hat{\beta}^2 \sum x_i^2 + \sum e_i^2$. The total variations are decomposed into two parts: (i) $\hat{\beta}^2 \sum x_i^2$: representing the estimated effect of X on the variations in Y. (ii) $\sum e_i^2$: representing the variations in Y which remain unexplained by the estimated relationship between Y and X. This decomposition of total variations in Y leads to a measure of the 'goodness of fit'- which is known as coefficient of determination and symbolized as R^2.

$$R^2 = \text{Variations explained/Variations required to be explained}$$

$$R^2 = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

This is the required classical measure of goodness of fit.
The corresponding robust coefficient of determination (RR^2) is given as follows:

$$RR^2 = 1 - \frac{median\left(d\{(|e_i| - |e_j|)|\}; 1 \le i < j \le n\right)}{median\left(d\{|y_i - y_j|\}; 1 \le i < j \le n\right)}$$

$$= 1 - \frac{\phi_e}{\phi_y} \qquad (1)$$

Where the constant d is selected to obtain a consistent estimator for σ, and equals d=2.219 at a normal model distribution.

It is easy to see in (1) that limits of RR^2 is zero and unity. If our fit is perfect, ϕ_e equal to zero as well as RR^2 equal to unity; indicating the best fit. At the other extreme if our estimated sample regression line is horizontal $(\beta = 0)$, then $\phi_e = \phi_y$ as well as RR^2 equal to zero. Thus, $0 \le RR^2 \le 1$.

3. EMPERICAL EXAMPLES

The well known data set consists of a research engineer is investigating the use of a windmill to generate electricity. He has collected data on the DC output from his windmill and the corresponding wind velocity (n=25), which has taken from [13]. Checking the goodness of fit of the fitted regression line to a set of data, that is, I will find out how well the sample re-

---

• *Lecturer, Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh. Email: sdoulah_brur@yahoo.com*

gression line fits the data. Next checking the outliers by the robust LTS method; it can detect 3 outliers (case 4, 8 and 25). Original data set and deleting these outliers, I recheck the appropriate model of the data sets, which results has shown below:

TABLE 1: PERFORMANCE OF DIFFERENT MEASURES FOR APPROPRIATE MODEL SELECTION

Model Selection Criteria	Data Type	Linear Model	Quadratic Model	Cubic Model	Reciprocal Model
R^2	WO	0.874	0.907	**0.986**	0.980
	WOO	0.902	0.967	**0.969**	0.965
MR^2	WO	0.804	0.851	**0.863**	0.841
	WOO	0.820	0.835	0.854	**0.878**
$\bar{R}^2$	WO	0.869	0.914	**0.973**	0.969
	WOO	0.897	0.964	**0.965**	0.963
AIC	WO	0.060	0.016	**0.013**	0.020
	WOO	0.027	0.011	0.010	**0.009**
SIC	WO	0.066	0.019	**0.016**	0.019
	WOO	0.029	0.011	0.012	**0.010**
RR^2	WO	0.834	**0.796**	0.881	**0.900**
	WOO	0.793	**0.790**	0.870	**0.883**

[Note: WO: With Outliers, WOO: Without Outliers]

From TABLE 1, shows an important property of R^2, MR^2 and $\bar{R}^2$ is that its are nondecreasing function as well as the property of AIC and SIC is that its are nonincreasing function when the number of explanatory variables or regressors added in the model but except RR^2. If the value of RR^2 decreases for adding regressors in the model, former model may correct or taking another form of models for appropriate selection. Notice that, when no outliers occurs in the data, the MR^2, AIC and SIC select the appropriate model. But, the only newly proposed measures of coefficient of determination $\left(RR^2\right)$ select the correct model when a small percentage of outliers are present or absent in the data set. According to [13], the reciprocal transformation model is appropriate for aforesaid data set.

As another example, I consider a famous data set found in [6] refers to the per capita consumption of cigarettes in various countries in 1930 and the death rates (number of deaths per million people) from lung cancer for 1950 (n=11). Checking the goodness of fit of the fitted regression line to a set of data, that is, I will find out how well the sample regression line fits the data. Next checking the outliers by the robust LTS method; it can detect 1 outlier (case 11). Actual data set and deleting these outliers, I revisit the appropriate model of the data sets, which results have shown in TABLE 2.

TABLE 2: PERFORMANCE OF DIFFERENT MEASURES FOR SUITABLE MODEL SELECTION

Model Selection Criteria	Data Type	Linear Model	Quadratic Model	Cubic Model	Reciprocal Model
R^2	WO	0.544	0.774	**0.899**	0.654
	WOO	0.888	0.905	**0.906**	0.799
MR^2	WO	0.444	0.563	**0.572**	0.535
	WOO	0.543	0.633	**0.711**	0.639
$\bar{R}^2$	WO	0.492	0.718	**0.856**	0.615
	WOO	0.874	0.878	**0.879**	0.773
AIC	WO	8203	4862	**2605**	6214
	WOO	**2279**	2374	2861	4123
SIC	WO	8819	5419	**3010**	6680
	WOO	**2422**	2600	3229	4381
RR^2	WO	**0.816**	**0.748**	0.802	0.730
	WOO	**0.837**	**0.754**	0.767	0.757

[Note: WO: With Outliers, WOO: Without Outliers]

From TABLE 2, when no outlier occurs in the data set, AIC and SIC select the correct model. But, the newly proposed RR^2 is most efficient measures of goodness of fit when an outlier present in the data set or absent in the data set. According to [6], linear model (LM) is appropriate for this data set.

