UNIT 12 SELECTION OF VARIABLES AND TESTING MODEL ASSUMPTIONS

Structure

12.1 INTRODUCTION

In Unit 11, we have discussed the multiple linear regression model and estimation of model parameters. You have learnt how to identify the important independent variables which contribute to the variation in the dependent variables. You have estimated their importance by estimating regression coefficients and their standard errors. You have also learnt how to examine the overall fit of the model by determining the coefficient of determination R^2 and Adjusted R^2.

In this unit, we discuss how to build a regression model that explains the variability in the dependent (response) variable Y. We assume that p independent variables $X_1, X_2, \ldots, X_p$ are available for inclusion in the model and we search for a set of variables that explains the variability in Y. Since we have to select a subset of these variables, we take care that the procedure of choosing it is such that the model is a good fit and the numbers of parameters is as small as possible. In Sec. 12.2, we discuss the test of significance for the extra sum of squares. In Sec. 12.3, we describe three selection procedures, namely, forward selection, backward elimination and stepwise selection procedures for selecting variables. We discuss various aspects of testing the model assumptions to check for multicollinearity, outlying observations, homogeneity of variances and normality of data in Sec.12.4.

Objectives

After studying this unit, you should be able to:

- calculate the extra sum of squares of the additional variable(s) in the model;
- test the usefulness of including additional variable(s) in the model;

- describe the forward selection, backward elimination and stepwise selection procedures for selecting variable(s);
- select an adequate model by using the methods of selection of independent variables; and
- test the model assumptions for multicollinearity, normality, outlying variables (outliers), homogeneity.

12.2 EXTRA SUM OF SQUARES

In the previous unit, you have learnt that in a regression model, the independent variables are responsible for satisfactory explanation of variability in the dependent (response) variable Y. If p independent variables X_1, X_2, …, X_p are available for inclusion in the model, each and every variable contributes to explanation of variability in the response variable Y in the model.

In this section, we discuss how to determine the variability accounted for by the variable which enters in the model at a later stage when one or more variables have already been included. This additional variability accounted for by the variable entered at a later stage over the variability accounted for by the variables already included in the model is called the e**xtra sum of squares**.

Suppose we have three independent variables X_1, X_2, and X_3. We can fit the following four types of models:

Model 1: $Y = B_0 + e$ which has only an intercept.

Model 2: $Y = B_0 + B_1X_1 + e$ which has one independent variable X_1.

Model 3: $Y = B_0 + B_1X_1 + B_2X_2 + e$ which has two independent variables X_1 and X_2.

Model 4: $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$ which has three independent variables X_1, X_2 and X_3.

We first fit Model 1 with only the intercept:

$$Y = B_0 + e \qquad \ldots(1)$$

The sum of squares due to the intercept B_0 is $n\overline{Y}^2$ and the residual sum of squares SS_{Res} is given by

$$SS_{Res}(B_0) = \sum\left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - n\overline{Y}^2 = \sum Y_i^2 - SS(B_0) \quad \ldots (2)$$

Next we fit Model 2. Let us denote the sum of squares due to B_0 and B_1 by $SS(B_0, B_1)$. The residual sum of squares is given by

$$SS_{Res}(B_0, B_1) = \text{å}\ Y_i^2 - SS(B_0, B_1) \qquad \ldots (3)$$

Hence, the additional (extra) sum of squares accounted for by B_1 is

$$SS(B_1|B_0) = SS_{Res}(B_0) - SS_{Res}(B_0, B_1) = SS(B_0, B_1) - SS(B_0) \quad \ldots (4)$$

If we have an estimate of σ^2, i.e., $\hat{\sigma}^2$, we can test whether it is worthwhile to include X_1 in the model. In other words, we test whether $B_1 = 0$ or not, i.e.,

$$H_0: B_1 = 0 \qquad \text{against} \qquad H_1 = B_1 \neq 0$$

If H_0 is accepted, $SS(B_1|B_0)$ is not significant. Here $SS(B_1|B_0)$ is the **extra sum of squares** accounted for by B_1 (or X_1).

We proceed in this way and fit Model (3). We find the sum of squares accounted for by B_0, B_1 and B_2, i.e., $SS(B_0, B_1, B_2)$. We also calculate the extra sum of squares due to B_2 after accounting for B_0 and B_1, i.e., $SS(B_2|B_0, B_1)$ by calculating $SS_{Res}(B_0, B_1, B_2)$ as follows:

$$SS_{Res}(B_0, B_1, B_2) = \sum Y_i^2 - SS(B_0, B_1, B_2) \quad \ldots (5)$$

and

$$\begin{aligned} SS(B_2|B_0, B_1) &= SS_{Res}(B_0, B_1) - SS_{Res}(B_0, B_1, B_2) \\ &= SS(B_0, B_1, B_2) - SS(B_0, B_1) \end{aligned} \quad \ldots (6)$$

and so on. At each stage we can test for the significance of the extra sum of squares by comparing it with $\hat{\sigma}^2$.

