CHAPTER V

EFFECT OF DROPPING VARIABLES ON REGRESSION COEFFICIENTS

CHAPTER - V

EFFECT OF DROPPING VARIABLES ON ESTIMATION OF REGRESSION COEFFICIENTS

5.1 Introduction :

In the Chapter II, III and IV we have presented various criteria and sequential procedures for subset selection in regression analysis. We have also discussed the advantages of subset selection methods. The main advantage is that these methods help to identify relevant variables to be included in the model. On other hand, as pointed out earlier if variables are dropped from the model then estimator of regression coefficients becomes biased and this is the main drawback of any subset selection method. For this reason, in this chapter we discuss about the effect of dropping variables and bias.

In section (5.2) we discuss the two types of bias namely 1) Omission bias 2) Selection bias. The effect of dropping variables on estimation of $\underline{\beta}$ and estimation of error variance which are discussed in section (5.3). Now here question arises as how many variables should be included in the model. The answer to this question is given by Groman and Thoman (1966), which is discussed in section (5.4). Since the estimation of bias can not be devied and also it cannot be removed totally, a question is whether it can be reduced siginificantly. We address this question in the last section. Wherein we discuss the two methods of bias reduction with illustration namely 1)Jackknife statistic and 2) Bootstrap method.

5.2 Bias :

In this section, we discuss various types of bias that arise due to subset selection.

We consider the full model

$$\underline{Y} = X \underline{\beta} + \underline{\varepsilon} \tag{5.2.1}$$

We partition the matrix X and vector β as follows,

$$X = (X_1 : X_2) \quad \text{and} \quad \underline{\beta} = (\underline{\beta}_1 : \underline{\beta}_2)',$$

where X is $n \times k+1$ matrix, X_1 is $n \times p$ matrix and β is $k+1\times1$ vector and β_1 is the $p\times1$ vector. Then the least square estimate of $\underline{\beta}_1$ is given by

$$\underline{b}_1 = (X_1' X_1)^{-1} X_1' \underline{Y},$$

Then

$$E(\underline{b}_1) = \underline{\beta}_1 + (X_1' X_1)^{-1} X_1' X_2 \underline{\beta}_2, \tag{5.2.2}$$

where $\underline{\beta}_1$ and $\underline{\beta}_2$ consist of the first p+1 and last k-p elements of $\underline{\beta}$. The second term of (5.2.2) is clearly the bias in the first (p+1) regression coefficients arising from the omission of the last (k-p) variables. Thus, we have the following definition

Definition (5.2.1) Omission bias : The difference between $E(\underline{b}_1)$ and $\underline{\beta}_1$ is called as the 'omission bias' which arises when (k-p) variables are deleted.

i.e.

$$\text{Omission bias} = E(\underline{b}_1) - \underline{\beta}_1,$$
$$= (X_1' X_1)^{-1} X_1' X_2 \underline{\beta}_2, \qquad (5.2.3)$$

Further if the subset of variables is chosen from the same data which are used to estimate the regression coefficient and the prediction, it introduce the anthor type of bias which is called the 'selection bias'. It is formally defined as:

Definition (5.2.2): Selection bias : The difference between the expected values of regression coefficients when the data are such as to satisfy the condition necessary for the selection of the subset X_1 and unconditional expected value of $\underline{b}_1$ is called the 'Selection bias'.i.e.

$$\text{Selection bias} = E(\underline{b}_1/\text{subset first selected}) - E(\underline{b}_1) \qquad (5.2.4)$$

The first term of the above expression is the expected value of $\underline{b}_1$ over all possible Y vectors which would lead to subset first being selected. The second term is the expected value over all Y irrespective of what subset is selected.

