Criteria for selection of stochastic linear model selection

Cite as: AIP Conference Proceedings **2177**, 020041 (2019); https://doi.org/10.1063/1.5135216
Published Online: 04 December 2019

B. Mahaboob, J. Peter Praveen, Ranadheer Donthi, S. Vijay Prasad, and B. Venkateswarlu

Criteria for Selection of Stochastic Linear Model Selection

B.Mahaboob[1], J. Peter Praveen[1], Ranadheer Donthi[2], S. Vijay Prasad[1], B.Venkateswarlu[3, a)]

*[1]Department of Mathematics, Koneru Lakshmaih Education Foundation,
Vaddeswaram, Guntur Dist.AP-522502-India*
*[2]Department of Mathematics, St. Martin's Engineering College, Dhulapally,
Kompally, Hyderabad, Telangana, India.*
[3]Department of Mathematics, Vellore Institute of Technology, Vellore-632014

[a)] Corresponding author: venkatesh.reddy@vit.ac.in

Abstract. This research article proposes criteria for the selection of the best stochastic linear regression model. This is one of the three special problems of stochastic linear regression model namely, model selection, misspecification of the model and selection of regressors. Selection of the best model is an important part of stochastic model building. Alarge number of methods have been developed in the literature for selecting best stochastic linear regression model. Y. Tuac et al[7], in 2017, in his research article, presented a small simulation study and real data example to illustrate the performance of the proposed method for dealing with the variable selection and the parameter estimation in restricted linear regression models. Jussi Matta, [6], in his paper, studied model selection methods for two domains linear regression and phylogenetic reconstruction focussing particularly on situations where the amount of data available in either small or very large .Guoqui et al.[9] in their paper presented several model selection criteria which generally can be classified as the penalized robust method. B.M. Potscher, [10], in his research study presented the more general case of regression selection in stochastic linear regression model. Timoterasvirta et al.[8] in his research paper discussed the problem choosing a linear model from a set of nested alternatives.

INTRODUCTION

Stochastic model building has an important role in analysing different research problems in the various fields of science. Under stochastic model building the empirical determinative of certain laws can be expressed in the form of some linear and nonlinearmodels. In recent years a great deal of research has been directed on stochastic linear regression models disturb very often the optimum propertiesof OLS estimators of the parameters of the model. Besides two problems of hetroscedastic and auto correlated errors, these are three special problems of stochastic linear regression model namely, model selection, misspecification of the model and selection of the repressors. Selection of the best model is an important part of the stochastic model building. A large number of methods have been developed in the literature for selecting best stochastic linear regression model.Some commonly used selection methods are: Coefficient of Multiple Determination (R^2), Adjusted R^2 or $\overline{R}^2$, Prediction Mean Squared Error (S_p) criterion, Mallows C_p criterion etc.

COEFFICIENT OF MULTIPLE DETERMINATION CRITERION (OR) R2 CRITERION FOR STOCHASTIC LINEAR REGRESSION MODEL SELECTION.

In the applied stochastic model building, a commonly used selection criterion is the R^2 criterion.

Recent Trends in Pure and Applied Mathematics
AIP Conf. Proc. 2177, 020041-1–020041-5; https://doi.org/10.1063/1.5135216
Published by AIP Publishing. 978-0-7354-1924-7/$30.00

Consider the standard stochastic linear regression model

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \varepsilon_{n\times 1}$$

The OLS estimator $\hat{\beta}$ which is the best linear unbiased estimator (BLUE) of β is given by

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

Define OLS residual vector as

$$e = \left[Y - \hat{Y}\right] = \left[Y - X\hat{\beta}\right]$$

$$Y = X\hat{\beta} + e$$

$$\text{(or)}\, Y'Y = \left[X\hat{\beta} + e\right]' \left[X\hat{\beta} + e\right]$$

$$= \hat{\beta}^1 X'X\hat{\beta} + e'e + 2\hat{\beta}X'e \;\; (\Theta\, (e'X\hat{\beta})\;\; is\ a\;\; scalar.$$

$$\Rightarrow Y'Y = \hat{\beta}'X'X\hat{\beta} + e'e \;\; (X'e = 0)$$

If X contains a constant term (intercept), then the Analysis of Variance (ANOVA) model can be expressed as

$$\left[Total\ sum\ of\ \ squares\right] = \begin{bmatrix} \text{Re}gression\ sum\ of\ squares \\ orExplained\ sum\ of\ squares \end{bmatrix} + \begin{bmatrix} \text{Re}sidual\ sum\ of\ squares \\ or\ un\exp ected\ \ sum\ of\ squares \end{bmatrix} \Rightarrow \left[Y'Y - n\bar{Y}^2\right] = \left[\hat{Y}'\hat{Y} - n\bar{Y}^2\right] + e'e$$

The R^{2}statstic is defined as the ratio of the regression sum of squares to the total sum of squares

$$R^2 = \left[\frac{\hat{Y}'\hat{Y} - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}\right]$$

$$R^2 = \left[\frac{\hat{\beta}'X'X\hat{\beta} - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}\right] = \left[\frac{\hat{\beta}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}\right]$$

or

$$(\Theta\; X'X\hat{\beta} = X'Y)$$

If the standard stochastic linear regression model is defined in deviation form or without constant term (intercept) then the R^2 is defined as

$$R^2 = \left[\frac{\hat{\beta}'X'Y}{Y'Y}\right]$$

Here R^2 can be interpreted as a measure of proposition of the variance in Y which is explained by the estimated equation. In other words R^2 is a measure of proportion of total variance accounted for by the linear influence of the independent variables. The value of R^2 always lies between 0 and 1.i.e, $0 \leq R^2 \leq 1$.