Note from equation (6) that the extra sum of squares depends on the variables, which have been accounted for. For example, in general, $SS(B_2|B_0, B_1)$ is not the same as $SS(B_1|B_0, B_2)$.

The test of significance of the extra sum of squares helps us in deciding whether it is worthwhile to include the additional variable and thereby an additional parameter or not. Let us try to understand this concept with the help of an example.

Example 1: A statistician collected data of 25 values with two independent variables X_1 and X_2. The following four models were considered one at a time:

1. $Y = B_0 + e$
2. $Y = B_0 + B_1X_1 + e$
3. $Y = B_0 + B_2X_2 + e$
4. $Y = B_0 + B_1X_1 + B_2X_2 + e$

The results obtained were:

$\hat{B}_0 = 22.38, \hat{B}_1 = 1.6161, \hat{B}_2 = 0.0144,\ SS(B_0) = 12526.08,$

$SS(B_0, B_1) = 17908.47,\ SS(B_0, B_2) = 17125.23,\ SS(B_0, B_1, B_2) = 18079.0,$

$\hat{\sigma}^2 = 10.53$, SE ($\hat{B}_1$) = 0.17 and SE ($\hat{B}_2$) = 0.0035.

Find the additional contribution of (i) X_2 over X_1 and (ii) X_1 over X_2. Test whether their inclusion in the model is justified or not.

Solution: Using the results given above, we calculate the extra sum of squares due to X_1 after accounting for X_0 :

$$\begin{aligned} SS(B_1|B_0) &= SS(B_0, B_1) - SS(B_0) \\ &= 17908.47 - 12526.08 = 5382.39 \end{aligned}$$

The extra sum of squares due to X_2 after accounting for X_0 is

$$\begin{aligned} SS(B_2|B_0) &= SS(B_0, B_2) - SS(B_0) \\ &= 17125.23 - 12526.08 = 4599.15 \end{aligned}$$

The extra sum of squares due to X_2 after accounting for X_0 and X_1 is:

$$SS(B_2|B_0, B_1) = SS(B_0, B_1, B_2) - SS(B_0, B_1)$$
$$= 18079.00 - 17908.47 = 170.53$$

The extra sum of squares due to X_1 and X_2 after accounting for X_0 is

$$SS(B_1, B_2|B_0) = SS(B_0, B_1, B_2) - SS(B_0)$$
$$= 18079.00 - 12526.08 = 5552.92$$

The extra sum of squares due to X_1 after accounting for X_0 and X_2 is

$$SS(B_1|B_0, B_2) = SS(B_0, B_1, B_2) - SS(B_0, B_2)$$
$$= 18079.00 - 17125.23 = 953.77$$

With only one term B_0 in the model, i.e., for Model 1, we see that the sum of squares due to X_0 is given by

$$SS(B_0) = 12526.08$$

If we add one more variable (X_1) to the model, (Model 2), the sum of squares due to X_0 and X_1 is

$$SS(B_0, B_1) = 17908.47$$

and the extra sum of squares accounted for by B_1 over B_0 is

$$SS(B_1|B_0) = 5382.39$$

If we test against $\hat{\sigma}^2 = 10.53$, the calculated value of F-statistic is

$$F = 5382.39/10.53 = 511.48$$

We compare this calculated value with the tabulated value $F_{(1,22)}$ for $\alpha = 0.05$, which is 4.30. Hence, the term is highly significant. This means that the contribution of X_1 is significant.

After X_1 has been included, it is of interest to see whether X_2 contributes significantly to the model. The additional or extra sum of squares due to Model 3 over Model 2, i.e., the extra sum of squares due to X_2 after accounting for X_0 and X_1 is given by

$$SS(B_2|B_0, B_1) = 170.53$$

If we test against $\hat{\sigma}^2 = 10.53$, the calculated value of F-statistic

$$F = 170.53/10.53 = 16.19$$

On comparing it with the tabulated value $F_{(1,22)}$ for $(\alpha = 0.05) = 4.30$, we find that the contribution of X_2 after X_1 is significant and it has to be included in the model for it to be worthwhile.

We may also be interested in testing for the contribution of X_1 when X_2 is already included in the model. The contribution in terms of the extra sum of squares due to X_1 after accounting for X_0 and X_2 is given by

$$SS(B_1|B_0, B_2) = 953.77$$

If we test against $\hat{\sigma}^2 = 10.53$, the calculated value of F-statistic is

$$F = 953.77/10.53 = 90.58$$

This is also significant.

