LINEAR REGRESSION ANALYSIS

MODULE – XIII

Lecture - 39

Variable Selection and Model Building

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

5. Akaike's information criterion (*AIC*)

The Akaike's information criterion statistic is given as

$$AIC_p = n\ln\left(\frac{SS_{res}(p)}{n}\right) + 2p$$

where

$$SS_{res}(p) = y'H_1y = y'X_1(X_1'X_1)^{-1}X_1'y$$

is based on the subset model $y = X_1\beta_1 + \delta$ derived from the full model $y = X_1\beta_1 + X_2\beta_2 + \varepsilon = X\beta + \varepsilon$.

Now we derive the given expression for *AIC*. In general, the *AIC* is defined as

AIC = -2(maximized log likelihood) + 2(number of parameters).

In linear regression model with $\varepsilon \sim N(0, \sigma^2 I)$, the likelihood function is

$$L(y, \beta, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma^2}\right]$$

and log – likelihood is

$$\ln L(y; \beta, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma^2}.$$

The log-likelihood is maximized at

$$\tilde{\beta} = (X'X)^{-1}X'y$$

$$\tilde{\sigma}^2 = \frac{n - p}{n}\hat{\sigma}^2$$

where $\tilde{\beta}$ is maximum likelihood estimate of β which is same as OLSE, $\tilde{\sigma}^2$ is maximum likelihood estimate of σ^2 and $\hat{\sigma}^2$ is OLSE of σ^2.

So

$$AIC = -2\ln L(y; \tilde{\beta}, \tilde{\sigma}^2) + 2p$$
$$= n\ln\left(\frac{SS_{res}}{n}\right) + 2p + n\left[\ln(2\pi) + 1\right]$$

where

$$SS_{res} = y'\left[I - X(X'X)^{-1}X'\right]y.$$

The term $n\left[\ln(2\pi)+1\right]$ remains same for all the models under comparison if same observations y are compared. So it is irrelevant for *AIC*. Hence we get the required expression for *AIC*. A model with smaller value of *AIC* is preferable.

6. Bayesian information criterion (BIC)

Similar to *AIC,* the Bayesian information criterion is based on maximizing the posterior distribution of model given the observations y. In the case of linear regression model with p selected explanatory variables is defined as

$$BIC = n\ln(SS_{res}) + (p-n)\ln n.$$

A model with smaller value of *BIC i*s preferable.

7. PRESS statistic

Since the residuals and residual sum of squares act as a criterion for subset model selection, so similarly the PRESS residuals and prediction sum of squares can also be used as a basis for subset model selection. The usual residual and PRESS residuals have their own characteristics which are used is regression modeling.

The PRESS statistic based on a subset model with p explanatory variable is given by

$$PRESS(p) = \sum_{i=1}^{n}\left[y_i - \hat{y}_{(i)}\right]^2$$

$$= \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

where h_{ii} is the i^{th} element in $H = X(X'X)^{-1}X$. This criterion is used on the similar lines as in the case of $SS_{res}(p)$.

A subset regression model with smaller value of $PRESS(p)$ is preferable.

Partial *F*- statistic

The partial *F*-statistic is used to test the hypothesis about a subvector of the regression coefficient. Consider the model

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

which includes an intercept term and (*k – 1)* explanatory variables. Suppose a subset of *r* < (*k* - 1) explanatory variables is to be obtained which contribute significantly to the regression model. So partition

$$X = \left(X_1 \;\; X_2 \right), \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

where X_1 and X_2 are matrices of order $n\times(k-r)$ and $n\times r$ respectively; β_1 and β_2 are the vectors of order $(k-r)\times 1$ and $r\times 1$ respectively.

The objective is to test the null hypothesis

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0.$$

Then

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

is the full model and application of least squares gives the OLSE of β as

$$b = (X'X)^{-1}X'y.$$

The corresponding sum of squares due to regression with *k* degrees of freedom is

$$SS_{reg} = b'X'y$$

and the sum of squares due to residuals with (*n* - *k*) degrees of freedom is

$$SS_{res} = y'y - b'X'y$$

and $MS_{res} = \dfrac{y'y - b'X'y}{n-k}$ is the mean square due to residual.