5.3 Some effects of dropping variables:

5.3.1 Effect on estimation of $\underline{\beta}$:

If some important variables are dropped from the true model then the estimator of the parameter becomes biased estimator of $\underline{\beta}$. Let us consider the true model as in (5.2.1) and the partition similar to section (5.2), then

$$\underline{Y} = X_1\underline{\beta}_1 + X_2\underline{\beta}_2 + \underline{\varepsilon} \tag{5.3.1}$$

From model (4.3.1), suppose X_2 is dropped out then we have

$$\underline{Y} = X_1\underline{\beta}_1 + \underline{\varepsilon} \tag{5.3.2}$$

The least square estimate of $\underline{\beta}_1$ is given by,

$$\underline{b}_1 = (X_1'X_1)^{-1}X_1'\ \underline{Y} \tag{5.3.3}$$

$$\begin{aligned}
E(\underline{b}) &= E(X_1'X_1')^{-1}X_1'\ \underline{Y}),\\
&= (X_1'X_1')^{-1}X_1'\ X\ \underline{\beta},\\
&= (X_1'X_1')^{-1}X_1'\ (X_1 : X_2)\begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix},\\
&= (X_1'X_1')^{-1}X_1'\ (X_1\underline{\beta}_1 + X_2\underline{\beta}_2),\\
&= (X_1'X_1')^{-1}X_1'\ X_1\underline{\beta}_1 + (X_1'X_1')^{-1}X_1'\ X_2\underline{\beta}_2,\\
&= \underline{\beta}_1 + (X_1'X_1')^{-1}X_1'\ X_2\underline{\beta}_2,
\end{aligned}$$

Thus, the least square estimate of $\underline{\beta}_1$ is obtained by least squares after deleting $X_2\underline{\beta}_2$ is biased by the amount $(X_1'X_1')^{-1}X_1'X_2\underline{\beta}_2$. Below, we give a condition for bias to be zero.

Suppose X_2 is orthogonal to X_1 such that $X_1'X_2 = 0$

Now

$$X'X = (\ X_1 : X_2)'(X_1 : X_2)$$

$$= \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}$$

$$= \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{bmatrix}$$

$$(X'X)^{-1}X'\underline{Y} = \begin{bmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{bmatrix} \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \underline{Y},$$

$$= \begin{bmatrix} (X_1'X_1)^{-1}X_1' \\ (X_2'\ X_2)^{-1}X_2' \end{bmatrix} \underline{Y}$$

We know that ,

$$\underline{b} = (X'X)^{-1}X'\underline{Y},$$

$$= \begin{bmatrix} (X_1'X_1)^{-1}X_1' \\ (X_2'\ X_2)^{-1}X_2' \end{bmatrix} \underline{Y},$$

$$= \begin{bmatrix} \underline{b}_1 \\ \underline{b}_2 \end{bmatrix}$$

Hence

$$\underline{b}_1 = (X_1'X_1)^{-1}X_1'\underline{Y},$$

$$\underline{b}_2 = (X_2'X_2)^{-1}X_2'\underline{Y},$$

and since $X'\underline{Y} = (X_1'\underline{Y} : X_2'\underline{Y})$.

It shows that $\underline{b}_1$ is actually same as first p components of $\underline{b}$. Thus, if X_1 and X_2 are orthogonal, dropping X_2 does not have any effect on $\underline{b}_1$. Of course, such a condition is quite unrealastic.

5.3.2 Effect on estimation of error variance

In this subsection, we discuss the effect of the dropping variables on estimation of variance of error. If we exclude the variables from the model, the estimate of variance of error increases. From the discussion in Chapter (I), we have the estimator of σ^2 based on full model (5.2.1) which is given by

$$\hat{\sigma}^2 = RSS_{k+1} / (n-k-1),$$

$$= (n-k-1)^{-1}\underline{Y}'(I - H)\underline{Y},$$

where $H = X(X'X)^{-1}X'$.

