2321-8045 Science Park Research Journal Vol-2, Issue-10, 18th Sept 2014
Impact Factor : 1.6200[UIF-2013]

Original Article

A Diagnostic Test For Selection Of Linear Regression Models Using Compound Linear Models

M.Ramesh[1] , D.Giri [2] , B.Sarojamma[3] , P.Srivyshnavi [4] , B.Sireesha[5] and P.Balasiddamuni[6]

ABSTRACT

The problem of selecting the best linear regression model has long been of special interest to theoretical and applied statisticians. In the present study an attempt has been made by suggesting a diagnostic test for selection lo linear statistical models. The proposed criterion is conceptually much simpler then most of the existing model selection criteria, This can be readily be implemented using computer software and handle several alternative models simultaneously

Keywords:
Diagnostic Test , Linear Regression Models , Compound Linear Models , statistical analysis.

I.INTRODUCTION

Model selection is one of the basic problems of any statistical analysis. The usual setting for this problem is one in which a statistician has several possible models for data to arise from a given sampling study. More specifically the problem may be that of identifying the underlying model structure say, to obtain a correct model for optimization or prediction purposes or to estimate a set of parameters. Although the choice of the model(s) must take into account subject matter and other non-statistical aspects, data based statistical methods are very useful techniques in the selection procedure.The emphasis in statistics research presently has shifted from merely estimating a given model to choosing among competing models. In this shift, several Statisticians have made contributions to the selection techniques for linear statistical models.

M.Ramesh[1] , D.Giri [2] , B.Sarojamma[3] , P.Srivyshnavi [4] , B.Sireesha[5] and P.Balasiddamuni[6]

From
[1]Data Scientist, Tech Mahindra , Hyderabad, India.
[2]Principal,P.R.R & V.S Govt College,Vidavalur, SPSR Nellore, Andhra pradesh, India
[3]Assistant Professor,Department of Statistics, S.V.University, Tirupati, Andhra pradesh, India
[4]Academic Consultant, Department of IT, SPMVV, Tirupati, Andhra Pradesh, India
[5]Research Scholar, Department of Statistics, S.V.University, Tirupati, Andhra Pradesh,India
[6]Professor ,Department of Statistics, S.V.University, Tirupati, Andhra Pradesh, India

The Article Is Published On September 2014 Issue & Available At www.scienceparks.in

DOI:10.9780/23218045/1202013/49

The problem of selecting the best linear regression model has long been of special interest to theoretical and applied statisticians. In applied statistics, most research workers are facing with uncertainity as to the correct statistical models. Efforts to validate model estimates, and subsequent model search or revision procedures are consequences of recognition by applied research workers that they typically deal with false models. The classical problem of choosing between two linear regression models was studied by Lien and Vuong (1987).

In the present study an attempt has been made by proposing a simple criteria for choosing between two linear statistical models using studentized residuals.

II. CRITERION FOR TESTING TWO SEPARATE LINEAR REGRESSION MODELS

The choice between two linear regression models is a perennial problem in statistical analysis.

Consider two separate independent linear regression models under the two hypotheses as

$H_1 : Y = X\beta + \in, \qquad \in \sim N(0, \sigma_{\in}^2 I_n) \qquad \ldots (2.1)$

$H_2 : Y = Z\gamma + u, \qquad u \sim N(0, \sigma_u^2 I_n) \qquad \ldots (2.2)$

Where

Y, $\in$ and u are (nx1) vectors ;

X is (n x r) matrix of non-stochastic regressors with rank r;

Z is (n x s) matrix of non-stochastic regressors with rank S;

β is (r x 1) vector of unknown parameters;

γ is (s x 1) vector of unknown parameters;

H_1 and H_2 are two hypotheses, each contains regressors which can not be expressed as linear combinations of the regressors of other model.

