Chapter 13
Variable Selection and Model Building

The complete regression analysis depends on the explanatory variables present in the model. It is understood in the regression analysis that only correct and important explanatory variables appear in the model. In practice, after ensuring the correct functional form of the model, the analyst usually has a pool of explanatory variables which possibly influence the process or experiment. Generally, all such candidate variables are not used in the regression modelling, but a subset of explanatory variables is chosen from this pool. How to determine such an appropriate subset of explanatory variables to be used in regression is called the problem of variable selection.

While choosing a subset of explanatory variables, there are two possible options:

1. In order to make the model as realistic as possible, the analyst may include as many as possible explanatory variables.
2. In order to make the model as simple as possible, one way includes only a fewer number of explanatory variables.

Both approaches have their consequences. In fact, model building and subset selection have contradicting objectives. When a large number of variables are included in the model, then these factors can influence the prediction of the study variable y. On the other hand, when a small number of variables are included then the predictive variance of $\hat{y}$ decreases. Also, when the observations on more number are to be collected, then it involves more cost, time, labour etc. A compromise between these consequences is striked to select the "best regression equation".

The problem of variable selection is addressed assuming that the functional form of the explanatory variable, e.g., x^2, $\frac{1}{x}, \log x$ etc., is known and no outliers or influential observations are present in the data. Various statistical tools like residual analysis, identification of influential or high leverage observations, model adequacy etc. are linked to variable selection. In fact, all these processes should be solved simultaneously. Usually, these steps are iteratively employed. In the first step, a strategy for variable selection is opted, and the model is fitted with selected variables. The fitted model is then checked for the functional form, outliers,

influential observations etc. Based on the outcome, the model is re-examined, and the selection of variable is reviewed again. Several iterations may be required before the final adequate model is decided.

There can be two types of incorrect model specifications.

1. Omission/exclusion of relevant variables.
2. Inclusion of irrelevant variables.

Now we discuss the statistical consequences arising from both situations.

1. Exclusion of relevant variables:

In order to keep the model simple, the analyst may delete some of the explanatory variables which may be of importance from the point of view of theoretical considerations. There can be several reasons behind such decision, e.g., it may be hard to quantify the variables like the taste, intelligence etc. Sometimes it may be difficult to take correct observations on the variables like income etc.

Let there be k candidate explanatory variables out of which suppose r variables are included and $(k-r)$ variables are to be deleted from the model. So partition the X and β as

$$\underset{n\times k}{X} = \begin{pmatrix} \underset{n\times r}{X_1} & \underset{n\times(k-r)}{X_2} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \underset{r\times 1}{\beta_1} \\ \underset{(k-r)\times 1}{\beta_2} \end{pmatrix}.$$

The model $y = X\beta + \varepsilon,\ E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I$ can be expressed as

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

which is called a **full model** or **true model.**

After dropping the r explanatory variable in the model, the new model is

$$y = X_1\beta_1 + \delta$$

which is called a **misspecified mode**l or **false model.**

Applying OLS to the false model, the OLSE of β_1 is

$$b_{1F} = (X_1'X_1)^{-1}X_1'y.$$

The estimation error is obtained as follows:

$$\begin{aligned} b_{1F} &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon \\ b_{1F} - \beta_1 &= \theta + (X_1'X_1)^{-1}X_1'\varepsilon \end{aligned}$$

where $\theta = (X_1'X_1)^{-1}X_1'X_2\beta_2$.

Thus

$$\begin{aligned} E(b_{1F} - \beta) &= \theta + (X_1'X_1)^{-1}E(\varepsilon) \\ &= \theta \end{aligned}$$

which is a linear function of β_2, i.e., the coefficients of excluded variables. So b_{1F} is biased, in general. The bias vanishes if $X_1'X_2 = 0$, i.e., X_1 and X_2 are orthogonal or uncorrelated.

The mean squared error matrix of b_{1F} is

$$\begin{aligned} MSE(b_{1F}) &= E(b_{1F} - \beta)(b_{1F} - \beta)' \\ &= E\left[\theta\theta' + \theta\varepsilon' X_1(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'\varepsilon\theta' + (X_1'X_1)^{-1}X_1'\varepsilon\varepsilon' X_1(X_1'X_1)^{-1}\right] \\ &= \theta\theta' + 0 + 0 + \sigma^2(X_1'X_1)^{-1}X_1'IX_1(X_1'X_1)^{-1} \\ &= \theta\theta' + \sigma^2(X_1'X_1)^{-1}. \end{aligned}$$

So efficiency generally declines. Note that the second term is the conventional form of MSE.