Suppose we delete X_2 in the full model. Then an estimate of σ^2 based on p independent variables is given by

$$\hat{\sigma}_p^2 = RSS_p / (n-p-1),$$

$$= (n-p-1)^{-1}(\underline{Y} - \hat{\underline{Y}}_p)'(\underline{Y} - \hat{\underline{Y}}_p),$$

$$= (n-p-1)^{-1}\left[\underline{Y}'\underline{Y} - 2\underline{Y}'\hat{\underline{Y}}_p + \hat{\underline{Y}}_p'\hat{\underline{Y}}_p\right],$$

$$= (n-p-1)^{-1}\left[\ \underline{Y}'\underline{Y}\ -2\underline{Y}'X_1 b_{1-1} + b_{-1}'X_1'X_1 b_{1-1}\right],$$

$$= (n-p-1)^{-1}\left[\ \underline{Y}'\underline{Y}\ -2\underline{Y}'X_1'(X_1'X_1)^{-1}X_1'\underline{Y}+\underline{Y}'X_1(X_1'X_1)^{-1}X_1'X_1(X_1'X_1)^{-1}X_1'\underline{Y}\ \right],$$

$$= (n-p-1)^{-1}\left[\ \underline{Y}'\underline{Y}\ -\underline{Y}'X_1(X_1'X_1)^{-1}X_1'\underline{Y}\ \right],$$

$$= (n-p-1)^{-1}\underline{Y}'\left[\ I-H_1\right]\ \underline{Y}.$$

where $H_1 = X_1(X_1'X_1)^{-1}X_1'$ and $\hat{\underline{Y}}_p = X_1 b_{1-1} = H_1\underline{Y}$.

The $\hat{\underline{Y}}_p$ is the predicted value of $\underline{Y}$ based on first P independent variables. Here note that

$$tr(I-H_1) = n-p-1.$$

Now,

$$E(\hat{\sigma}_p^2) = E\left[(n-p-1)^{-1}\underline{Y}'(\ I-H_1)\ \underline{Y}\right],$$

$$= (n-p-1)^{-1}E\left[\underline{Y}'(I-H_1)\ \underline{Y}\right],$$

$$= (n-p-1)^{-1}E\left[tr(\ I-H_1)\underline{Y}\ \underline{Y}'\right],$$

$$= (n-p-1)^{-1}tr\left[(\ I-H_1)(\sigma^2 I + (X\underline{\beta})(X\underline{\beta})')\right],$$

$$= (n-p-1)^{-1}\left[(\ n-p-1)\sigma^2 + tr(I-H_1)(X\underline{\beta})(X\underline{\beta})'\right],$$

$$= (n-p-1)^{-1}\left[(n-p-1)\sigma^2 + \underline{\beta}'X'(I-H_1)X\underline{\beta}\right].$$

Hence,

$$E(\hat{\sigma}^2_p) = \sigma^2 + (n-p-1)^{-1}\underline{\beta}'X'(I-H_1)X\underline{\beta} \qquad (5.3.4)$$

and $$E(\hat{\sigma}^2) = \sigma^2 .$$

It follows that,

$$E(\hat{\sigma}^2_p-\hat{\sigma}^2) = (n-p-1)^{-1}\underline{\beta}'X'(I-H_1)X\underline{\beta}.$$

We know that $(I - H_1)$ is idempotent matrix. Therefore

$$X'(I-H_1)X' \geq 0 \quad \text{for all } X .$$

Hence,

$$(n-p-1)^{-1}\underline{\beta}'X'(I-H_1)X\underline{\beta} \geq 0$$

From above inquality,

$$E(\hat{\sigma}^2_p-\hat{\sigma}^2) \geq 0,$$

$$E(\hat{\sigma}^2_p)-E(\hat{\sigma}^2) \geq 0,$$

$$E(\hat{\sigma}^2_p) \geq E(\hat{\sigma}^2) \qquad (5.3.5)$$

The experssion (5.3.5) shows that $\hat{\sigma}^2$ is biased estimator of σ^2 and equation (5.3.5) shows that $E(\hat{\sigma}^2_p)$ increases when variables are deleted from the model.

