UNIT 5 SIMPLE REGRESSION MODEL: INFERENCE*

Structure

5.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the concept of Testing of Hypothesis;
- derive the confidence interval for the slope coefficient in a simple linear regression model;
- explain the approach of 'test of significance' for testing the hypothesis on the estimated slope coefficient;
- describe the concept of Analysis of Variance (ANOVA);
- state the Gauss Markov Theorem with its properties; and
- derive the confidence interval for the predicted value of Y in a simple regression model.

5.1 INTRODUCTION

In Unit 4 we discussed the procedure of estimation of the values of the parameters. In this unit, we focus upon how to make inferences based on the estimates of parameters obtained. We consider a simple linear regression model with only one independent variable. This means we have one slope coefficient associated with the independent variable and one intercept term. We begin by recapitulating the basics of 'hypothesis testing'.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

5.2 TESTING OF HYPOTHESIS

Testing of hypothesis refers to assessing whether the observation or findings are compatible with the stated hypothesis or not. The word compatibility implies "sufficiently close" to the hypothesized value. It further indicates that we do not reject the stated hypothesis. The stated hypothesis is also referred to as 'Null Hypothesis' and it is denoted by H_0. The null hypothesis is usually tested against the 'alternative hypothesis', also known as maintained hypothesis. The alternative hypothesis is denoted by H_1. For instance, suppose the given population regression function is given by the equation:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{... (5.1)}$$

where X_i is personal disposable income (PDI) and Y_i is expenditure. Now, the null hypothesis is:

$$H_0 : \beta_2 = 0 \qquad \text{... (5.2)}$$

while the alternative hypothesis is:

$$H_1 : \beta_2 \neq 0 \qquad \text{... (5.3)}$$

We deliberately set the null hypothesis to 'zero' in order to find out whether *Y* is related to *X* at all. If *X* really belongs to the model, we would fully expect to reject the zero-null hypothesis H_0 in favour of the alternatives hypothesis H_1. The alternative hypothesis implies that the slope coefficient is different from zero. It could be positive or it could be negative. Similarly, the true population intercept can be tested by setting up the null hypothesis:

$$H_1 : \beta_1 = 0 \qquad \text{… (5.4)}$$

while the alternative hypothesis is:

$$H_1 : \beta_1 \neq 0 \qquad \text{… (5.5)}$$

The null hypothesis states that the true population intercept is equal to zero, while the alternative hypothesis states that it is not equal to zero. In case of both the null hypotheses,, i.e., for true population parameter or slope and the intercept, the null hypothesis as stated is a 'simple hypothesis'. The alternative hypothesis is composite. It is also known as a **two-sided hypothesis.** Such a two-sided alternative hypothesis reflects the fact that we do not have a strong apriori or theoretical expectation about the direction in which the alternative hypothesis must move from the null hypothesis. However, when we have a strong apriori or theoretical expectations, based on some previous research or empirical work, then the alternative hypothesis can be one-sided or unidirectional rather than two-sided. For instance, if we are sure that the true population value of slope coefficient is positive then the best way to express the two hypotheses is

$$H_0 : \beta_2 = 0$$

$$H_1: \beta_2 > 0$$

Let us take an example from macroeconomics. The prevailing economic theory suggests that marginal propensity to consume is positive. This means that the slope coefficient is positive. Now, suppose that the given population regression function is estimated by using a sample regression by adopting Ordinary Least Squares estimate. Let us also suppose that the results of sample regression yield the value of estimated slope coefficient as $b_2 = 0.0814$. This numerical value will change from sample to sample. We know that β_2 follows normal distribution, i.e., $b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\Sigma x_i^2}\right)$. There are two methods of testing the null hypothesis that the true population slope coefficient is equal to zero. The next two sections of this unit describe the two methods of testing of hypothesis of regression parameters.

