LINEAR REGRESSION ANALYSIS

MODULE – XIII

Lecture - 38

Variable Selection and Model Building

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Evaluation of subset regression model

A question arises after the selection of subsets of candidate variables for the model, how to judge which subset yields better regression model. Various criteria have been proposed in the literature to evaluate and compare the subset regression models.

1. Coefficient of determination

The coefficient of determination is the square of multiple correlation coefficient between the study variable *y* and set of explanatory variables $X_1, X_2, ..., X_p$ denotes as R_p^2. Note that $X_{i1} = 1$ for all $i = 1, 2, ..., n$ which simply indicates the need of intercept term in the model without which the coefficient of determination can not be used. So essentially, there will be a subset of $(p-1)$ explanatory variables and one intercept term in the notation R_p^2.

The coefficient of determination based on such variables is

$$R_p^2 = \frac{SS_{reg}(p)}{SS_T}$$

$$= 1 - \frac{SS_{res}(p)}{SS_T}$$

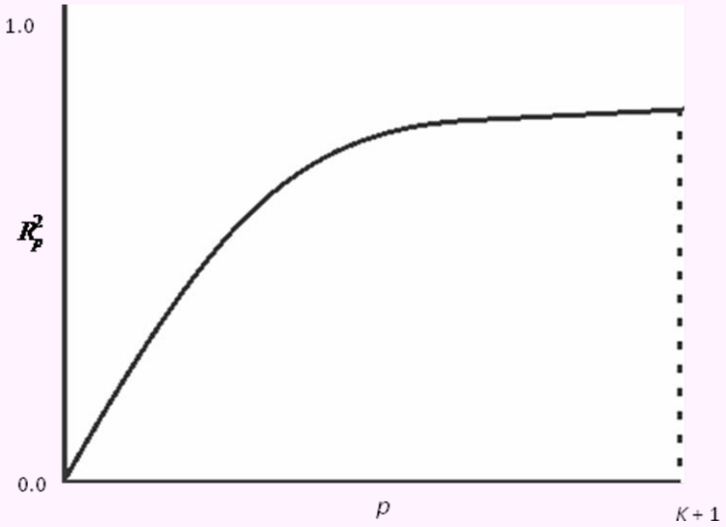
where $SS_{reg}(p)$ and $SS_{res}(p)$ are the sum of squares due to regression and residuals, respectively in a subset model based on $(p-1)$ explanatory variables.

Since there are *k* explanatory variables available and we select only $(p-1)$ out of them, so there are $\binom{k}{p-1}$ possible choices of subsets. Each such choice will produce one subset model. Moreover, the coefficient of determination has a tendency to increase with the increase in *p.*

So proceed as follows:

- Choose any appropriate value of *p*, fit the model and obtain R_p^2.
- Add one variable, fit the model and again obtain R_{p+1}^2 .
- Obviously $R_{p+1}^2 > R_p^2$. If $R_{p+1}^2 - R_p^2$ is small, then stop and choose the value of *p* for subset regression.
- If $R_{p+1}^2 - R_p^2$ is high, then keep on adding variables upto a point where an additional variable does not produces a large change in the value of R_p^2 or the increment in R_p^2 becomes small.

To know such value of *p*, create a plot of R_p^2 versus p. For example, the curve will look like as in the following figure



Choose the value of *p* corresponding to a value of R_p^2 where the "knee" of the curve is clearly seen. Such choice of may not be unique among different analyst. Some experience and judgment of analyst will be helpful is finding the appropriate and satisfactory value of *p*.

To choose a satisfactory value analytically, a solution is a test which can identify the model with R^2 which does not significantly differ from the R^2 based on all the explanatory variables.

Let

$$R_0^2 = 1-(1-R_{k+1}^2)(1+d_{\alpha,n,k})$$

where

$$d_{\alpha,n,k} = \frac{kF_\alpha(n,n-k-1)}{n-k-1}$$

and R_{k+1}^2 is the value of R^2 based on all $(k+1)$ explanatory variables. A subset with $R^2 > R_0^2$ is called an R^2-adequate(α) subset.

2. Adjusted coefficient of determination

The adjusted coefficient of determination has certain advantages over the usual coefficient of determination. The adjusted coefficient of determination based on *p*-term model is

$$R^2_{adj}(p) = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p).$$

An advantage of $R^2_{adj}(p)$ is that it does not necessarily increases as *p* increases.