The residual sum of squares is

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-r} = \frac{e'e}{n-r}$$

where $e = y - X_1 b_{1F} = \bar{H}_1 y,$

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

Thus

$$\begin{aligned} \bar{H}_1 y &= \bar{H}_1(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= 0 + \bar{H}_1(X_2\beta_2 + \varepsilon) \\ &= \bar{H}_1(X_2\beta_2 + \varepsilon). \end{aligned}$$

$$\begin{aligned} y'\bar{H}_1 y &= (X_1\beta_1 + X_2\beta_2 + \varepsilon)\bar{H}_1(X_2\beta_2 + \varepsilon) \\ &= (\beta_2'X_2\bar{H}_1'\bar{H}_1X_2\beta_2 + \beta_2'X_2'\bar{H}_1\varepsilon + \beta_2'X_2\bar{H}_1'X_2\beta_2 + \beta_1'X_1'\bar{H}_1\varepsilon + \varepsilon'\bar{H}_1'X_2\beta_2 + \varepsilon'\bar{H}_1\varepsilon). \end{aligned}$$

$$
\begin{aligned}
E(s^2) &= \frac{1}{n-r}\left[E(\beta_2' X_2' \bar{H}_1 X_2 \beta_2) + 0 + 0 + E(\varepsilon' \bar{H} \varepsilon)\right] \\
&= \frac{1}{n-r}\left[\beta_2' X_2' \bar{H}_1 X_2 \beta_2) + (n-r)\sigma^2\right] \\
&= \sigma^2 + \frac{1}{n-r}\beta_2' X_2' \bar{H}_1 X_2 \beta_2.
\end{aligned}
$$

Thus s^2 is a biased estimator of σ^2 and s^2 provides an overestimate of σ^2. Note that even if $X_1' X_2 = 0,$ then also s^2 gives an overestimate of σ^2. So the statistical inferences based on this will be faulty. The t-test and confidence region will be invalid in this case.

If the response is to be predicted at $x' = (x_1', x_2'),$ then using the full model, the predicted value is

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

with

$$
\begin{aligned}
E(\hat{y}) &= x'\beta \\
Var(\hat{y}) &= \sigma^2\left[1 + x'(X'X)^{-1}x\right].
\end{aligned}
$$

When the subset model is used then the predictor is

$$\hat{y}_1 = x_1' b_{1F}$$

and then

$$
\begin{aligned}
E(\hat{y}_1) &= x_1'(X_1'X_1)^{-1}X_1'E(y) \\
&= x_1'(X_1'X_1)^{-1}X_1'E(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\
&= x_1'(X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2) \\
&= x_1'\beta_1 + x_1'(X_1'X_1)^{-1}X_1'X_2\beta_2 \\
&= x_1'\beta_1 + x_i'\theta.
\end{aligned}
$$

Thus $\hat{y}_1$ is a biased predictor of y. It is unbiased when $X_1' X_2 = 0.$ The MSE of predictor is

$$MSE(\hat{y}_1) = \sigma^2\left[1 + x_1'(X_1'X_1)^{-1}x_1\right] + \left(x_1'\theta - x_2'\beta_2\right)^2.$$

Also

$$Var(\hat{y}) \geq MSE(\hat{y}_1)$$

provided $V(\hat{\beta}_2) - \beta_2\beta_2'$ is positive semidefinite.

2. Inclusion of irrelevant variables

Sometimes due to enthusiasm and to make the model more realistic, the analyst may include some explanatory variables that are not very relevant to the model. Such variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom $(n-k)$ and consequently, the validity of inference drawn may be questionable. For example, the value of the coefficient of determination will increase, indicating that the model is getting better, which may not really be true.

Let the true model be

$$y = X\beta + \varepsilon,\ E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I$$

which comprise k explanatory variable. Suppose now r additional explanatory variables are added to the model and the resulting model becomes

$$y = X\beta + Z\gamma + \delta$$

where Z is a $n \times r$ matrix of n observations on each of the r explanatory variables and γ is $r \times 1$ vector of regression coefficient associated with Z and δ is disturbance term. This model is termed as a **false model.**

Applying OLS to false model, we get

$$\begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix} \begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\Rightarrow X'Xb_F + X'ZC_F = X'y \qquad (1)$$

$$Z'Xb_F + Z'ZC_F = Z'y \qquad (2)$$

where b_F and C_F are the OLSEs of β and γ respectively.