4. REPORT OF MONTE CARLO SIMULATION

In this section, I discuss a Monte Carlo simulation study which is planned to evaluate the performance of the newly proposed RR^2 with five other popular and frequently used model selection criteria, i.e., the R^2, MR^2, $\bar{R}^2$, AIC and SIC. In order to compare the appropriate model identification power performance of R^2, MR^2, $\bar{R}^2$, AIC, SIC and RR^2, I simulate artificial data sets. So that, I find out from them who can caught the right model. The following procedures are as follows: I simulate data based on one model and run the data sets four aforementioned models as well as compare the percentage which measure can detect the correct model how many times. Firstly, I simulate linear model (LM) data and generating samples from uniform distribution. Next, I simulate quadratic model (QM) data and generating samples from same distribution. Again then, I simulate cubic model (CM) data and generating samples from aforesaid distribution. And then, I simulate reciprocal model (RM) data and generating samples from aforementioned distribution. In my simulation experiment, I have taken different sample sizes, n=50, 100, 200 and 500. Each experiment is run 10,000 times and the outcomes are given TABLE 3.

TABLE 3 POWER PERFORMANCE COMPARISONS OF DIFFERENT MODEL SELECTION CRITERIA

Model Selection Criteria	Power (in percentage) Linear Model	Quadratic Model	Cubic Model	Reciprocal Model
		n=50		
R^2	0	0	100	13.90
MR^2	0	0	100	13.90
$\overline{R}^2$	0	0	100	13.90
AIC	0	0	100	18.15
SIC	0	0	100	18.15
RR^2	**100**	**100**	**100**	**100**
		n=100		
R^2	0	0	100	16.81
MR^2	0	0	100	16.81
$\overline{R}^2$	0	0	100	16.81
AIC	0	0	100	19.26
SIC	0	0	100	19.26
RR^2	**100**	**100**	**100**	100
		n=200		
R^2	0	0	100	21.59
MR^2	0	0	100	21.59
$\overline{R}^2$	0	0	100	21.59
AIC	0	0	100	25.98
SIC	0	0	100	25.98
RR^2	**100**	**100**	**100**	100
		n=500		
R^2	0	0	100	29.09
MR^2	0	0	100	29.09
$\overline{R}^2$	0	0	100	29.09
AIC	0	0	100	36.89
SIC	0	0	100	36.89
RR^2	**100**	**100**	**100**	100

From TABLE 3 shows that the right model selection performance of R^2, MR^2, $\overline{R}^2$, AIC and SIC are necked zero for linear and quadratic models in different sample sizes. But correct model detection power of aforesaid tools is perfect for cubic model as well as identification power is very poor for reciprocal model in different samples. Alternatively, the correct model declaration power of my newly propose coefficient of determination (RR^2) is perfect for all models in different samples. Therefore, I can say that my newly proposed tool RR^2 is batter than any other techniques for exact model identification.

5. CONCLUSION

In this paper, shows that the proposed measure RR^2 appears to perform much better than the other measures of coefficient of determination for appropriate model selection. This technique has very good power against a variety of sizes and is capable of clear-cut selection of accurate model in regression and other applications. However, both the real data sets and simulation study demonstrate that the robust coefficient of determination (RR^2) is more accurate measure in a variety of situations. Since RR^2 perform superbly here and hence can be recommended to use an effective measure of goodness of fit.

REFERENCES

[1] A. Buse, "Goodness of Fit in Generalized Least Square Estimation," *American Statistician*, 27, pp. 106-108, 1973.
[2] S. Cameron, "Why is the R Squared Adjusted Reported?" *Journal of Quantitative Economics*, 9(1), pp. 183-186, 1993.
[3] N.D. Draper and H. Smith, "*Applied Regression Analysis*," Third Edition, John Wiley & Sons, New York, 1998.
[4] Dielman and E. Terry, "*Applied Regression Analysis for Business and Economics*," PWS-Kent, Boston, 1991.
[5] Davidson and James, "*Econometric theory*," Blackwell Publisher, Oxford, U.K, 2000.
[6] E.A. Freedman, "*Statistics*," John Wiley & Sons, 1991.
[7] W.H. Greene, "*Econometric Analysis*," Fifth Edition, New Jersey: Prentice-Hall, 2008.
[8] A.S.Goldberger, "*A Course in Econometrics*," Harvard University Press, Cambridge, Mass, 1991.
[9] D.N. Gujarati, "*Basic Econometrics*," Fourth Edition, McGraw Hill, New York, 2010.
[10] Hayashi and Fumio, "*Econometrics*," Princeton University Press, Princeton, N.J., 2000.
[11] G.M.K. Madnani, "*Introduction to Econometrics Principles and Applications*," Seventh Edition, Oxford & IBH Publishing Co. Pvt. Ltd, New Delhi, 2005.
[12] G. Maddala, "*Introduction to Econometrics*," Second Edition, Macmillan, New York, 1992.
[13] D.C. Montgomery and E.A. Peck, "*Introduction to Linear Regression Analysis*," John Wiley & Sons, 1981.
[14] H. Theil, "*Introduction to Econometrics*," Prentice Hall, Englewood Cliffs, N.J. 1978.
[15] Verbeek and Marno, "*A Guide to Modern Econometrics*," John Wiley & Sons, New York, 2000.
[16] F. Windmeijer, "*Goodness of Fit Measures in Binary Choice Models*," Econometric Reviews 14, 101-116, 1995.
[17] Wooldridge and M. Jeffrey, "*Introductory Econometrics*," South-Western College Publishing, 2000.