If there are *r* more explanatory variables which are added to a *p* - term model then

$$R^2_{adj}(p+r) > R^2_{adj}(p)$$

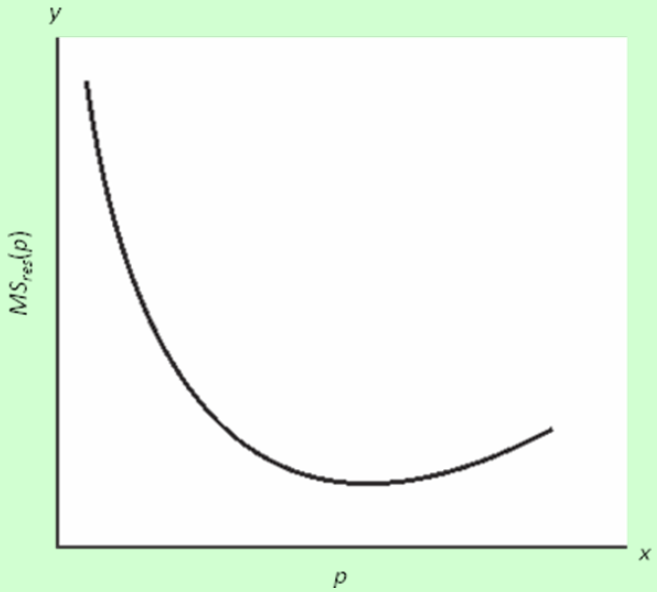
if and only if the partial *F* - statistic for testing the significance of *r* additional explanatory variables exceeds 1. So the subset selection based on $R^2_{adj}(p)$ can be made on the same lines are in R^2_p. In general, the value of *p* corresponding to maximum value of $R^2_{adj}(p)$ is chosen for the subset model.

3. Residual mean square

A model is said to have a better fit if residuals are small. This is reflected in the sum of squares due to residuals SS_{res}. A model with smaller SS_{res} is preferable. Based on this, the residual mean square based on a *p* variable subset regression model is defined as

$$MS_{res}(p) = \frac{SS_{res}(p)}{n-p}.$$

So $MS_{res}(p)$ can be used as a criterion for model selection like SS_{res}. The $SS_{res}(p)$ decreases with an increase in p. So similarly as p increases, $MS_{res}(p)$ initially decreases, then stabilizes and finally may increase if the model is not sufficient to compensate the loss of one degree of freedom in the factor $(n - p)$. When $MS_{res}(p)$ is plotted versus p, the curve look like as in the following figure.



So

- plot $MS_{res}(p)$ versus p.
- Choose p corresponding to minimum value of $MS_{res}(p)$.
- Choose p corresponding to which $MS_{res}(p)$ is approximately equal to MS_{res} based on full model.
- Choose p near the point where the smallest value of $MS_{res}(p)$ turns upward.

Such minimum value of $MS_{res}(p)$ will produce a $R^2_{adj}(p)$ with maximum value. So

$$R^2_{adj}(p) = 1 - \frac{n-1}{n-p}(1 - R^2_p)$$

$$= 1 - \frac{n-1}{n-p} . \frac{SS_{res}(p)}{SS_T}$$

$$= 1 - \frac{n-1}{SS_T} . \frac{SS_{res}(p)}{n-p}$$

$$= 1 - \frac{MS_{res}(p)}{SS_T / (n-1)} .$$

Thus the two criterion, viz, minimum $MS_{res}(p)$ and maximum $R^2_{adj}(p)$ are equivalent.

4. Mallow's C_p statistics

Mallow's C_p criterion is based on the mean squared error of a fitted value.

Consider the model $y = X\beta + \varepsilon$ with partitioned $X = (X_1, X_2)$ where X_1 is $n \times p$ matrix and X_2 is $n \times q$ matrix, so that

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon,\ E(\varepsilon) = 0,\ V(\varepsilon) = \sigma^2 I$$

where $\beta = (\beta_1', \beta_2')'$.

Consider the reduced model

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

and predict y based on subset model as

$$\hat{y} = X_1\hat{\beta}_1, \text{ where } \hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y.$$

The prediction of y can also be seen as the estimation of $E(y) = X\beta$, so the expected outweighed squared error loss of $\hat{y}$ is given by

$$\Gamma_p = E\left[\left(X_1\hat{\beta}_1 - X\beta\right)'\left(X_1\hat{\beta}_1 - X\beta\right)\right].$$

So the subset model can be considered as an appropriate model if Γ_p is small.

Since $H_1 = X_1(X_1'X_1)^{-1}X_1'$, so

$$\Gamma_p = E(y'H_1y) - 2\beta'X'H_1X\beta + \beta'X'X\beta.$$

where $E(y'H_1y) = E\left[(X\beta + \varepsilon)'H_1(X\beta + \varepsilon)\right]$

$$= E\left[\beta'X'H_1X\beta + \beta'X'H_1\varepsilon + \varepsilon'H_1X\beta + \varepsilon'H_1\varepsilon\right]$$

$$= \beta'X'H_1X\beta + 0 + 0 + \sigma^2\, tr\, H_1$$

$$= \beta'X'H_1X\beta + \sigma^2 p.$$

Thus

$$\Gamma_p = \sigma^2 p + \beta' X' H_1 X \beta_1 - 2\beta' X' H_1 X \beta + \beta' X' X \beta$$