By applying OLS estimation, the estimated equations are given by

i. $\hat{Y}_1 = X\hat{\beta}$, Where $\hat{\beta} = (X^{|}X)^{-|}X^{|}Y$

and ii. $\hat{Y}_2 = Z\hat{\gamma}$, Where $\hat{\gamma} = (Z^{|}Z)^{-|}Z^{|}Y$

obtain the OLS residual vectors as

$e_1 = [Y - X\hat{\beta}] = [Y - \hat{Y}_1]$

and $e_2 = [Y - Z\hat{\gamma}] = [Y - \hat{Y}_2]$

Consider the new linear regression models,

(a) $Y = X\beta + \alpha e_2 + \nu_1$, , $\nu_1 \sim N(0, \sigma^2_{\nu_1} I_n)$ ... (2.3)

and (b) $Y = Z\gamma + \delta e_1 + \nu_2$, , $\nu_2 \sim N(0, \sigma^2_{\nu_2} I_n)$ ... (2.4)

Where, δ_1 and δ_2 are the regression coefficients of e_1 and e_2 respectively. Now, estimate these two linear regression models by using the OLS method and test for the statistical significance of α and δ in the above models (2.3) and (2.4) by using the t-test.

α is significant $\Rightarrow$ Rejecting H_1 and accepting the linear model under H_2.

is significant $\Rightarrow$ Rejecting H_2 and accepting the linear model under H_1.

This test can be further extended by the following procedure :

By introducing Quadratic and Cubic forms of residuals as additional regressors, the models can be written as

(i) $Y = X\beta + \alpha_1 e_2^2 + \alpha_2 e_2^3 + u_1$, $u_1 \sim N(0, \sigma^2_{u_1} I_n)$ ... (2.5)

and (ii) $Y = Z\gamma + \delta_1 e_1^2 + \delta_2 e_1^3 + u_2$, $u_2 \sim N(0, \sigma^2_{u_2} I_n)$ ... (2.6)

We obtain the coefficients of determination R^2 from the regressions (2.1), (2.2), (2.5) and (2.6) as R_I^2, R_{II}^2, R_I^{2*} and R_{II}^{2*} respectively.

For the selection between two linear regression models, we compute the F-statistics as follows :

(a) $$F_I = \frac{\left[R_I^{2*} - R_I^2\right] / 2}{\left[1 - R_I^{2*}\right] / (n - r - 2)} \sim F_{[2,(n-r-2)]} \quad (2.7)$$

(b) $$F_{II} = \frac{\left[R_{II}^{2*} - R_{II}^2\right] / 2}{\left[1 - R_{II}^{2*}\right] / (n - s - 2)} \sim F_{[2,(n-s-2)]} \quad (2.8)$$

F_I is significant $\Rightarrow$ Rejecting H_1 and accepting model under H_2.

F_{II} is significant $\Rightarrow$ Rejecting H_2 and accepting model under H_1.

We regress e_2 on all X – regressors and e_1 on Z – regressors as follows :

(i) $e_2 = X\Gamma + W_1$, $W_1 \sim N(0, \sigma^2_{w_1} I_n)$ ... (2.9)

and (ii) $e_1 = Z\xi + W_2$, $W_2 \sim N(0, \sigma^2_{w_2} I_n)$ ... (2.10)

Now, we obtain R^2 by using regressions (2.9) and (2.10) as R_I^{2**} , R_{II}^{2**} ; and use F-test for their significance :

(i) $$F_I^{**} = \left[\frac{R_I^{2**} / r-1}{\left(1 - R_I^{2**}\right) / n-r} \right] \sim F_{[(r-1,\ n-r]} \quad \ldots (2.11)$$

and (ii) $$F_{II}^{**} = \left[\frac{R_{II}^{2**} / S-1}{\left(1 - R_{II}^{2**}\right) / n-S} \right] \sim F_{[(s-1,\ n-s]} \quad \ldots (2.12)$$

F_I^{**} is significant $\Rightarrow$ Rejecting H_1 and accepting model under H_2.