Hence, it is worthwhile to include both X_1 and X_2 in the model. Therefore, the final model is

$$Y = B_0 + B_1X_1 + B_2X_2 + e$$

You may now like to solve the following exercises which will help you to assess your understanding of Sec.12.2.

E1) A statistician collected data of 78 values with two independent variables X_1 and X_2, the four models considered are the same as in Example 1 and the results are: $SS(B_0) = 652.42$, $SS(B_0, B_1) = 679.34$, $SS(B_0, B_2) = 654.00$, $SS(B_0, B_1, B_2) = 687.79$ and $\hat{s}^2 = 0.91$.

Find the additional contribution of (i) X_2 over X_1 and (ii) X_1 over X_2. Test whether their inclusion in the model is justified.

12.3 SELECTION OF INDEPENDENT VARIABLES

In Unit 11, you have learnt that the regressor variables included in the multiple regression model are considered to be important since the variability in the response variable is accounted for only by them. So far we have focused on the techniques to ensure that the functional form of the fitted model is correct and the underlying assumptions are not violated.

Along with the above considerations and assumptions we have set the following main objectives to obtain an adequate multiple regression model:

i) to cover as much information as possible in the model by including as many regressor variables as possible, and

ii) to increase the variability accounted for by the regressor variables by including only a few important regressor variables, so that fitted model is simple and adequate.

These two objectives contradict each other since adding a large number of regressor variables increases the accounted variability, and selecting only a few of them may reduce the accounted variability and may cause loss of useful information.

The problem of building an adequate model by striking a balance between both these objectives is not trivial. In multiple linear regression, we determine a group of regressor variables, which are proved to be important significantly, to find an adequate model. In some cases, theoretical considerations and personal experience may be helpful in selecting the important regressor variables to be included in the model. However, in most cases, the appropriate subset of the important variables to be included in the model needs to be examined. The methods of finding an appropriate subset of regressor variables, which adequately explain the variability in the response variable and need to be included in the model are called **variable selection methods**.

When a number of independent variables X_1, X_2, …, X_p are available, we may be interested in retaining only a few of them in the model, without sacrificing much of the explanatory power of the model. In this section, we describe some methods that help us in selecting important variables, which adequately explain the variability in the dependent variable Y. The main criterion is that the number of parameters should be as small as possible and all the variables which contribute significantly should be included in the model. When a variable X_i is included in the model, then any other variable, say, X_k, which is

highly correlated with X_i, need not be included as its contribution has already been accounted for by X_i. In the same way if X_k is included in the model first, then X_i may not be included as its contribution has already been taken into account by X_k. Thus, the order in which variables enter in the model is important.

We now describe three methods for selection of variables and model building: The forward selection method, the backward elimination method and the stepwise selection method.

12.3.1 Forward Selection Method

In this method, the most important variable is included first. The most important variable is the one which accounts for a maximum sum of squares (SS). After including this variable, the next most important variable is chosen which accounts for a maximum *additional* sum of squares. The contribution is tested at each stage and the most important variable, which accounts for a maximum additional sum of squares is included in the model. When the maximum contribution is not significant, the process is stopped. Otherwise, the process continues by including the variables with significant contributions.

In the forward selection method, variables are added to the model one at a time until the addition of a variable does not significantly improve the model. Thus, variables are added one at a time until further addition of a variable results in the acceptance of the reduced model.

Let us explain this method for two variables.

Suppose we have two possible regressor variables X_1 and X_2, which are linearly related with the response variable Y. The steps followed to obtain the adequate final model using forward selection method are explained below:

We start with two one variable models:

$$Y = B_0 + B_1 X_1 \quad \ldots (i)$$

$$Y = B_0 + B_2 X_2 \quad \ldots (ii)$$

We fit these models to the given data. The value of the coefficient of determination (R^2) is calculated for each model. The one with the highest R^2 is chosen and compared to the reduced model:

$$Y = B_0$$

Let the model $Y = B_0 + B_2 X_2$ have larger R^2. In this case, we test

H_0: $B_2 = 0$ (reduced model $Y = B_0$ is appropriate)

H_1: $B_2 \neq 0$ (model $Y = B_0 + B_2 X_2$ is needed)

If H_0 is rejected, the variable X_2 is included in the model. Otherwise, it is not included.

We now try to include the variable X_1. For this, we fit the two variable model to the given data:

$$Y = B_0 + B_1 X_1 + B_2 X_2$$

The value of the coefficient of determination (R^2) is calculated for the model $Y = B_0 + B_1 X_1 + B_2 X_2$ and compared to the reduced model $Y = B_0 + B_2 X_2$.