5.4 How many variables in prediction equation ?

If we are planning to use a linear regression equation to predict a future observation, we are faced with a problem of selecting an adequate set of independent variables to include in the equation. Obviously the set of variables should be selected such that it minimizes the variance of predicted value $\hat{Y}$ and bias of regression coefficents. This idea is used by Groman and Thoman (1966). In this section we discuss about the variance of prediction.

As pointed out earlier, if some variables are dropped from the model, then variance of $\hat{Y}$ decreases. In this situation a question arises as to how many variables should be used for prediction. Below, we attempt to answer this question. As usual, let

$$Y = \beta_o + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \qquad (5.4.1)$$

The least square estimator of the regression coefficent is,

$$\underline{b} = (X'X)^{-1}X'\underline{Y}$$

Now the predicted vector $\hat{\underline{Y}}$ for a given vector $\underline{x}$ where

$$\underline{x}' = (1, x_1, x_2, \ldots x_k)$$

is given by

$$\hat{Y} = \underline{x}' \ \underline{b},$$

Further,

$$V(\hat{\underline{Y}}) = V(\underline{x}' \ \underline{b}),$$

$$= \underline{x}'(X'X)^{-1}X'V(\underline{Y})X(X'X)^{-1}\underline{x},$$

$$= \sigma^2 \ \underline{x}'(X'X)^{-1}\underline{x}.$$

Now, we find an upper triangular matrix R of order (k+1)x(k+1) by using the Cholesky Factorization of X'X such that

$$(X'X)^{-1} = R^{-1}R^{-T}$$

where the superscript -T denotes the inverse of the transope.

Then it follows that,

$$V(\hat{\underline{Y}}) = \sigma^2 \underline{x}'(R^{-1}R^{-T})\underline{x}$$

$$= \sigma^2 \underline{x}'R^{-1}(\underline{x}'R^{-T})' \qquad (5.4.2)$$

Note that $\underline{x}'R^{-1}$ is a vector of length (k+1) so that the variance of the predicted values of $\underline{Y}$ is the sum of squares of its elements.

Now consider $\underline{Y}$ using only the first p of the X-variables where p < k. Write $X = (X_1 : X_2)$, where X_1 consists of the first (p+1) columns of X and X_2 consists of the remaining (k-p) columns. Then, it is well known that if we obtain the choslesky factorization,

$$X_1' X_1 = R_1' R_1$$

then R_1 consists of the first (p+1) rows and columns of R and that the inverse R_1^{-1} is identical with the same rows and columns of R^{-1}. If $\underline{x}_1$ consists of the first (p+1) elements of $\underline{x}$ and $\underline{b}_1$ is the corresponding vector of least square estimator of regression coefficents for the model with only p variables, we have similar to (5.4.2)

$$V(\underline{x}_1'\underline{b}_1) = \sigma^2 (\underline{x}_1'R_1^{-1})(\underline{x}_1'R_1^{-1})' \qquad (5.4.3)$$

From (5.4.3), we see that the variance of the predicted values of $\underline{Y}$ is the sum of squares of the first (p+1) elements which are summed to obtain the variance of $\underline{x}'\underline{b}$, and hence,

$$V(\underline{x}'\underline{b}) \geq V(\underline{x}_1'\underline{b}_1) \qquad (5.4.4)$$

The relation (5.4.4) shows that, the variance of predicted values increases monotonically with the number of variables used in the prediction. Suppose if variables are deleted, the variance decreases but bias increases. If we consider the model without independent variables, the variance of prediction is zero but bias is probably large. Thus in such situation, one must seek a method to reduce the bias.

5.5 Bias Reduction Methods :

One of the major problems which needs substancial attention in subset selection is that of handling bias. Since it cannot be made zero, a question arises as to whether it can be reduced significantly. Miller (1990) briefly gives some methods of bias reduction. We discuss in this section two important methods of bias reductions namely (i) Jackknife Statistics and (ii) Bootstrap Method.