5.3 CONFIDENCE INTERVAL

In this section, we shall derive the confidence interval for the slope parameter in equation (5.1) above. Note that the confidence interval approach is a method of testing of hypothesis. This is because it refers to the probability that a population parameter falls within the set of critical values from the Table. We make two assumptions, viz. (i) α, the level of significance on probability of committing type I error, is fixed at 5% level and (ii) the alternative hypothesis is two sided. From the *t*-table (given at the end of the book) we find the critical value of *t* at (n – k) degrees of freedom (d.f.) at $\alpha = 5\%$ is:

$$P(-2.306 \leq t \leq 2.306) = 0.95 \qquad \text{... (5.6)}$$

Substituting for 't', equation (5.6) can be re-written as:

$$P\left(2.306 \leq \frac{b_2 - \beta_2}{\hat{\sigma} / \sqrt{\Sigma x_i^2}} \leq 2.306\right) = 0.95 \qquad \text{... (5.7)}$$

Hence, the probability that *t* value lies between the limits –2.306, +2.306) is 0.95 or 95%. These are the critical *t* values. Substituting the value of *t* into equation (5.6) and rearranging the terms in (5.7) we get:

$$P[(b_2 - 2.306 .SE(b_2) \leq \beta_2 \leq b_2 + 2.306\, SE(b_2))] = 0.95$$

The above equation provides a 95% confidence interval for β_2. Such a confidence interval is known as the region of acceptance (for H_0) and the area outside the confidence interval is known as the region of rejection [for (H_0)]. If this interval includes the value of β_2 we do not reject the hypothesis; but if it lies outside the confidence interval, we reject the null hypothesis.

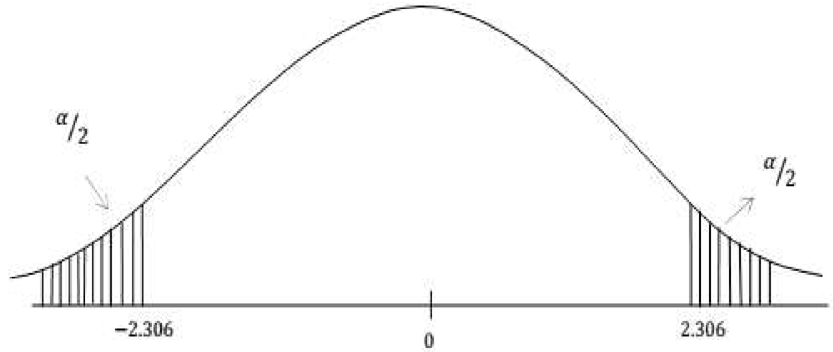

Fig 5.1: t-Distribution

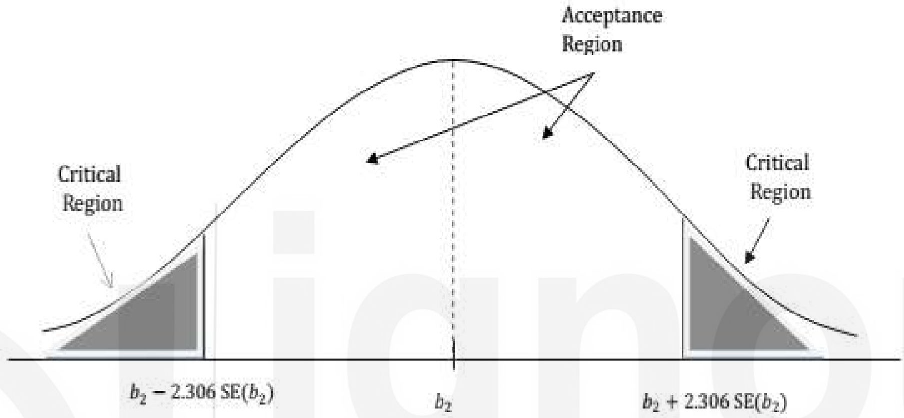

Fig 5.2: Confidence Interval for β_2

Check Your Progress 1 [answer questions in about 50-100 words]

1) State the difference between a simple and a composite hypothesis.