At this stage we again test

H_0: $B_1 = 0$ (reduced model $Y = B_0 + B_2 X_2$ is appropriate)

H_1: $B_1 \neq 0$ (model $Y = B_0 + B_1 X_1 + B_2 X_2$ is needed)

If H_0 is rejected, at this stage the variable X_1 is included in the model. The final model we obtain is the two variable model:

$$Y = B_0 + B_1 X_1 + B_2 X_2$$

12.3.2 Backward Elimination Method

In this method, all p variables are included in the model. Then the least important variable is deleted if its contribution to the sum of squares is not significant. At the second stage, the next least important variable whose contribution is not significant is deleted. This process goes on till a variable cannot be deleted from the model.

In backward elimination method, we begin with the model that includes all potential regressor or independent variables. Variables are deleted from the model one at a time until further deletion of a variable results in the rejection of the reduced model.

It is considered to be an economical procedure as it attempts to examine only the best regression equation containing a certain number of variables.

Suppose we have two potential regressor or independent variables X_1 and X_2. We obtain a best fitted model using the backward elimination method as follows:

First of all, we fit a full model with two regressor variables X_1 and X_2:

$$Y = B_0 + B_1 X_1 + B_2 X_2$$

Then we determine the value of R^2.

Now, we reduce the model by discarding one variable at a time. Suppose the two reduced models obtained on discarding the variables X_1 and X_2 one at a time are, respectively, given by

$$Y = B_0 + B_1 X_1$$

$$Y = B_0 + B_2 X_2$$

The value of R^2 and the extra sum of squares is obtained for both the reduced models and the model with higher extra sum of squares and R^2 is chosen and compared with the full model.

Let the reduced model $Y = B_0 + B_2 X_2$ have a higher value of R^2. Then, in this case, we test

H_0: $B_1 = 0$ (reduced model $Y = B_0 + B_2 X_2$ is adequate)

H_1: $B_1 \neq 0$ (full model $Y = B_0 + B_1 X_1 + B_2 X_2$ is needed)

If H_0 is accepted, we delete the regressor variable X_1 from the model since it appears to be unnecessary.

Now to test the significance of the variable X_2 in model $Y = B_0 + B_2 X_2$, we fit the model without any regressor variable, i.e.,

$$Y = B_0$$

We calculate R^2 and test it against the model with only variable X_2, i.e.,

H_0: $B_2 = 0$ (reduced model $Y = B_0$ is adequate)

H_1: $B_2 \neq 0$ (model $Y = B_0 + B_2 X_2$ is needed)

If H_0 is rejected, we are left with the model that contains one regressor variable X_2, i.e.,

$$Y = B_0 + B_2X_2$$

12.3.3 Stepwise Selection Method

The stepwise selection method begins with selecting an equation containing the single best independent variable X. Then, the second most important variable is included in the model. But at this stage, each of the variables selected earlier in the model are tested for their significance. If they are found to be significant, then we try to build an equation with the subsequent inclusion of other regressor variables one at a time. These additional variables are included as long as their inclusion is significant. A variable is excluded from the model if found insignificant. The selection process continues till no additional variable can be included or deleted in the model.

The order of addition is determined by using a test of significance based on partial F-test values. In order to decide the next variable to be included, the highest partial F-value is compared with the selected F-tabulated value (for the corresponding variable to be included). After a variable has been included, the fitted equation is examined to check whether some variable(s) can be deleted.

Stepwise selection method is a modified forward selection method. In the forward selection method, once a variable enters the model, it stays. Unfortunately, it is possible for a variable entering at a later stage to render a previously selected variable unimportant when these variables are themselves closely related. The forward selection method does not consider this possibility. In stepwise selection method, each time a new variable is entered into the model, all previous variables in the model are checked for continued importance.

It may be noted here that these procedures may not select a unique model. There could be several models which may explain the variability adequately and could be considered equivalent.

Suppose we have two regressor variables X_1 and X_2 and both are closely related. Let X_2 be the best single predictor. We now explain the stepwise selection method for this problem.

First of all, we fit two single variable models to the given data:

$$Y = B_0 + B_1 X_1$$

$$Y = B_0 + B_2 X_2$$

The value of R^2 is computed for each model and the model with larger R^2 and extra sum of squares is chosen and compared with the model with no

variables, i.e., $Y = B_0$. Let the selected model be $Y = B_0 + B_1 X_1$. We test this model with the reduced model as follows:

$H_0 : B_1 = 0$ (reduced model $Y = B_0$ is adequate)

$H_1 : B_1 \neq 0$ (selected model $Y = B_0 + B_1 X_1$ is needed)

If H_0 is rejected, we are left with the model that contains one regressor variable X_1, i.e.,

$$Y = B_0 + B_1 X_1$$

Now we fit the model with two variables X_1 and X_2, i.e.,

$$Y = B_0 + B_1 X_1 + B_2 X_2$$

and calculate R^2 and the extra sum of squares for this model. We test this model against the previously selected model as follows:

$H_0: B_2 = 0$ (model $Y = B_0 + B_1 X_1$ is adequate)

$H_1: B_2 \neq 0$ (model $Y = B_0 + B_1 X_1 + B_2 X_2$ is needed)

Let H_0 be rejected. This means that X_2 is to be entered in the model.