F_{II}^{**} is significant $\Rightarrow$ Rejecting H_2 and accepting model under H_1.

III. SELECTION CRITERION FOR CHOOSING BETWEEN TWO LINEAR STATISTICAL MODELS BY USING STUDENTIZED RESIDUALS

It is very often in empirical research that the applied statistician faces the task of testing separate hypotheses. The poineering work on testing such hypotheses was carried out by Cox (1962) and his general results have been applied to various models of interest to statisticians. Pesaran (1974) derived statistics for testing non-nested single equation linear regression models; and Pesaran and Deaton (1978) obtained the corresponding expressions for systems of separate nonlinear regression models.

Consider two linear models each of which is to be estimated by the OLS method :

$M_1 : Y = X\beta + \in,$ $\in \sim N(0, \sigma_\in^2 I_n)$ ... (3.1)

$M_2 : Y = Z\gamma + u,$ $u \sim N(0, \sigma_u^2 I_n)$ ... (3.2)

Where Y , $\in$ and u are (nx1) vectors;

X is (n x r) matrix with rank r;

Z is (n x s) matrix with rank s.

Suppose the regressors of the Models M_1 and M_2 are nonstochastic. OLS estimation can be used to estimate the parameters of these two models. The OLS estimators of β and γ are given by

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

and $\hat{\gamma} = (Z'Z)^{-1} Z'Y$ respectively.

The validity of M_1 could be checked by testing $H_0 : \alpha = 0$ in

$Y = X\beta + A\alpha + \in$ … (3.3)

Where A being (n x p) matrix of selected test variables.

The relevant test statistics are under appropriate conditions, asymptotically distributed as χ^2_p , when $\alpha = 0$.

A general approach which does use information about M_1 and M_2 when testing either of them involves forming a comprehensive model

M_3 : $Y = X\beta + \alpha Z\ \hat{\gamma} + \nu_1$ … (3.4)

or $Y = Z\gamma + \delta X\ \hat{\beta} + \nu_2$ … (3.5)

and carrying out tests of $\alpha = 0$ and $\delta = 0$.

Define the projection matrices for the two linear regression models as

$P_X = [I - X (X^{|}X)^{-|} X^{|}]$ … (3.6)

$P_Z = [I - Z (Z^{|}Z)^{-|} Z^{|}]$ … (3.7)

Assume that Plim $\left(\frac{X^{|} Z}{n}\right)$, Plim $\left(\frac{X^{|} P_Z X}{n}\right)$ and

Plim $\left(\frac{Z^{|} P_X Z}{n}\right)$ exist and are not null matrices.

We have, $P_Z X\ \hat{\beta} = [I - Z (Z^{|}Z)^{-|} Z^{|}] X\ \hat{\beta}$

$= X\hat{\beta} - Z (Z^{|}Z)^{-|} Z^{|} X\ \hat{\beta}$

$= X\hat{\beta} - Z\ \hat{\Gamma}$,

$= e^{**}_X$

Where, $\hat{\Gamma} = (Z^{|}Z)^{-|} Z^{|} (X\hat{\beta})$

Here, $X\hat{\beta}$ is regressed on Z i.e.,

$X\hat{\beta} = Z\Gamma + \xi$ ... (3.8)

$\xi \sim N(0, \sigma_{\xi}^2 I_n)$

e_X^{**} is the OLS residual vector from regression (3.8).

Similarly, we can have,

$P_X Z\hat{\gamma} = [I - X(X'X)^{-1}X'] Z\hat{\gamma}$

$= Z\hat{\gamma} - X(X'X)^{-1}X' Z\hat{\gamma}$

$= Z\hat{\gamma} - X\hat{\eta}$

$= e_Z^{**}$ ... (3.9)

Where $\hat{\eta} = (X'X)^{-1}X' Z\hat{\gamma}$.