Premultiply equation (2) by $X'Z(Z'Z)^{-1}$, we get

$$X'Z(Z'Z)^{-1}Z'Xb_F + X'Z(Z'Z)^{-1}Z'ZC_F = X'Z(Z'Z)^{-1}Z'y. \qquad (3)$$

Subtracting equation (1) from (3), we get

$$\left[X'X - X'Z(Z'Z)^{-1}Z'X\right]b_F = X'y - X'Z(Z'Z)^{-1}Z'y$$

$$X'\left[I - Z(Z'Z)^{-1}Z'\right]Xb_F = X'\left[I - Z(Z'Z)^{-1}Z'\right]y$$

$$\Rightarrow b_F = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y$$

where $\bar{H}_Z = I - Z(Z'Z)^{-1}Z'$.

The estimation error of b_F is

$$\begin{aligned} b_F - \beta &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z(X\beta + \varepsilon) - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z\varepsilon. \end{aligned}$$

Thus

$$E(b_F - \beta) = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z E(\varepsilon) = 0$$

so b_F is unbiased even when some irrelevant variables are added to the model.

The covariance matrix is

$$\begin{aligned} V(b_F) &= E\left(b_F - \beta\right)\left(b_F - \beta\right)^1 \\ &= E\left[\left(X'\bar{H}_Z X\right)^{-1}X'\bar{H}_Z\varepsilon\varepsilon'\bar{H}_Z X(X'\bar{H}_Z X)^{-1}\right] \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1}X'\bar{H}_Z I\bar{H}_Z X\left(X'\bar{H}_Z X\right)^{-1} \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1}. \end{aligned}$$

If OLS is applied to true model, then

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

with $E(b_T) = \beta$

$$V(b_T) = \sigma^2(X'X)^{-1}.$$

To compare b_F and b_T, we use the following result.

Result: If A and B are two positive definite matrices then $A - B$ is at least positive semi-definite if $B^{-1} - A^{-1}$ is also at least positive semi-definite.

Let

$$A = (X'\bar{H}_Z X)^{-1}$$
$$B = (X'X)^{-1}$$
$$\begin{aligned} B^{-1} - A^{-1} &= X'X - X'\bar{H}_Z X \\ &= X'X - X'X + X'Z(Z'Z)^{-1}Z'X \\ &= X'Z(Z'Z)^{-1}Z'X \end{aligned}$$

which is at least positive semidefinite matrix. This implies that the efficiency declines unless $X'Z = 0.$ If $X'Z = 0,$ i.e., X and Z are orthogonal, then both are equally efficient.

The residual sum of squares under the false model is

$$SS_{res} = e_F' e_F$$

where

$$\begin{aligned} e_F &= y - Xb_F - ZC_F \\ b_F &= (X\bar{H}_Z X)^{-1} X'\bar{H}_Z y \\ C_F &= (Z'Z)^{-1}Z'y - (Z'Z)^{-1}Z'Xb_F \\ &= (Z'Z)^{-1}Z'(y - Xb_F) \\ &= (Z'Z)^{-1}Z'\left[I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_z\right]y \\ &= (Z'Z)^{-1}Z'\bar{H}_{XZ}y \\ \bar{H}_Z &= I - Z(Z'Z)^{-1}Z' \\ \bar{H}_{Zx} &= I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_Z \\ \bar{H}_{ZX}^2 &= \bar{H}_{ZX} : \text{ idempotent.} \end{aligned}$$

So

$$\begin{aligned} e_F &= y - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y - Z(Z'Z)^{-1}Z'\bar{H}_{ZX}y \\ &= \left[I - X(X'\bar{H}_Z X)^{-1}X'\bar{H}_Z - Z(Z'Z)^{-1}Z'H_{ZX}\right]y \\ &= \left[\bar{H}_{ZX} - (I - \bar{H}_Z)\bar{H}_{ZX}\right]y \\ &= \bar{H}_Z\bar{H}_{ZX}y \\ &= \bar{H}_{ZX}^* y \quad \text{where } \bar{H}_{ZX}^* = \bar{H}_Z\bar{H}_{ZX}. \end{aligned}$$

Thus

$$\begin{aligned} SS_{res} &= e_F' e_F \\ &= y'\bar{H}_Z\bar{H}_{ZX}\bar{H}_{ZX}\bar{H}_Z y \\ &= y'\bar{H}_Z\bar{H}_{ZX}y \\ &= y'\bar{H}_{ZX}^* y \end{aligned}$$

$$E(SS_{res}) = \sigma^2 tr(\bar{H}_{ZX}^*)$$
$$= \sigma^2 (n-k-r)$$
$$E\left(\frac{SS_{res}}{n-k-r}\right) = \sigma^2.$$

So $\frac{SS_{res}}{n-k-r}$ is an unbiased estimator of σ^2.