....

....

....

....

2) Null hypothesis is the indicator of simple and composite hypothesis. Is this statement true? Justify.

....

....

....

....

3) What is meant by a 'confidence interval'?

....

....

....

....

4) Why do we say that the interval contains the hypothesized value of true population parameter?

....

....

....

....

5.4 TEST OF SIGNIFICANCE

Test of significance approach is another method of testing of hypothesis. The decision to accept or reject H_0 is made on the basis of the value of *t*- test. It is computed by the statistic from the sample data as:

$$t = \frac{b_2 - \beta_2}{SE(b_2)} \qquad ...(5.8)$$

Equation (5.8) follows *t*-distribution with (n – k) degrees of freedom. The null hypothesis that we are testing here is:

$$H_0 : \beta_2 = \beta_2^* \qquad ... (5.9)$$

Note that β_2^* is some specific numerical value of β_2. Thus, the computed value of the test-statistic '*t*' will be like:

$$t = \frac{b_2 - \beta_2^*}{SE(b_2)} \qquad ...(5.10)$$

= [(estimated value) – (hypothesized value)] ÷ (standard error of estimator)

This can be computed from sample data as all values are available. The *t* value computed from (5.10) follows t distribution with (n – k) degrees of freedom (d.f.). This testing procedure is called the *t*-test. Fig. 5.3 depicts the region of rejection and the region of acceptance. One method of deciding on the result of the testing is to compare the computed value with the tabulated value (also called the 'critical value'). If the computed value of *t* is greater than the critical value of *t* then we reject the null hypothesis. This means we are rejecting the hypothesis that the true population parameter, or the slope coefficient, is zero. It implies that the explanatory variable plays a significant role in determining dependent variable. On the other hand, if the computed *t* value is less than critical value of *t*, then we do not reject the null hypothesis that the true value of the population parameter (or the slope coefficient) is zero. Not rejecting the null hypothesis implies that the value of slope coefficient is zero and that the explanatory variable does not play any significant role in determining the dependent variable.

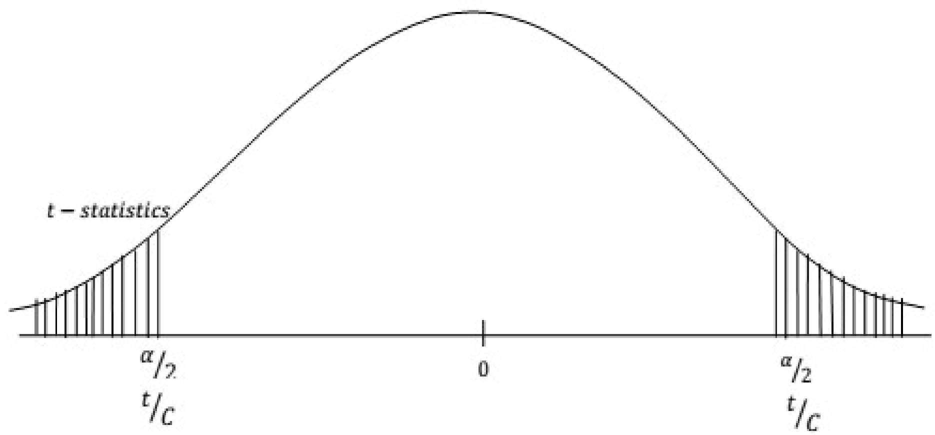

Fig 5.3: Test of Significance

In present times, when the results of the regression are obtained by computer, we usually get the *p*-value for the computed statistic. The p-value indicates the probability that the null hypothesis is true. If $p < 0.05$, we reject the null hypothesis and accept the alternative hypothesis. If $p > 0.05$, then we accept the null hypothesis. This means we base our test result at 5 percent level of significance. This also means that in 95 out of 100 independent samples, our result of the test will be similar. In other words, in 5 out of 100 cases, we could be coming to a wrong conclusion.