Here, $Z\hat{\gamma}$ is regressed on X.

i.e., $Z\hat{\gamma} = X\eta + w$, $w \sim N(0, \sigma_w^2 I_n)$ ... (3.10)

e_Z^{**} is the OLS residual vector from regression..... (3.10)

Define the Internally studentized residuals as

$$\tilde{e}_{X_i} = \frac{e_{X_i}^{**}}{\hat{\sigma}_{\xi}\sqrt{(1-h_{ii}(X))}}, \qquad i=1, 2, \ldots, n$$

and $$\tilde{e}_{Z_i} = \frac{e_{Z_i}^{**}}{\hat{\sigma}_{\omega}\sqrt{(1-h_{ii}(Z))}} \qquad i=1, 2, \ldots, n$$

Where $\hat{\sigma}_{\xi}^2 = \frac{\sum e_{X_i}^{**2}}{n-s}$ and $\hat{\sigma}_{\omega}^2 = \frac{\sum e_{Z_i}^{**2}}{n-r}$.

Here, H(X) = $\left(\left(h_{ij}(X)\right)\right)$ and H(Z) = $\left(\left(h_{ij}(Z)\right)\right)$ are the Hat matrices of the regressions (3.8) and (3.10) respectively.

Also denote $\tilde{e}_X$ and $\tilde{e}_Z$ are the internally studentized residual vectors.

For testing the validity of two models, we consider the two linear regression models with artificial test variables ($\tilde{e}_X$ and $\tilde{e}_Z$) as

$Y = X\beta + \tilde{e}_X\ \theta_1 + \in$ ... (3.11)

$Y = Z\gamma + \tilde{e}_Z\ \theta_2 + u$ ... (3.12)

The validity of M_1 can be tested by testing $H_0 : \theta_1 = 0$

When the alternative is

$H_1 : Y = X\beta + \tilde{e}_X\ \theta_1 + \in$

Here, the t-statistic for testing H_0: $\theta_1=0$ will have a central t-distribution with (n–r–1) degrees of freedom, when the null hypothesis is true.

Similarly the validity of M_2 can be tested by testing $H_0 : \theta_2=0$. When the alternative is

$H_1 : Y = Z\gamma + \tilde{e}_Z\ \theta_2 + u$

Here, the t-statistic for testing $H_0 : \theta_2 = 0$ will have a central t-distribution with (n-s-1) degrees of freedom, when the null hypothesis is true

IV. MODEL SELECTION BETWEEN TWO LINEAR STATISTICAL MODELS BY USING COMPOUND LINEAR STATISTICAL MODELS

Consider two linear regression models as

(i) $M_I : Y = X_1\beta_1 + \in_1$ ... (4.1)

and (ii) $M_{II} : Y = X_2\beta_2 + \in_2$ ... (4.2)

Where Y is (nx1) vector of observations on dependent variable;

X_1 and X_2 are (nxK_1) and (nxk_2) data matrices of known constants respectively;

β_1 and β_2 are (k_1x1) and (k_2x1) vectors of unknown parameters respectively;

and $\in_1$ and $\in_2$ are (nx1) vectors of disturbances.

It is assumed that X_1, X_2 are non-stochastic; and $\in_1$ ~N (0, σ_1^2 I_n) and $\in_2$ ~N (0, σ_2^2 I_n).

Here, σ_1^2 and σ_2^2 are unknown error variances. Now one can think of M_I and M_{II} as special cases of the linear regression model,

M_{III} : $Y = X_1 \beta_1 + X_2 \beta_2 + \in$, $\in$~N (0, $\sigma^2 I_n$) ... (4.3)

When $\beta_2 = 0$, M_{III} reduces to M_I ;

and $\beta_1 = 0$, M_{III} reduces to M_{II} .