At this stage we check the importance of X_1 in this model. Therefore, we test the full model against the model with X_2 as the only variable, i.e.,

$H_0: B_1 = 0$ (model $Y = B_0 + B_2 X_2$ is needed)

$H_1: B_1 \neq 0$ (model $Y = B_0 + B_1 X_1 + B_2 X_2$ is adequate)

If H_0 is not rejected, it means that after entering the variable X_2 into the model, the variable X_1 does not have any importance. Therefore, we delete the variable X_1 from the model.

We end this section by applying all three methods to a problem.

Example 2: Consider the results of Example 1. Apply all three selection procedures to choose an appropriate model.

Solution: From the results of Example 1, we have

$$SS_{Res} = \sum Y_i^2 - \hat{B}\, Y'X = 231.63$$

Therefore, on putting the value of SS_{Res} in the expression of $\hat{\sigma}^2$, we get

$$\hat{\sigma}^2 = 231.63/(25-3) = 10.529$$

Using the results of Example 1, we get

$$\hat{B}_0 = 22.38, \quad \hat{B}_1 = 1.6161 \quad \text{and} \quad \hat{B}_2 = 0.0144$$

Using the above results along with those of Example 1 of this unit, we apply variable selection methods, one at a time, to find the adequate model as follows:

1. Forward Selection Method

We start with a model with only the intercept, i.e.,

$$Y = B_0 + e$$

and try to see if other variables can be added. For this, we fit two models with one variable each:

$$Y = B_0 + B_1 X_1 + e$$

and $$Y = B_0 + B_2 X_2 + e$$

According to the calculations and the models fitted by including the extra variables X_1 and X_2, separately and one at a time, the extra sum of squares due to X_1 over X_0, and X_2 over X_0 are:

$$SS(B_1|B_0) = 5382.39$$
$$SS(B_2|B_0) = 4599.15$$

As the extra sum of squares due to X_1 over X_0, i.e., $SS(B_1|B_0)$ is larger than the extra sum of squares due to X_2 over X_0, i.e., $SS(B_2|B_0)$, we test the hypothesis

$H_0: B_1 = 0$ against $H_1: B_1 \neq 0$

and get the t-statistic value

$$t = \frac{\hat{B}_1}{SE(\hat{B}_1)} = \frac{1.6161}{0.17} = 9.15$$

This value is significant in comparison with the t-tabulated value $t_{(0.5, 22)}$=2.074. Thus, the first variable to enter in the model is X_1. Hence at this stage, the selected model is

$$Y = B_0 + B_1X_1 + e$$

We now check whether the variable X_2 can be included in the model. The extra residual sum of squares due to X_2, after including X_1, is given by

$$SS(B_2|B_0, B_1) = 170.53$$

Then the value of F-statistic is calculated as

$$F = 170.53/10.53 = 16.19$$

which is significant. Hence, it is worthwhile to include X_2 after X_1 has been included in the model. The final model is

$$Y = B_0 + B_1X_1 + B_2X_2 + e$$

2. Backward Elimination Method

Here we consider the full model

$$Y = B_0 + B_1X_1 + B_2X_2 + e$$

and then try to identify and delete the least important variable if its contribution is not significant. For this, we fit two one variable models

$$Y = B_0 + B_1X_1 + e$$

and

$$Y = B_0 + B_2X_2 + e$$

We calculate the sum of squares due to X_1 and X_2 for the respective models. According to the calculation and model fitted by excluding the extra variables X_1 and X_2, one at a time, from the model, the sum of squares due to X_0, X_1 and X_2 , the sum of squares due to X_2 over X_0 and X_1, and the sum of squares due to X_1 over X_0 and X_2 are given as:

$$SS(B_0, B_1, B_2) = 18079.00$$

$$SS(B_2|B_0, B_1) = 170.53$$

$$SS(B_1 | B_0, B_2) = 953.77$$

From the above results, we see that the least important variable is X_2 since the extra sum of squares is least for X_2. We test its significance against $\hat{\sigma}^2$ =10.53 and get the F-statistic as:

$$F = 170.53/\hat{\sigma}^2 = 170.53/10.53 = 16.50$$

which is significant since the tabulated value $F_{(1, 22)} = 4.33$ at $\alpha = 0.05$.

We also test the hypothesis

$$H_0: B_2 = 0 \text{ against } H_1: B_2 \neq 0$$

and get the t-statistic value as

$$t = \frac{\hat{B}_2}{SE(\hat{B}_2)} = \frac{0.0144}{0.0035} = 4.11$$

which is significant in comparison to the tabulated value $t_{(0.5, 22)} = 2.074$.