5.5 ANALYSIS OF VARIANCE (ANOVA)

Analysis of Variance (ANOVA) is a statistical tool used to analyse the given data for variations caused by several factors. These factors are divided into two parts: one is called the deterministic (or the systematic) part and the other is called the random part. This method of analysing the variance was developed by Ronald Fisher in 1918. Hence, this is also known as Fisher's analysis of variance. The ANOVA method separates the observed variance in the data into different components. It is used to determine the influence that the independent variables have on the dependent variable in a regression analysis. In a regression analysis ANOVA identifies the variability within a regression. Note that the total variability of dependent variable can be expressed in two parts as follows:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \qquad \text{... (5.11)}$$

Equation (5.11) distributes the total variation in the dependent variable *Y* into two parts, i.e., the variation in mean and the residual value. Squaring each of the terms in equation (5.11) and adding over all the *n* observations, we get the following equation.

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 \qquad \text{... (5.12)}$$

The above equation can be written as: TSS = ESS + RSS where TSS is the Total Sum of Squares, ESS is the Explained Sum of Squares, and RSS is the Residual

Sum of Squares. The RSS is also called the 'Sum of Squares due to Error (SSE)'. The ratio ESS / TSS is defined as the coefficient of determination R^2. The R^2 indicates the proportion of total sum of squares explained by the regression model. An ANOVA analysis is carried out with the help of a table (Table 5.1). From such a table of analysis of variance, the *F*-statistic can be computed as: ESS/RSS. This *F*-statistic is used to test the overall level of significance of the model. The null hypothesis and the alternative hypothesis for testing the overall significance using ANOVA are given by:

H_0: Slope coefficient is zero

H_1: Slope coefficient is not equal to zero.

Table 5.1: Format of a Typical ANOVA Table

Sources	**Degrees of Freedom (df)**	**Sum of Squares**	**Mean Square**	***F* Statistics = ESS /RSS**
Model	1	$\sum(\hat{Y}_i - \bar{Y})^2$	ESS / df	
Error	*n-2*	$\sum(Y_i - \hat{Y}_i)^2$	RSS / df	
Total	*n-1*	$\sum(Y_i - \bar{Y})^2$	TSS / df	

F = $\frac{ESS/(k-1)}{RSS/(n-k)}$ gives the observed value. The *F*-critical value at (*k-1*) and (*n-k*) degrees of freedom can be located from the statistical table. When the computed *F* is > than *F*-critical, the null hypothesis is rejected. Since the alternative hypothesis is accepted, the inference is that the explanatory variable plays a crucial role in determining the dependent variable. Similarly, when the *F* computed is < than the *F*-critical, the null hypothesis is not rejected. In this case, the hypothesis that the explanatory variable plays no role in determining the dependent variable is accepted. Again, here also, we can base our inference based on the *p*-value. This means if $p < 0.05$, we reject the null hypothesis.

Check Your Progress 2 [answer questions in about 50-100 words]

1) What is ment by the 'test of significance approach' to hypothesis testing?

....

....

....

....

....

....

2) What does the 'level of significance' indicate?

....

....

....

....

3) What purpose does an ANOVA serve?

....

....

....

....

4) Distinguish between *t*-test in a regression model.

....

....

....

....

5.6 GAUSS-MARKOV THEOREM

This is an important theorem which gives us the condition under which the least squares estimator is the best estimator. When the assumptions of the classical linear regression model are not violated, the least-squares estimator fulfils certain optimum properties. These properties are summarised in the Gauss-Markov theorem which is stated as follows:

Gauss-Markov Theorem: Given the assumptions of classical linear regression model, the least-squares estimators, have minimum variance, in the class of all unbiased linear estimators, i.e., they are BLUE [best linear unbiased estimator(s)]. The characteristic of BLUE implies that the estimator obtained by the OLS method has the following properties.

a) It is *linear*, i.e., the estimator is a linear function of a random variable (such as the dependent variable *Y* in the regression model).
b) It is *unbiased*, i.e., its average or expected value is equal to true value [$E(b_2) = \beta_2$].
c) It has *minimum variance* in the class of all such linear unbiased estimators. In other words, such an estimator with the least variance is an efficient estimator.

