Chapter 7
Variable Selection and Model Building

The complete regression analysis depends on the explanatory variables present in the model. In regression analysis, only relevant explanatory variables are included in the model. In practice, after ensuring the model's correct functional form, the analyst usually has a pool of explanatory variables that may influence the process or experiment. Generally, not all such candidate variables are used in the regression modelling, but a subset of explanatory variables is chosen from this pool. Determining an appropriate subset of explanatory variables for regression is called variable selection.

Evaluation of the subset regression model:

A question arises after selecting subsets of candidate variables for the model: how to judge which subset yields the 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 the multiple correlation coefficient between the study variable y and a 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 for an 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

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

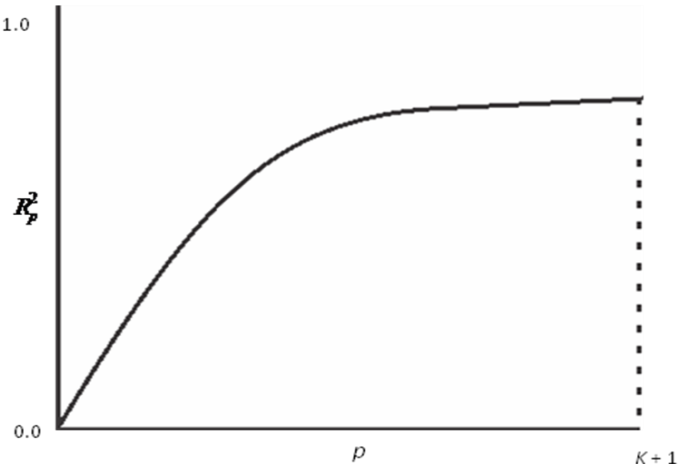
where $SS_{reg}(p)$ and $SS_{res}(p)$ are the sum of squares due to regression and residuals, respectively in a subset model based on $(p-1)$ explanatory variables.

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 a value of p, create a plot of R_p^2 versus p. For example, the curve will look 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 a choice of p may not be unique among different analysts. Some experience and judgment of the 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 **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

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

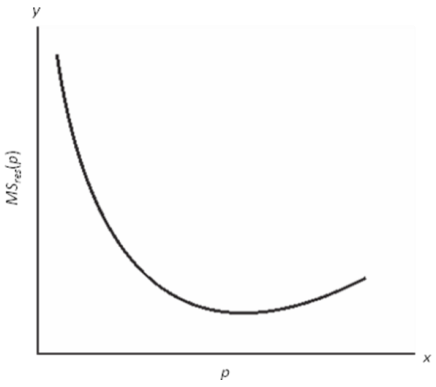
if and only if the partial $F-$statistic for testing the significance of r additional explanatory variables exceeds 1. So the subset selection based on $R_{adj}^2(p)$ can be made on the same lines as in R_p^2. In general, the value of p corresponding to the maximum value of $R_{adj}^2(p)$ is chosen for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

3. Residual mean square:

A model is said to have a better fit if the 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. Similarly, as p increases, $MS_{res}(p)$ initially decreases, then stabilizes and finally may increase if the model is not sufficient to compensate for the loss of one degree of freedom in the factor $(n-p)$. When $MS_{res}(p)$ is plotted versus p, the curve looks 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 a 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

$$\Gamma_p = \sigma^2(2p-n) + E(y'\bar{H}_1y).$$

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 $SS_{res}(p) = y'H_1y$ is the residual 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, the models with the smallest C_p are considered to be better than those models with higher C_p. So lower C_p is preferable.

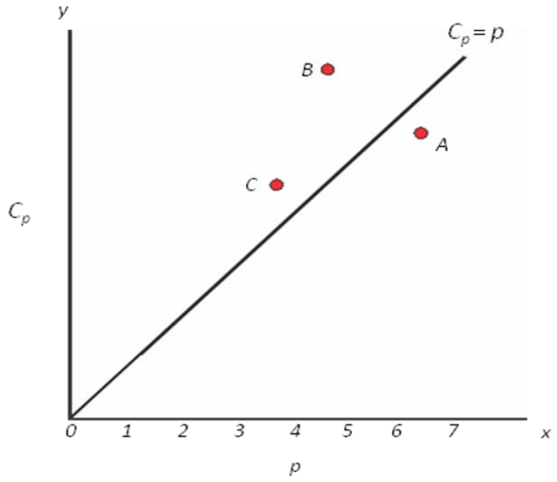
If the subset model has a negligible bias (in case of b, then bias is zero), then

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

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 will look as follows:



Points with smaller bias will be closer to the line, and those with significant bias will lie above it. For example, the point A has little bias, so it is closer to the 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 on C_p depends on the good choice of the estimator of σ^2.

5. Akaike's information criterion (AIC):

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

So

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

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 a linear regression model, it is defined as

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

A model with a lower *BIC* value *i*s preferable.

7. PRESS statistic:

Since the residuals and residual sum of squares serve as criteria for subset model selection, the PRESS residuals and prediction sum of squares can also be used as bases for subset model selection. The usual residuals and PRESS residuals have their own characteristics used in regression modelling.

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

$$\begin{aligned} PRESS(p) &= \sum_{i=1}^{n}\left[y_i - \hat{y}_{(i)}\right]^2 \\ &= \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2. \end{aligned}$$

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 are to be obtained which contribute significantly to the regression model. So partition

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

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 are

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

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

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

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 the 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 is 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 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: either select all possible explanatory variables, or select them 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 equations 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 via an efficient computational algorithm on a computer.

2. Stepwise regression techniques:

This methodology selects explanatory variables in the subset model in steps, either by adding one variable at a time or by deleting one at a time. Based on this, there are three procedures.

- Forward selection,
- backward elimination and
- stepwise regression.

These procedures are primarily computer-intensive and executed using the software.

Forward selection procedure:

This methodology assumes that the model contains no explanatory variables other than 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 the 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 recompute the correlations of the 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 the explanatory variable corresponding to the highest partial correlation with y is chosen to be added to 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 to the model if partial F-statistic exceeds F_{IN}.
- Continue with such a selection as long as, 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 variables in the model and adds 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 variable 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 than 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 is the stepwise regression procedure. It is a modification of the forward selection procedure and follows these steps.

- Consider all the explanatory variables entered into the model at the previous step.
- Add a new variable and regress 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 it 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 variables.
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 it is removed from the model.
4. All procedures may lead to different models.
5. Different model selection criteria 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 the maximum among them is examined.
- Some analysts prefer small values of F_{IN} and F_{OUT} whereas some prefer extreme values. A popular choice is $F_{IN} = F_{OUT} = 4$ which corresponds to 5% level of significance of $F-$distribution.