Hence, we cannot delete any variable in the model and the final model is

$$Y = B_0 + B_1X_1 + B_2X_2 + e$$

3. Stepwise Selection Method

Here we start with the model with only the intercept, i.e.,

$$Y = B_0 + e$$

and try to see if regressor variables can be added. For this, we fit two models with one variable

$$Y = B_0 + B_1X_1 + e$$

and

$$Y = B_0 + B_2X_2 + e$$

According to the calculations and the models fitted by including the extra variables X_1 over X_0 separately, one at a time, the extra sum of squares due to X_1 over X_0, and X_2 over X_0 are as follows:

$$SS(B_1 | B_0) = 5382.39$$

$$SS(B_2 | B_0) = 4599.15$$

Since the extra sum of squares due to X_1 over X_0, i.e., $SS(B_1 | B_0)$ is larger than the extra sum of squares due to X_2 over X_0, i.e., $SS(B_2, | B_0)$, we now test the hypothesis

$$H_0: B_1 = 0 \text{ against } H_1: B_1 \neq 0$$

and get the t-statistic value as follows:

$$t = \frac{\hat{B}_1}{SE(\hat{B}_1)} = \frac{1.6161}{0.17} = 9.15$$

This value is significant in comparison with the tabulated value $t_{(0.5, 22)}$=2.074. Thus, the first variable to enter in the model is X_1. Hence, at this stage the selected model is

$$Y = B_0 + B_1X_1 + e$$

We now check whether the variable X_2 can be included in the model. The extra residual sum of squares due to X_2, after including X_1, is given by

$$SS(B_2|B_0, B_1) = 170.53$$

Then we calculate the value of F-statistic:

$$F = 170.53/10.53 = 16.19$$

which is significant. We also test the hypothesis

$H_0: B_2 = 0$ against $H_1: B_2 \neq 0$

and get the t-statistic value:

$$t = \frac{\hat{B}_2}{SE(\hat{B}_2)} = \frac{0.0144}{0.0035} = 4.11$$

which is significant in comparison with the tabulated value $t_{(0.5, 22)}$=2.074. Hence, it is worthwhile to include X_2 after X_1 has been included in the model.

At this stage we have to check whether X_1 can be excluded from the model. For this we calculate the sum of squares due to X_1 over X_0 and X_2, i.e., $SS(B_1|B_0, B_2)$ as follows:

$$SS(B_1|B_0, B_2) = 953.77$$

We now have to test whether this is significant or not. In case it is not significant, we can exclude X_1 from the model. We test its significance against $\hat{\sigma}^2 = 10.53$ and get

$$F = 953.77/10.53 = 90.58$$

which is significant. This suggests that we cannot delete any term from the model. Hence, the selected model is:

$$Y = B_0 + B_1X_1 + B_2X_2 + e$$

Therefore, all three procedures select the same final model.

You may now like to solve the following exercise which will help you assess your understanding of the three methods.

E2) Using the data and the results of E1), apply all three methods of selection of variables and comment on the selected models.

12.4 TESTING MODEL ASSUMPTIONS

So far we have discussed the linear regression model when certain assumptions are satisfied. Many problems arise when we fit the model to actual data and make inferences. There may be situations when the assumptions are violated due to some problems. Then we may need to test the model assumptions and modify the model. In this section, we consider the major problems and briefly discuss methods of overcoming them. Some of the major problems are associated with:

- Multicollinearity;
- Outlying observations;
- Fitting the correct model;
- Homogeneity of variances; and
- Normality of data.

12.4.1 Multicollinearity

This problem often occurs when the number of independent variables is large and some of them are linearly related. When this happens, it is difficult to solve normal equations. The solution of equations is not very stable and estimates are not very accurate. The variables Y and X may show very high correlation but their contribution, as shown by t values of test of significance, may be very small. Eliminating some variables from the model may make a very large change in the coefficients of other variables in the model. In such a situation, the prediction may not be very reliable. To overcome this problem we have to examine the correlations among independent variables and delete the variables which are highly correlated. However, this may not be very easy. In such cases it may be worthwhile to include only those variables which contribute significantly after accounting for the other variables in the model. To a great extent, this problem can be overcome by using an appropriate model selection procedure.

12.4.2 Outlying Observations

Sometimes when we collect observations, we may commit mistakes in recording them. In some cases the variable (Y) may have a large error component e. Such extreme observations have great influence on the estimates $\hat{B}$. In case such observations are removed, the estimates obtained may differ considerably. Therefore, it may be better to remove them from the data under consideration. For detecting such observations, we calculate the residuals $Y - \hat{Y}$ for each observation and standardise them by dividing them by $\hat{\sigma}$. If the absolute value of a standardised residual is larger than 3, we call it an outlying observation. Thus, if

$$\left| \frac{Y - \hat{Y}}{\hat{\sigma}} \right| \geq 3$$

we consider it is as an outlying observation. We examine such observations carefully and in some cases we may decide to analyse the data after neglecting such observations. However, in some cases, such observations may suggest that the model itself is not correct and we may revise the model.