5.5.1 Jackknife Statistic

Quenouille (1949) introduced a technique for reducing bias. The 'Jackknife' procedure is as follows :

Suppose N data points are grouped into n groups of k observations each. Let $\hat{\theta}$ be the biased estimator of θ based on all N observations and let $\hat{\theta}_{.i}$, $i = 1,2,\ldots,n$ denote the estimator of θ obtained after deleting the i^{th} group observation. Define

$$\tilde{\theta}_i = n\,\hat{\theta} - (n-1)\,\hat{\theta}_i \qquad i = 1,2,\ldots,n.$$

Definition (5.5.1) Jackknife statistic : The Jackknife statistic (or esimator) $\tilde{\theta}$ to estimate θ is defined as,

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i$$

Note that

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[n\,\hat{\theta} - (n-1)\,\hat{\theta}_i\right],$$

$$\tilde{\theta} = n\,\hat{\theta} - (n-1)\,\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i.$$

Thus,

$$\tilde{\theta} = n\,\hat{\theta} - (n-1)\bar{\hat{\theta}},$$

where

$$\bar{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$$

The estimator $\tilde{\theta}$ is also called as "First order Jackknife estimator" of θ. The Jackknife technique is quite useful in subset selection. To see this, as usual consider the linear regression model of k variables

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}\ . \qquad (5.5.1)$$

The least square estimator of $\underline{\beta}$ is

$$\underline{b} = (X'X)^{-1}X'\underline{Y}$$

Now, we consider the subset linear regression model containing p (p<k) variables

$$\underline{Y} = X_1\underline{\beta}_1 + \underline{\varepsilon}.$$

The least square estimator of $\underline{\beta}_1$ is

$$\underline{b}_1 = (X_1'X_1)^{-1}X_1'\underline{Y}$$

The Jackknife estimator for $\underline{\beta}_1$ is given as follows:

Let $\underline{b}_{1(i)}$ be the estimate of $\underline{\beta}_1$ obtained after recomputing $\underline{b}_1$ with i^{th} pair (Y_i, X_i) deleted from the sample. Define p_i as

$$\underline{p}_i = n\,\underline{b}_1 - (n-1)\,\underline{b}_{1(i)} \qquad i = 1,2,\ldots n$$

The average of the p_i is the Jackknife estimator of $\underline{\beta}$ that is,

$$\hat{\underline{b}} = \sum_{i=1}^{n} p_i / n$$

The Jackknife estimator $\hat{\underline{b}}$ eliminates the n^{th} term from the bias (Gray & Schucany 1972) if

$$E\,(\hat{b} - \beta_1) = \sum_{i=1}^{n} a_r / n^r \quad \text{for all } n \qquad (5.5.2)$$

Using the above, we get

$$E(\hat{\underline{b}}) = E\left[\sum_{i=1}^{n} p_i/n\right],$$

$$= \sum_{i=1}^{n}\left[n\underline{b}_1 - (n-1)\underline{b}_{1(i)}\right]/n\,,$$

$$= n\,E\,(\underline{b}_1) - (n-1)\,E\,(\underline{b}_{1(i)}),$$

$$= n\left[\beta_{1}+a_{1}/n+a_{2}/n^{2}+\ldots\right]-(n-1)\left[\beta_{1}+a_{1}/(n-1)+a_{2}/(n-1)^{2}+\ldots\right]$$

$$= \beta_{1}+a_{1}-a_{1}+a_{2}/n-a_{2}/(n-1)+\ldots$$

$$= \beta_{1}-a_{2}/n(n-1)+O(1/n^{2}) \qquad (5.5.3)$$

On comparing (5.5.2) with (5.5.3), we see that the terms of order n^{-1} in the bias are eliminated, while those of order n^{-2} are reversed in sign and incereased very slightly in magnitude.