A comparison of exclusion and inclusion of variables is as follows:

	Exclusion type	**Inclusion type**
Estimation of coefficients	Biased	Unbiased
Efficiency	Generally declines	Declines
Estimation of the disturbance term	Over-estimate	Unbiased
Conventional test of hypothesis and confidence region	Invalid and faulty inferences	Valid though erroneous

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

$$R_p^2 = \frac{SS_{reg}(p)}{SS_T}$$
$$= 1 - \frac{SS_{res}(p)}{SS_T}$$

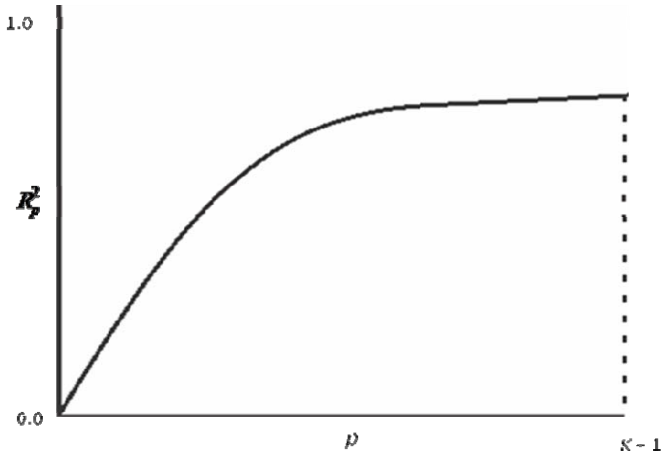
where $SS_{reg}(p)$ and $SS_{res}(p)$ are the sum of squares due to regression and residuals, respectively in a subset model based on $(p-1)$ explanatory variables.

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 an 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 up to a point where an additional variable does not produce 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 p may not be unique among different analyst. Some experience and judgment of analyst will be helpful in 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 $d_{\alpha,n,k} = \dfrac{kF_\alpha(n,n-k-1)}{n-k-1}$ 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 **$\mathbf{R^2}$ - 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

$$R_{adj}^2(p) = 1-\left(\frac{n-1}{n-p}\right)(1-R_p^2).$$

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

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

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

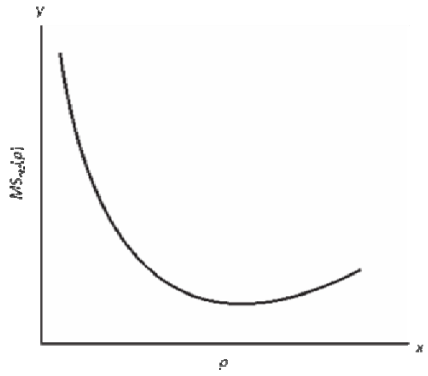
if and only if the partial $F-$statistic for testing the significance of r additional explanatory variables exceeds 1. So the subset selection based on $R_{adj}^2(p)$ can be made on the same lines are in R_p^2. In general, the value of p corresponding to the maximum value of $R_{adj}^2(p)$ is chosen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



So

- plot $MS_{res}(p)$ versus p.
- Choose p corresponding to the minimum value of $MS_{res}(p)$.
- Choose p corresponding to which $MS_{res}(p)$ is approximately equal to MS_{res} based on the 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

$$\begin{aligned} R^2_{adj}(p) &= 1-\frac{n-1}{n-p}(1-R^2_p) \\ &= 1-\frac{n-1}{n-p}.\frac{SS_{res}(p)}{SS_T} \\ &= 1-\frac{n-1}{SS_T}.\frac{SS_{res}(p)}{n-p} \\ &= 1-\frac{MS_{res}(p)}{SS_T/(n-1)}. \end{aligned}$$

Thus the two criteria, 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 the 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 $E(y'H_1y) = E\left[(X\beta+\varepsilon)'H_1(X\beta+\varepsilon)\right]$

$$\begin{aligned}
&= E\left[\beta'X'H_1X\beta + \beta'X'H_1\varepsilon + \varepsilon'H_1X\beta + \varepsilon'H_1\varepsilon\right] \\
&= \beta'X'H_1X\beta + 0 + 0 + \sigma^2\, tr\, H_1 \\
&= \beta'X'H_1X\beta + \sigma^2 p.
\end{aligned}$$