12.4.3 Fitting the Correct Model

When we fit a model given by

$$Y = B_0 + B_1X_1 + ... + B_pX_p + e \qquad \ldots (1)$$

we assume that e is a random error with mean zero. However, very often the correct form of the model is not known and the error component e may not behave as random error with zero mean. This may create problem in the estimation of parameters and they may not give unbiased estimates. In such cases, inference obtained from the data may not be correct. Let us consider an example. Suppose a set of data comes from a quadratic model:

$$Y = B_0 + B_1X + B_2X^2 + e \quad \ldots (2)$$

However, a wrong model is chosen and a straight line is fitted:

$$Y = B_0 + B_1X + e \quad \ldots (3)$$

This means that now the error component e' is

$$e' = B_2X^2 + e \quad \ldots (4)$$

It not only contains the error term e but also the term B_2X^2. Hence, its expected value is given by

$$E(e') = B_2X^2$$

This is not zero but depends upon X. If we estimate B_0 and B_1 using the wrong model (3), then the estimates of B_0 and B_1 may not be unbiased and the fitted model may not be correct.

In case the correct model is used, we expect that the residuals r given by

$$r = Y - \hat{Y}$$

should behave like random error. In case they do not, it suggests that the correct form of the model has not been used. For this, we calculate the residuals r from the proposed model and then plot them against all regressor variables. If we find a random pattern of residuals, we may accept the model. Otherwise, we may modify it by including appropriate terms.

12.4.4 Homogeneity of Variances

One of our assumptions is that the variances of Y are constant. Very often this assumption is violated. Consequently, estimates of parameters are not very efficient. Moreover, the distribution of the test statistic may not follow the desired distribution. Very often the variances of Y may depend upon E(Y) and when E(Y) increases, variances may also increase. For checking this, we plot the residuals against the estimated means $\hat{Y}$ (Fig. 12.1). If the spread of residuals increases or decreases, it indicates that variances may depend on means.

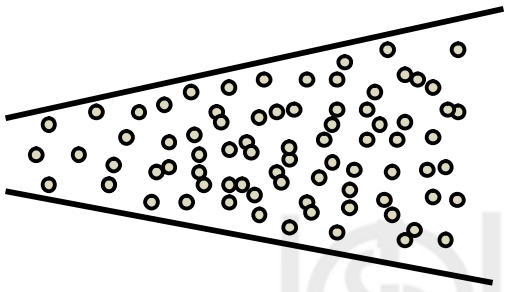

Fig.12.1: Residual plot indicating the increase in variability.

Fig. 12.1 shows that variances are increasing with estimated values of the dependent variable.

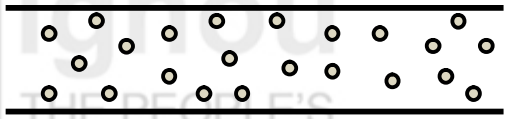

Fig.12.2: Residual plot of constant variability.

Fig. 12.2 shows that variances may be constant (homogeneous). In case variances are not constant, we may have to consider suitable transformation of Y. Sometimes heterogeneity of variances gives an indication that the model may not be a good fit.

12.4.5 Normality of Data

We have also assumed in our models that the observations (Y) are coming from a normal population. If this assumption is violated, the estimation of parameters may not be affected but the inferences regarding the test of hypotheses and confidence intervals are affected. The level of significance (α) and the probability level of confidence interval are affected. Moreover, the level of departure may not be known. For diagnosing their departure, we usually calculate the standardised residuals:

$$d_i = \frac{Y_i - \hat{Y}_i}{\hat{s}}, \quad i = 1, 2, ..., n$$

These are plotted on a normal probability plot. In case of normality, these residuals lie on a straight line of slope 1. If there is a significant departure from a straight line, it is an indication of departure from normality. Sometimes, the transformation of data with a change of variable using the log, inverse, square root functions may help them to behave like normally distributed variables.

We stop here and summarise the concepts we have discussed in this unit.

12.5 SUMMARY

1. The variability accounted for by the variable which enters in the model at a later stage over the variability accounted for by the variables already present in the model is called the **extra sum of squares**. Three methods are used for choosing important variables from a large number of independent variables: **forward selection method**, **backward elimination method** and **stepwise selection method**.
2. In a regression model, the independent variables are responsible for the satisfactory explanation of variability in the dependent (response) variable Y. If p independent variables X_1, X_2, …, X_p are available for inclusion in the model, each and every variable contributes for explanation of variability in the response variable Y in the model. In multiple linear regression, a group of significant regressor variables needs to be selected to find an adequate model.
3. In the forward selection method, variables are added to the model one at a time until the addition of a variable does not significantly improve the model. Thus, variables are added one at a time until further addition of a variable results in the acceptance of the reduced model.