Miller(1984) used the Jackknifed technique in subset selection. The discussion is given below:

Suppose b_{1} is the least square estimate of the regression coefficent for a subset of variables selected using a particular procedure such as forward selection, All-subsets, etc. for n observations. As seen earlier, this regression coefficent b_{1} is biased because some of the variables are deleted from the model. Let $b_{1(i)}$ be the least square estimate of the regression coefficent of β_{1} obtained from (n-1) observations out of the n observations. Note that sample of (n-1) observations can be obtained in n different ways by deleting one-out-of the n observations. Consider all n such samples and apply the same selection procedure to each selected sample.

Suppose that in m out of the n cases the subset of interest is selected. Then we can use the m estimators of $b_{1(i)}$ for these cases in the Jackknife and average the results. In limited simulations, Miller (1984) has shown that the value of m has usually been close to n and rarely less than n/2. The Jackknife estimators $nb_{-1}-(n-1)\ b_{-1(i)}$ may be fairly successful at removing bias but the variance of Jackknife estimates is very large. The selection bias may be roughly proportional to $n^{-1/2}$ when the predictor variables are orthogonal and are all equally good choices. To eliminate this types of bias, the Jackknife statistic is modified as

$$\left[\ n^{1/2}b_{-1} - (n-1)^{1/2}\ b_{-1(i)}\ \right] \Big/ \left(\ n^{1/2}-(n-1)^{1/2}\ \right).$$

Below, we illustrate the use of Jackknife technique.

Example (5.5.1) We have generated five sets of random samples each of size 30 from $N_4(\underset{\sim}{\mu},\Sigma)$. where $\underset{\sim}{\mu}$ = [10 15 20 25]' and

$$\Sigma = \begin{bmatrix} 1 & 0.7 & 0.8 & -0.5 \\ & 1 & 0.2 & -0.8 \\ & & 1 & -0.1 \\ & & & 1 \end{bmatrix}$$

Further, we selected the best subset for each set by using forward selection method of subset selection. Then, on applying the Jackknife procedure, the following results are obtained.

Population regression coefficients	Sample regression coefficients	Jackknife regression coefficients	modified Jackknife regression coefficients
0.6053	0.588	0.5943	0.6009
0.6842	0.719	0.7104	0.7016

Quite evidently, the modified Jackknife estimators are 'closer' to population values.

5.5.2 The Bootstrap Method :-

Efron (1979) suggested another method for bias reduction which is called a "Bootstrap method". We discuss it below: Bootstrap method is a resampling scheme in which a one attempts to learn the sampling properties of a statistic by recomputing its value on the basis of a new sample realized from the orignal one. Efron (1979) points out that Bootstrap method is more widely applicale that Jackknife.

The general steps involved in bootstrap method are as follows:

Step 1: Resample a bootstrap sample from a given sample, that is draw a simple random sample of size n with replacement from the data of size n.

Step 2: Caluclate the value of statistic under consideration using this bootstrap sample.

Step 3: Repeat the Step-1 and 2 say 'r' times or required number of times.

Step 4: Take the average of all bootstrap sample statistics.

Consider the model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad (5.5.4)$$

The LSE $\underline{\beta}$ is

$$\underline{b} = (X'X)^{-1}X'\underline{Y} \qquad (5.5.5)$$

Bootstrap method is used in regression analysis as follows:

Step I: Regress Y on $X_1, X_2, \ldots, X_n$ such as

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

Step II: Compute the residual vector

$$e_i = Y_i - \hat{Y}_i \qquad i=1,2,\ldots,n. \qquad (5.5.6)$$

where $\hat{Y}_i$ denote the predicted value of Y_i.

Step III: Draw the simple random sample from residual vector e_i. It is denoted by e_i^*

Step IV: Obtain the Bootstrap observations Y_i^* as

$$Y_i^* = X_i b + e_i^* \qquad (5.5.7)$$

Step V : Obtain the LSE of $\underline{\beta}$ using Y_i^* ,say $\underline{b}^*$

Where

$$b_j^* = (X'X)^{-1}Y_j^* \qquad j=1,2,\ldots \qquad (5.5.8)$$

Step VI : Repeat the Step III to Step V for m required times.