Thus

$$\begin{aligned}
\Gamma_p &= \sigma^2 p + \beta'X'H_1X\beta_1 - 2\beta'X'H_1X\beta + \beta'X'X\beta \\
&= \sigma^2 p + \beta'X'X\beta - \beta'X'H_1X\beta \\
&= \sigma^2 p + \beta'X'(I - H_1)X\beta \\
&= \sigma^2 p + \beta'X'\bar{H}_1X\beta
\end{aligned}$$

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

Since

$$\begin{aligned}
E(y'H_1y) &= E\left[(X\beta+\varepsilon)'H_1(X\beta+\varepsilon)\right] \\
&= \sigma^2 tr\bar{H}_1 + \beta'X'\bar{H}_1X\beta \\
&= \sigma^2(n-p) + \beta'X'\bar{H}_1X\beta \\
\Rightarrow \beta'X'\bar{H}_1X\beta &= E(y'\bar{H}_1y) - \sigma^2(n-p)
\end{aligned}$$

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 the 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.

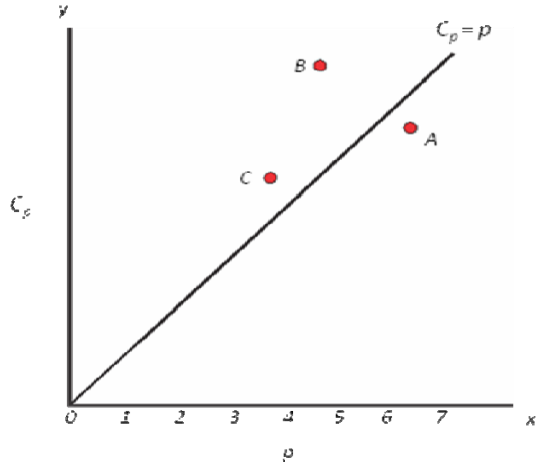
If the subset model has a negligible bias, (in case of b, then bias is zero), then

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

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 the 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, the point A has little bias, so it is closer to line A whereas points B and C have a substantial bias, so they are above the line. Moreover, the point C is above point A, and it represents a model with a 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 a 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 the estimator of σ^2.

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 $SS_{res}(p) = y' H_1 y = y' X_1 (X_1' X_1)^{-1} X_1' y$

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

The *AIC* is defined as

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

In the 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 of $L(y, \beta, \sigma^2)$. 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 the same for all the models under comparison if the same observations y are compared. So it is irrelevant for AIC.

6. Bayesian information criterion *(BIC)*

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

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

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

7. PRESS statistic

Since the residuals and residual sum of squares act as a criterion of 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 use 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 similar lines as in the case of $SS_{res}(p)$. A subset regression model with a 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

$$\underset{n\times 1}{y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

where $p = k+1$ which includes an intercept term and k explanatory variables. Suppose a subset of $r < k$ 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 $n\times(p-r)$ and $n\times r$, respectively; β_1 and β_2 are the vectors of order $(p-1)\times 1$ 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 p degrees of freedom is

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

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

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

and $$MS_{res} = \frac{y'y - b'X'y}{n-p}$$

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 the corresponding sum of squares due to regression with $(p-r)$ 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 $SS_{reg}(\beta_1)$ are the sum of squares due to regression with all explanatory variables corresponding to β is the model and the 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. $p-(p-r) = r.$ 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 $(n-p)$ degrees of freedom under H_0. The decision rule is to reject H_0 whenever

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

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 the 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 the 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 an increase in k. So such models can be evaluated using a 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 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 the 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 tests the fitted model at each step using some suitable criterion. It has the 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 the largest correlation with y.
- Suppose x_1 is the variable which has the 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 a regression.
- Choose a prespecified value of F value, say $F_{IN}(F-\text{to}-\text{enter})$.
- If $F > F_{IN}$, then accept x_1 and so x_1 enters into the model.
- Adjust the effect of x_1 on y 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 the second-largest correlation with y, i.e., the variable with the 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 the highest partial correlation with y is chosen to be added into the model. Equivalently, the partial F-statistics are calculated, and the 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 a particular step, the partial $F-$statistic does not exceed F_{IN} or when the least explanatory variable is added to the model.

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

Backward elimination procedure:

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

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

It is based on the 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_{0UT} ($F-$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 the smallest partial $F-$statistic exceeds F_{OUT}.

Stepwise regression procedure:

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

- Consider all the explanatory variables entered into the model at the 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 the 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 the 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 the analyst to specify these values directly.
- Some algorithms require type I errors to generate F_{IN} or/and F_{OUT}. Sometimes, taking α as the 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. A popular choice is $F_{IN} = F_{OUT} = 4$ which is corresponding to 5% level of significance of $F-$distribution.