4. In backward elimination method, we begin with the model that includes all potential regression or independent variables. Variables are deleted from the model one at a time until further deletion of a variable results in rejection of the reduced model.
5. Stepwise selection is a modified forward selection method. In forward selection, once a variable enters the model it stays. It is possible for a variable entering at a later stage to render a previously selected variable unimportant. Forward selection method does not consider this possibility. In stepwise regression, each time a new variable is entered into the model, all the variables in the previous model are checked for continued importance.
6. Very often the correct form of the model is not known and the error component may not behave as random error with mean value zero. This may create problems with the estimation of parameters and they may not give unbiased estimates. In such cases inference obtained from the data may not be correct.
7. The residuals plot is used for identifying the departure from assumptions, such as normality and homogeneity of variances.

12.6 SOLUTIONS/ANSWERS

E1) The following extra sums of squares for the data are given as:

$$SS = (B_1, B_2 / B_0) = 35.37,\ SS = (B_1 | B_0) = 26.82,\ SS = (B_2 | B_0) = 1.58$$

Therefore, the extra sum of squares due to X_2 over X_0 and X_1 and the extra sum of squares due to X_1 over X_0 and X_2 are

$$SS(B_2 | B_0, B_1) = SS(B_1, B_2 | B_0) - SS(B_1 | B_0) = 8.55$$

$$SS(B_1 | B_0, B_2) = SS(B_1, B_2 | B_0) - SS(B_2 | B_0) = 33.79$$

and the estimated value of σ^2 is $\hat{s}^2 = 0.91$ at 76 d.f.

Testing for X_2(after X_1):

$$F = SS(B_2 | B_0, B_1) / \hat{\sigma}^2 = 9.39$$

where $F_{1,86}$ at $\alpha = 0.05 = 3.95$. Hence, the additional contribution of X_2 over X_1 is significant.

Testing for X_1(after X_2):

$$F = SS(B_1 | B_0, B_2) / \hat{\sigma}^2 = 33.79 / 0.91 = 37.13$$

where $F_{1,86}$ at $\alpha = 0.05 = 3.95$. Hence, the additional contribution of X_1 over X_2 is also significant.

Hence their inclusion in the model is justified.

E2) Forward Selection Method

First we include X_1 as the extra sum of squares due to X_1 over X_0 is 26.82 and is much larger than the extra sum of squares due to X_2 over X_0, which is 1.58.

After X_1 has been included, the extra sum of squares due to X_2 is

$$SS(B_2|B_0, B_1) = 8.55$$

We now test for X_2 by comparing it with the estimate of σ^2:

$$F = 8.55/0.91 = 9.39$$

It is significant in comparison with $F_{(1, 86)}$ at $\alpha = 0.05$, which is 3.95. Hence, the selected model should contain both X_1 and X_2.

Backward Deletion Method

In this method, we consider the full model with X_1 and X_2:

$$Y = B_0 + B_1X_1 + B_2X_2 + e$$

We now try to delete the variable which accounts for the least sum of squares. We fit the two one variable models by deleting one variable, one at a time

$$Y = B_0 + B_1X_1 + e$$

$$Y = B_0 + B_2X_2 + e$$

We calculate the extra sum of squares due to X_1 and X_2 over X_0. From E1, we have

$$SS(B_2|B_0, B_1) = 8.55,\ SS(B_1|B_0, B_2) = 33.79$$

We can observe from the above results that X_2 is the least important variable in the model. Hence, we test for X_2 and calculate the F-statistic:

$$F = 8.55/0.91 = 9.39$$

It is significant in comparison with $F_{(1, 86)}$ at $\alpha = 0.05$, which is 3.95. Hence, we cannot delete X_2 from the model. Thus, both variables X_1 and X_2 are important.

Stepwise Selection Method

In this method, we start with the forward selection method and include X_1 in the model. Then we add X_2 at the second stage and examine whether X_1 can be excluded from the model. For this, we calculate the extra sum of squares due to X_1 after X_2 has been included in the model, i.e., $SS(B_1|B_0, B_2) = 33.79$.

We test for X_1 by comparing it to the estimate of σ^2 and get the F-statistic

$$F = 33.79/0.91 = 37.13$$

This is also significant in comparison to $F_{(1, 86)}$ at $\alpha = 0.05$, which is 3.95. Hence, we cannot delete X_1 after X_2 has been selected. Thus, the final model contains both variables X_1 and X_2.

Hence, all three methods select the model with both variables X_1 and X_2, i.e.,

$$Y = B_0 + B_1X_1 + B_2X_2 + e$$