Thus, in the context of regression, the OLS estimators are BLUE. This is the essence of Gauss-Markov Theorem.

5.7 PREDICTION

So far we have spoken about estimation of population parameters. In the two variable model, we derived the OLS estimators of the intercept (β_1) and slope (β_2) parameters. Prediction refers to estimation of the dependent value at a particular value of the independent variable. In other words, we use the estimated regression model to predict the value of Y corresponding to a given value of X.

Prediction is important to us for two reasons: First, it helps us in policy formulation. On the basis of the econometric model, we can find out the impact of changes in the explanatory variable on the dependent variable. Second, we can find out the robustness of our estimated model. If our econometric model is correct, the error between forecast value and actual value of the dependent variable should be small. Prediction could be of two types, as mentioned below.

5.7.1 Individual Prediction

If we predict an individual value of the dependent variable corresponding to a particular value of the explanatory variable, we obtain the individual prediction. Let us take a particular value of X, say X = X_0. Individual prediction of Y at X = X_0 in obtained by:

$$Y_0 = \beta_1 + \beta_0 X_0 + u_0 \qquad \dots (5.13)$$

We know that b_1 and b_2 are unbiased estimators of β_1 and β_2. Hence, $\hat{Y}_0$ is an unbiased predictor of E (Y | X_0).

Therefore,

$$\hat{Y}_0 = b_1 + b_2 X_0 \qquad \dots (5.14)$$

Since $\hat{Y}_0$ is an estimator, the actual value Y_0 will be different from $\hat{Y}_0$, and there will be certain 'prediction error'.

The prediction error in $[\hat{Y}_0 - Y_0]$ is given by

$$\hat{Y}_0 - Y_0 = (b_1 + b_2 X_0) - (\beta_1 + \beta_2 X_0 + u_0) \qquad \dots (5.15)$$

We can re-arrange the terms in equation (5.15) to obtain

$$\hat{Y}_0 - Y_0 = (b_1 - \beta_1) + (b_2 - \beta_2)X_0 - u_0$$

Let us take expected value of (5.15).

$$E(\hat{Y}_0 - Y_0) = E(b_1 - \beta_1) + E(b_2 - \beta_2)X_0 - E(u_0) \qquad \dots (5.16)$$

We know that $E(b_1) = \beta_1$, $E(b_2) = \beta_2$ and $E(u_0) = 0$.

Thus, we find that expected value of prediction error is zero.

Now let us find out the variance of the prediction error.

The variance of the prediction error,

$$V(\hat{Y}_0 - Y_0) = V(b_1 - \beta_1) + V(b_2 - \beta_2)X_0$$

$$+2\, X_0\, cov(b_1 - \beta_1, b_2 - \beta_2) + V(u_0) \qquad \ldots (5.17)$$

We know that

$$V(b_1) = \sigma^2 \frac{\Sigma X_i^2}{n\Sigma x_i^2} \qquad \ldots (5.18)$$

$$V(b_2) = \frac{\sigma^2}{\Sigma x_i^2} \qquad \ldots (5.19)$$

$$Cov(b_1, b_2) = -\bar{X}\left(\frac{\sigma^2}{\Sigma x_i^2}\right) \qquad \ldots (5.20)$$

By combining the above three equations and re-arranging terms, we obtain

$$V(\hat{Y}_0 - Y_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})}{\Sigma x_i^2}\right] \qquad \ldots (5.21)$$

Thus, Y_0 follows normal distribution with mean $\beta_1 + \beta_0 X_0$ and variance $\sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}\right]$.