$$= \sigma^2 p + \beta' X' X \beta - \beta' X' H_1 X \beta$$

$$= \sigma^2 p + \beta' X' (I - H_1) X \beta$$

$$= \sigma^2 p + \beta' X' \bar{H}_1 X \beta$$

where

$$\bar{H}_1 = I - X_1 (X_1' X_1)^{-1} X_1'.$$

Since

$$E(y' H_1 y) = E\left[(X\beta + \varepsilon)' H_1 (X\beta + \varepsilon)\right]$$

$$= \sigma^2 tr\bar{H}_1 + \beta' X' \bar{H}_1 X \beta$$

$$= \sigma^2 (n - p) + \beta' X' \bar{H}_1 X \beta$$

$$\Rightarrow \beta' X' \bar{H}_1 X \beta = E(y' \bar{H}_1 y) - \sigma^2 (n - p).$$

Thus

$$\Gamma_p = \sigma^2 (2p - n) + E(y' \bar{H}_1 y).$$

Note that Γ_p depends on β and σ^2 which are unknown. So Γ_p can not be used in practice. A solution to this problem is to replace β and σ^2 by their respective estimators which gives

$$\hat{\Gamma}_p = \hat{\sigma}^2 (2p - n) + SS_{res}(p)$$

where $SS_{res}(p) = y' H_1 y$ is the residuals sum of squares based on the subset model.

A rescaled vision of $\hat{\Gamma}_p$ is

$$C_p = (2p - n) + \frac{SS_{res}(p)}{\hat{\sigma}^2}$$

which is the Mallow's C_p statistic for the model $y = X_1\beta_1 + \delta,$ the subset model. Usually

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

$$\hat{\sigma}^2 = \frac{1}{n-p-q}(y - X\hat{\beta})'(y - X\hat{\beta})$$

are used to estimate β and σ^2 respectively which are based on full model.

When different subset models are considered, then the models with smallest C_p are considered to be better than those models with higher C_p. So lower C_p is preferable.

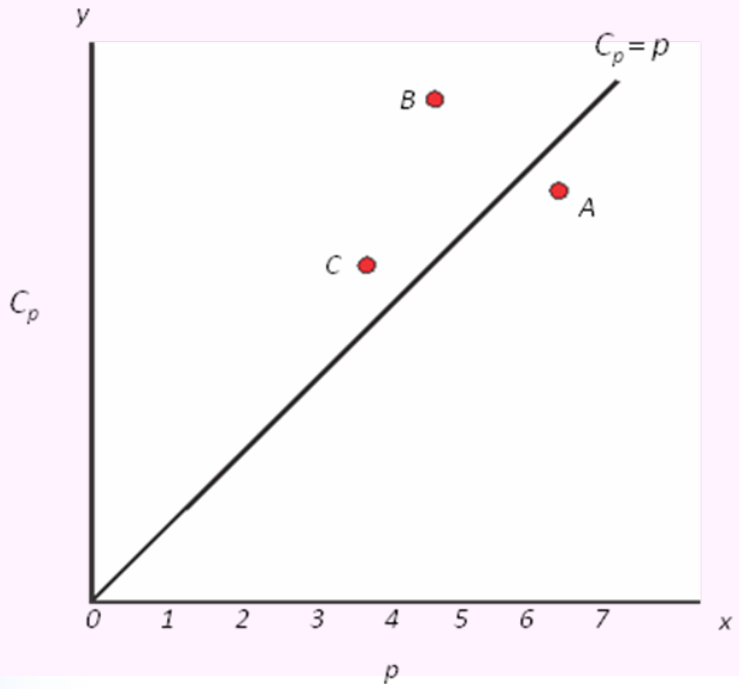
If the subset model has negligible bias, (in case of *b*, then bias is zero), then

$$E\left[SS_{res}(p)\right] = (n-p)\sigma^2$$

and

$$E\left[C_p \mid Bias = 0\right] = 2p - n - \frac{(n-p)\sigma^2}{\sigma^2} = p.$$

The plot of C_p versus p for each regression equation will be a straight line passing through origin and look like as follows:



Those points which have smaller bias will be near to line and those points with significant bias will lie above the line. For example, point *A* has little bias, so it is closer to line whereas points *B* and *C* have substantial bias, so they are above the line. Moreover, point *C* is above point *A* and it represents a model with lower total error. It may be preferred to accept some bias in the regression equation to reduce the average prediction error.

Note that an unbiased estimator of σ^2 is used in $C_p = p$ which is based on the assumption that the full model has negligible bias. In case, the full model contains non-significant explanatory variables with zero regression coefficients, then the same unbiased estimator of σ^2 will overestimate σ^2 and then C_p will have smaller values. So working of C_p depends on the good choice of estimator of σ^2.