The contribution of explanatory variables in β_2 in the regression can be found by considering the full model under $H_0 : \beta_2 = 0$. Assume that $H_0 : \beta_2 = 0$ is true, then the full model becomes

$$y = X_1\beta_1 + \delta,\ E(\delta) = 0,\ Var(\delta) = \sigma^2 I$$

which is the reduced model. Application of least squares to reduced model yields the OLSE of β_1 as

$$b_1 = (X_1'X_1)^{-1}X_1'y$$

and corresponding sum of squares due to regression with $(k-r)$ degrees of freedom is

$$SS_{reg} = b_1'X_1'y.$$

The sum of squares of regression due to β_2 given that β_1 in already in the model can be found by

$$SS_{reg}(\beta_2 \mid \beta_1) = SS_{reg}(\beta) - SS_{reg}(\beta_1)$$

where $SS_{reg}(\beta)$ and $SS_{reg}(\beta_1)$ respectively are the sum of squares due to regression with all explanatory variables corresponding to β is the model and the sum of squares due to explanatory variables corresponding to β_1 in the model.

The term $SS_{reg}(\beta_2 \mid \beta_1)$ is called as the **extra sum of squares** due to β_2 and has degrees of freedom $k-(k-r) = r.$ It is independent of MS_{res} and is a measure of regression sum of squares that results from adding the explanatory variables $X_{k-r+1},...,X_k$ in the model when the model has already $X_1, X_2,..., X_{k-r}$ explanatory variables.

The null hypothesis $H_0 : \beta_2 = 0$ can be tested using the statistic

$$F_0 = \frac{SS_{res}(\beta_2 \mid \beta_1)/r}{MS_{res}}$$

which follows *F* - distribution with *r* and (*n* – *k*) degrees of freedom under H_0. The decision rule is to reject H_0 whenever

$$F_0 > F_\alpha(r, n-k).$$

This is known as the **partial *F* – test.** It measures the contribution of explanatory variables in X_2 given that the other explanatory variables in X_1 are already in the model.

Computational techniques for variable selection

In order to select a subset model, several techniques based on computational procedures and algorithm are available. They are essentially based on two ideas – select all possible explanatory variables or select the explanatory variables stepwise.

1. Use all possible explanatory variables

This methodology is based on following steps:

- Fit a model with one explanatory variable.
- Fit a model with two explanatory variables.
- Fit a model with three explanatory variables.

and so on.

Choose a suitable criterion for model selection and evaluate each of the fitted regression equation with the selection criterion. The total number of models to be fitted sharply rises with increase in k. So such models can be evaluated using model selection criterion with the help of an efficient computation algorithm on computers.

2. Stepwise regression techniques

This methodology is based on choosing the explanatory variables in the subset model in steps which can be either adding one variable at a time or deleting one variable at a times. Based on this, there are three procedures.

- forward selection
- backward elimination and
- stepwise regression.

These procedures are basically computer intensive procedures and are executed using a software.

Forward selection procedure

This methodology assumes that there is no explanatory variable in the model except an intercept term. It adds variables one by one and test the fitted model at each step using some suitable criterion. It has following steps.

- Consider only intercept term and insert one variable at a time.
- Calculate the simple correlations of $x_i's (i = 1,2,...,k)$ with y.
- Choose x_i which has largest correlation with y.
- Suppose x_1 is the variable which has highest correlation with y. Since F - statistic given by $F_0 = \frac{n-k}{k-1}.\frac{R^2}{1-R^2}$, so x_1 will produce the largest value of F_0 in testing the significance of regression.
- Choose a prespecified value of F value, say F_{IN} (F – to – enter).
- If $F > F_{IN}$, then accept x_1 and so x_1 enters into the model.
- Adjust the effect of x_1 on and re-compute the correlations of remaining x_i 's with y and obtain the partial correlations as follows.
  - Fit the regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ and obtain the residuals.
  - Fit the regression of x_1 on other candidate explanatory variables as
    $$\hat{x}_j = \hat{\alpha}_{oj} + \hat{\alpha}_{1j} x_1, \quad j = 2,3,...,k$$
    and obtain the residuals.
  - Find the simple correlation between the two residuals.
  - This gives the partial correlations.
- Choose x_i with second largest correlation with y, i.e., the variable with highest value of partial correlation with y.
- Suppose this variable is x_2. Then the largest partial F - statistic is