If we take estimator for σ^2, then we have

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_0 X_0)}{SE(\hat{Y}_0)} \qquad \ldots (5.22)$$

On the basis of (5.22) we can construct confidence interval for $\hat{Y}_0$. Since

$t = \frac{\hat{Y}_0 - E(Y/\alpha_0)}{SE(\hat{Y}_0)}$, we have

$$P\left[-t_{\alpha/2} \leq t \leq t_{\alpha/2}\right] = 1 - \alpha$$

Thus, the confidence interval of $\hat{Y}_0$ is

$$P\left[(b_1 + b_2 X_0) - t_{\alpha/2} SE(\hat{Y}_0) \leq (\beta_1 + \beta_2 X_0) \leq (b_1 + b_2 X_0) + t_{\alpha/2} SE(\hat{Y}_0)\right] = 1 - \alpha \qquad \ldots (5.23)$$

Let us lok into equation (5.21) again. We see that the variance of $\hat{Y}_0$ increases with $(X_0 - \bar{X})^2$. Thus, there is an increase in variance if X_0 is farther away from $\bar{X}$, the mean of the sample on the basis of which $b_1$and $b_2$are computed. In Fig. 5.4 we depict the confidence interval for $\hat{Y}_0$ (see the dotted line)

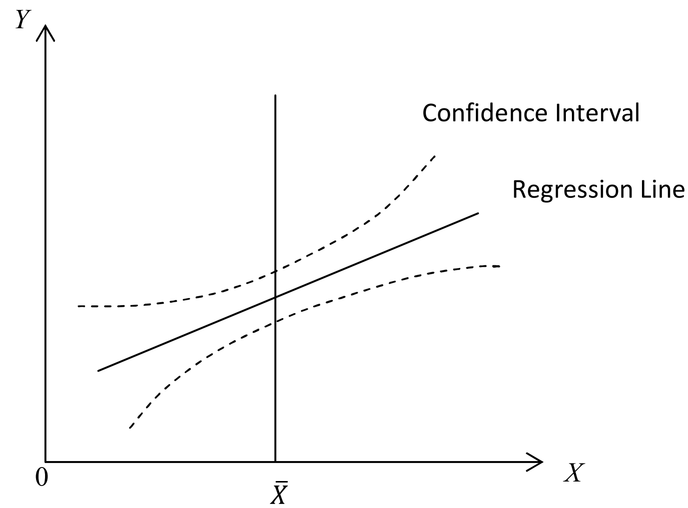

Fig. 5.4:Confidence Interval for Individual Prediction

5.7.2 Mean Prediction

It refers to prediction of expected values of Y_0, not the individual value. In other words, we are predicting the following:

$$\hat{Y}_0 = b_1 + b_2 X_0$$

Thus the error term u_0 is not added.

In the case of mean prediction, the prediction error in $[\hat{Y}_0 - Y_0]$ is given by

$$\hat{Y}_0 - Y_0 = (b_1 + b_2 X_0) - (\beta_1 + \beta_2 X_0) \qquad \dots (5.24)$$

We can re-arrange the terms in equation (5.24) to obtain

$$\hat{Y}_0 - Y_0 = (b_1 - \beta_1) + (b_2 - \beta_2) X_0$$

If we take the expected value of (5.24)

$$\mathrm{E}(\hat{Y}_0 - Y_0) = E(b_1 - \beta_1) + E(b_2 - \beta_2) X_0 \qquad \dots (5.25)$$

Thus, we find that expected value of prediction error is zero.

Now let us find out the variance of the prediction error in thecase of mean prediction.