The closer value of R^2 to 1, the better the performance of the independent variables

$$\mathrm{R}^2 = \frac{Cov^2(Y,\hat{Y})}{Var(Y)Var(\hat{Y})}$$

R^2 can be viewed as a measure of the stochastic linear regression model's predictive ability over the sample period.

A relationship between R^2 and F –test statistic is given by

$$F = \left[\frac{R^2/(k-1)}{(1-R^2)((n-k))}\right]$$

The inclusion of new independent variables will always increase the value of R^2

$$R^2 = \sum_{i=1}^{k} I(i) = \sum_{i=1}^{k}\left[R^2 - R_{(i)}^2\right]$$

where $I_{(i)} =$ Incremental contribution of every ith independent variable in the model

$R^2 =$ Coefficient of multiple determination with the inclusion of ith independent variable in the model

$R_{(i)}^2 = R^2$ with the exclusion of ith independent variable in the model

ADJUSTED R^2 OR $\overline{R}^2$ CRITERION FOR STOCHASTIC LINEAR REGRESSION MODEL SELECTION

Define R^2 statistic as $R^2 = \left[\frac{\hat{\beta}X'Y}{Y'Y}\right]$

$$\text{or}\, 1 - R^2 = \left[1 - \frac{\hat{\beta}X'Y}{Y'Y}\right] = \left[\frac{Y'Y - \hat{\beta}X'Y}{Y'Y}\right]$$

$$\Rightarrow 1 - R^2 = \left[\frac{ee'}{Y'Y}\right]$$

Consider unbiased estimator for error variance σ^2 and total variance (Y) as

$$\hat{\sigma}^2 = \frac{e'e}{n-k} \text{ and } Va\hat{r}(Y) = \frac{Y'Y}{n-1}$$

By adopting these unbiased estimators as adjustment factors in 1- R^2 , one may obtain the adjusted R^2 or $\overline{R}^2$ as

$$\left[1-\overline{R}^2\right]=\left[\frac{e'e/(n-k)}{Y'Y/(n-1)}\right]=\left[\frac{n-1}{n-k}\right]\left[\frac{e'e}{Y'Y}\right]$$

$$\Rightarrow\left[1-R^2\right]=\left[\frac{n-1}{n-k}\right]\left(1-R^2\right)$$

$$\text{or }\overline{R}^2=1-\left(\left(1-R^2\right)\left(\frac{n-1}{n-k}\right)\right)$$

$$\overline{R}^2=R^2-\left[\frac{k-1}{n-k}\right]\left(1-R^2\right)$$

Thus $\overline{R}^2<R^2$ except when k=1.

Also $\overline{R}^2$=1 $\Rightarrow\overline{R}^2=R^2$

It should be noted that $\overline{R}^2$ statistic is not an unbiased estimator. Further $\overline{R}^2$ sometimes may be negative. The concept of $\overline{R}^2$ was due to Theil (1961). Generally the value of $\overline{R}^2$ may be increased with the inclusion of additional independent variable whether include variable may be relevant or irrelevant variable. But the value of $\overline{R}^2$ may not be increased with the inclusion of additional independent variable if the additional included variable is irrelevant variable. $\overline{R}^2$ can be used as better criterion than R^2 for selecting the best stochastic linear regression model.

CONCLUSIONS

In the above monograph an attempt has been made by developing a criterion for stochastic linear regression model selection namely coefficient of multiple determination criterion or R^2 criterion. In addition to this adjusted R^2 or $\overline{R}^2$ criterion is also presented for stochastic linear regression model selection. Owing to the deficiencies of the criteria R^2 and $\overline{R}^2$ as they are not most powerful criteria based on Mean square Error prediction theGeneralized Mean Squared Error could be evaluated in the context of future research.

REFERENCES

1. Agostinelli. C, Statistics and Probability Letters, (2002).
2. Audley, R.J Psychol, and Rev.67:1-15, reprinted in Luce, R.D. Bush, R.R and Galanter. E. ads ," Readings in Mathematical Psychology", Wiley.(1963)
3. Byson J.T Morgan, "CRC Press, 978-1-58488-666-2, (2008).
4. Berry L. Nelson Mc-Graw-Hill, 978-007046213, (1995).
5. Kadane, J. B and Lazar, N. A, J. of the American Mathematical Association, **l99**, 279-290, (2004).
6. Justi Matta, Series of Publications, Report A (2016).
7. Y. Tuac, O. Arslan , Ankara University. (2017).
8. Timo Teravirta, and Ilikkamellin, Scandinavian journal of Statistics, 13, 159-171 (1988).

9. Qian Guoqi, Kunsch Hansruedi, ETH Zurich, Research collection, Research paper (1996).
10. B. M. Potschar, The Annals of Statistics, **17**, 1257-1274 (1989).