$$F = \frac{SS_{reg}(x_2 \mid x_1)}{MS_{res}(x_1, x_2)}.$$

- It $F > F_{IN}$ then x_2 enters into the model.
- These steps are repeated. At each step, the partial correlations are computed and explanatory variable corresponding to highest partial correlation with y is chosen to be added into the model. Equivalently, the partial F -statistics are calculated and largest F - statistic given the other explanatory variables in the model is chosen. The corresponding explanatory variable is added into the model if partial F-statistic exceeds F_{IN}.
- Continue with such selection as long as either at particular step, the partial F - statistic does not exceed F_{IN} or when the last explanatory variable is added to the model.

Note: The SAS software choose F_{IN} by choosing a type I error rate α so that the explanatory variable with highest partial correlation coefficient with y is added to model if partial F - statistic exceeds $F_{\alpha}(1, n-p)$.

Backward elimination procedure

This methodology is contrary to forward selection procedure. The forward selection procedure starts with no explanatory variable in the model and keep on adding one variable at a time until a suitable model is obtained .

The backward elimination methodology begins with all explanatory variables and keep on deleting one variable at a time until a suitable model is obtained.

It is based on following steps:

- Consider all k explanatory variables and fit the model.
- Compute partial F - statistic for each explanatory variables as if it were the last variable to enter in the model.
- Choose a preselected value F_{OUT} (F - to-remove).
- Compare the smallest of the partial F - statistics with F_{OUT}. If it is less than F_{OUT}, then remove the corresponding explanatory variable from the model.
- The model will have now (k - 1) explanatory variables.
- Fit the model with these (k - 1) explanatory variables, compute the partial F - statistic for the new model and compare it with F_{OUT}. If it is less them F_{OUT}, then remove the corresponding variable from the model.
- Repeat this procedure.
- Stop the procedure when smallest partial F - statistic exceeds $F_{OUT.}$

Stepwise regression procedure

A combinations of forward selection and backward elimination procedure is the stepwise regression. It is a modification of forward selection procedure and has following steps.

- Consider all the explanatory variables entered into to the model at previous step.
- Add a new variable and regresses it via their partial F - statistics.
- An explanatory variable that was added at an earlier step may now become insignificant due to its relationship with currently present explanatory variables in the model.
- If partial F - statistic for an explanatory variable is smaller than F_{OUT}, then this variable is deleted from the model.
- Stepwise needs two cut-off values, F_{IN} and F_{OUT} . Sometimes $F_{IN} = F_{OUT}$ or $F_{IN} > F_{OUT}$ are considered. The choice $F_{IN} > F_{OUT}$ makes relatively more difficult to add an explanatory variable than to delete one.

General comments

1. None of the methods among forward selection, backward elimination or stepwise regression guarantees the best subset model.
2. The order in which the explanatory variables enter or leave the models does not indicate the order of importance of explanatory variable.
3. In forward selection, no explanatory variable can be removed if entered in the model. Similarly in backward elimination, no explanatory variable can be added if removed from the model.
4. All procedures may lead to different models.
5. Different model selection criterion may give different subset models.

Comments about stopping rules

- Choice of F_{IN} and/or F_{OUT} provides stopping rules for algorithms.
- Some computer software allows analyst to specify these values directly.
- Some algorithms require type I errors to generate F_{IN} or/and F_{OUT}. Sometimes, taking α as level of significance can be misleading because several correlated partial F - variables are considered at each step and maximum among them is examined.
- Some analyst prefer small values of F_{IN} and F_{OUT} whereas some prefer extreme values.

  Popular choice is $F_{IN} = F_{OUT} = 4$ which is corresponding to 5% level of significance of F - distribution.