Step VII: Calculate $\quad \underline{b}^* = [\underline{b}_1^* + \underline{b}_2^* + \quad + \underline{b}_m^*] / m$

.Mean and variance of Bootstrap estimator:

Let $b^* = (X'X)^{-1}X'Y^*$ be the Bootstrap estimator of $\underline{\beta}$ when the model is $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$.

Then,

$$E(\underline{b}^*) = E\left[(X'X)^{-1}X'\underline{Y}^*\right],$$

$$= (X'X)^{-1}X'E(\underline{Y}^*),$$

$$= (X'X)^{-1}X'X\underline{b},$$

Thus,

$$E(\underline{b}^*) = \underline{b}$$

Further,

$$\begin{aligned}
Cov(\underline{b}^*) &= Cov\ (X'X)^{-1}X'\underline{Y}^* \ , \\
&= (X'X)^{-1}X'cov(\underline{Y}^*)X(X'X)^{-1}, \\
&= (X'X)^{-1}X'\hat{\sigma}^2\ X(X'X)^{-1},
\end{aligned}$$

Therefore,

$$cov(\underline{b}^*) = \hat{\sigma}^2(X'X)^{-1}$$

(Since the Bootstrap value e^*_i use in (5.5.7) are independent with mean zero and variance

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i-X_i\underline{b})^2/n.$$

The implication that $\underline{b}$ is unbiased for $\underline{\beta}$ with covariance matrix approximetly equal to $\hat{\sigma}^2(X'X)^{-1}$.)

Bootstrap method is used in subset selection as follows:

Step I: Select a subset model from the full model containing k variables, such as $\underline{Y}=X\ \underline{\beta} + \underline{\varepsilon}$. Suppose selected subset model containing p variables be $\underline{Y} = X\beta_{1-1} + \underline{\varepsilon}$.

Step II: Regress Y on $X_1, X_2, \ldots X_p$ such as

$$\underline{Y} = X\beta_{1-1} + \underline{\varepsilon}.$$

Step III: Compute the residual vector $e_i = Y_i - \hat{Y}_i$, i=1,2,...n.

Step IV: Draw a simple random sample of size n from residual vector e_i. It is denoted by e^*_i.

Step V : Obtain the Bootstrap observations Y_i^* as

$$Y_i^* = \hat{Y}_i + e_i^* \qquad i=1,2,....n.$$

Step VI: Select a subset of p variables from the model$\underline{Y}^* = X \underline{\beta} + \underline{\varepsilon}$

Step VII: If the selected subset is same as that selected at step I. Then obtain the LSE of the selected subset, it is denoted by b_{1j}^*.

Step VIII: Repeat the Steps IV to Steps VI required number of times

Step IX: Calculate the b_1^* by taking avarage of LSE which are obtained at step VI.

Below, we illustrate the above procedure.

Example (5.5.2): We have generated a random sample of size 30 from multivariate normal distribution $N_4(\underline{\mu},\Sigma)$ where $\underline{\mu}$ and Σ are as given in example (5.5.1). We select a subset from the sample by using forward selection method. Compute the residual vector e from the selected model. Then draw the 20 Bootstrap sample from the original sample $\underline{e}$. Fit the model such as,

$$\underline{Y}^* = X\underline{\beta} + \underline{\varepsilon}.$$

The subset by using forward selection method are selected for each sample. Then consider the samples which are selected the same subset in original subset and take the average of that sample estimators. The average of the ten sample estimators which are selected same subset is given below.

Population regression coefficients	Sample regression coefficients	Bootstrap regression coefficients
0.6053	0.5420	0.5559
0.6841	0.7058	0.6926

In the above discussion, we have seen that if variables are dropped from the model then the estimator of parameter becomes biased estimator of $\underline{\beta}$. Now-a-days 'bias' is an important problem in subset selection. Here, we discuss two methods of bias reduction with illustrations, namely Jackknife statistic and Bootstrap method. The results of these methods show that bias decreases. Lastly, we conclude that these two methods are most useful for subset selection in regression analysis.