Under the proposed criterion, one can consider the following two tests :

(a) H_0 : $Y = X_1 \beta_1 + \in_1$

H_1 : $Y = X_1 \beta_1 + X_2 \beta_2 + \in$

To test H_0, the F – statistic is given by

$$F = \frac{(n-k_1-k_2)(\hat{\sigma}_1^2 - \hat{\sigma}^2)}{k_2\ \hat{\sigma}_2} \quad \text{... (4.4)}$$

Where $\hat{\sigma}_1^2 = \frac{e_1^{\square 1} e_1^{\square}}{n-k_1}$

$\hat{\sigma}^2 = \frac{e^{\square 1} e^{\square}}{n-k}$, $k = k_1+k_2$

Here,

$\tilde{e}_1^1 \tilde{e}_1$ = Internally studentized Residual sum of squares obtained by estimating the model $Y= X_1 \beta_1 + \in_1$

$\tilde{e}_1^1 \tilde{e}_1$ = Internally studentized Residual sum of squares obtained by estimating the model $Y= X_1 \beta_1 + X_2 \beta_2 + \in$

Choose the model M_I : $Y = X_1 \beta_1 + \in_1$

If $$F \le \left[\frac{2(n-1)(n-k_1-k_2)}{(n+k_1)(n-k_1-k_2-2)} \right] \quad \text{... (4.5)}$$

(b) H_0 : $Y = X_2 \beta_2 + \in_2$

H_1 : $Y = X_1 \beta_1 + X_2 \beta_2 + \in$

To test H_0, the F – statistic is given by

$$F = \frac{(n - k_1 - k_2)(\hat{\sigma}_2^2 - \hat{\sigma}^2)}{k_1 \hat{\sigma}^2} \qquad \ldots (4.6)$$

V. CONCLUSIONS

Choosing between two linear models is an important topic in linear regression analysis. The selection of the linear statistical model that is consistent with the sampling process whereby the data are generated, is an old and important problem in statistics. For the last four decades, the various criteria, search processes, empirical rules and testing methods have been proposed as aids in the choice process.

In the present study, an attempt has been made by suggesting some criteria for selection between two linear statistical models by using the Internally studentized residuals.

The proposed criteria are conceptually much simpler than most of the existing model selection criteria, can readily be implemented using computer software and can handle several alternative models simultaneously.

BIBLIOGRAPHY

1. Amemiya, T.C. (1980), "Selection of Regressors", *International Economic Review*, 21, 331-354.

2. Atkinson, A.C. (1970), "A Method for Discriminating Between Models", *Journal of the Royal Statistical Society, Series B, 32,* 323-345.

3. Chow, G.C. (1988), "Econometrics", IV Printing, Mc Graw-Hill Book Company Tokyo.

4. Cox, D.R. (1962), "Further Results on Tests of Separate Families of Hypothesis," *Journal of the Royal Statistical Society, Series B, 24,* 406-424.

5. Draper, N.R. and Smmith, H. (1998), "*Applied Regression Analysis*", (Third Edition), John Wiley & Sons, Inc., New York.

6. Gelfand, A.E. and Ghosh, S.K. (1998), "Model Choice : A Minimum Posterior Predictive Loss Approach", *Biometrika*, 85, 1-11.

7. Giri,D.(2006), “Some New Selection Techniques for Linear Statistical Models”, unpublished Ph.D., Thesis, S.V. University, Tirupati, Andhra Pradesh State, India.

8. Hill, R.C., Judge, G.G., and Fomby, T.B. (1978), “On Testing the Adequacy of a Regression Model”, *Technometrics*, 20, No.4, 491-494.

9. Lien, D. and Q.H. Vuong (1987), "Selecting the Best Linear Regression Model: A Classical Approach, *Journal of Econometrics, 35*, 3-23.

10. Novotny, T.J., and McDonald, L.L. (1986), “Model Selection Using Discriminate Analysis”, *Journal of Applied Statistics*, 13, 159-165.

11. Pesaran, M.H. (1974), “On the General Problem of Model Selection”, *Review of Economic Studies*, 41, 153-171.

12. Pesaran, M.H., and A.S. Deaton (1978), “Testing Non-nested Nonlinear Regression Model”, *Econometrica*, 46, 677-694.