The variance of the prediction error,

$$\mathrm{V}(\hat{Y}_0 - Y_0) = V(b_1 - \beta_1) + V(b_2 - \beta_2) X_0$$
$$+2\, X_0\, cov(b_1 - \beta_1, b_2 - \beta_2) \qquad \dots (5.26)$$

If we compare equations (5.17) and (5.26) we notice an important change – the term $V(u_0)$ is not there in (5.26). Thus the variance of the prediction error in the case of mean prediction is less compared to individual prediction. There is a change in the variance of $\hat{Y}_0$ in the case of mean prediction, however. Variance of the prediction error, in the case of mean prediction is given by

$$V(\hat{Y}_0 - Y_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})}{\Sigma x_i^2}\right] \quad \text{... (5.27)}$$

Again, there is an increase in the variance of prediction error if X_0 is farther away from $\bar{X}$, the mean of the sample on the basis of which b_1 and b_2 are computed. It will look somewhat like the confidence interval we showed in Fig. 5.4, but the width of the confidence interval will be smaller.

An inference we draw from the above is that we can predict or forecast the value of the dependent variable, on the basis of the estimated regression equation, for a particular value of the explanatory variable (X_0). The reliability of our forecast, however, will be lesser if the particular value of X is away from $\bar{X}$.

Check Your Progress 3 [answer questions within the given space in about 50-100 words]

1) State Gauss-Markov Theorem.

....

....

....

....

....

2) Differentiate between the two types of prediction possibilities in forecasting.

....

....

....

....

....

5.8 LET US SUM UP

This unit explains how to make inference on the estimated results of a simple regression model. After presenting an account of hypothesis testing to recapitulate the basics, it explains the two approaches for deciding on the validation of estimated results. The two methods are: confidence interval approach and test of significance approach. The testing of overall significance of the model is explained by the technique of ANOVA. Here, the application of F – statistic is explained. The assumptions of classical linear regression model leads to the estimated parameters enjoying some unique properties. In light of this, the estimates are called BLUE (best linear unbiased estimates). This fact is stated in a result called the Gauss Markov theorem. The unit concludes with a detailed account of the concept of forecasting. This is once again a technique in which we have presented a confidence interval wherein the predicted or forecasted value of the dependent variable is shown to lie.

5.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) In case of both the null hypothesis (, i.e., for true population parameters of slope and intercept), the null hypothesis are a simple hypothesis, whereas the alternative hypothesis are composite. The former is usually equated to zero (unless equated to a known value) and the latter in stated in inequality terms. The latter is also known as **two-sided hypothesis** when stated in 'not equal to' terms**.** It is considered one sided if stated in > or < terms.

2) False. It is the alternative hypothesis that decides whether it is composite or one-sided hypothesis. If the alternative hypothesis is stated as not equal to zero then it is composite or two-tailed test. Otherwise, i.e., if the alternative hypothesis is stated in positive or negative terms, then it will be a one-sided test.

3) The confidence interval approach is a method of testing of hypothesis. It refers to the probability that a population parameter falls within the set of critical values drawn from the Table.

4) We say that the hypothesised value is contained in the interval because the value of the interval depends upon the sample or the data used for estimation. The true population parameter value is fixed but the interval changes depending on the sample.

Check Your Progress 2

1) The test of significance approach is another method of testing of hypothesis. The decision to accept or reject H_0 is made on the basis of the value of test statistic obtained from the sample data. This test statistic is given by: $t = \frac{b_2 - \beta_2}{SE(b_2)}$ and it follows $t -$distribution with (n −1 d.f.)

2) It is a measure of the strength of evidence when the null hypothesis is rejected It concludes that the effect is statistically significant. It is the probability of rejecting the null hypothesis when it is true. This is a grave error to commit and hence is chosen in a small measure like 1% or 5%.

3) Analysis of Variance (ANOVA) is a technique or a tool used to analyse the given data in two ways or direction. One is attributed to the deterministic factors, also called the explained part or the systematic part. The other is called the random or the unexplained part. This method of analysis of variance method was developed by Ronald Fisher in 1918.

4) The *t*-test is used to test the significance of estimated individual coefficients. It is distributed as t with $(k - 1)$ degrees of freedom (d.f.). where k is the number of parameters estimated including the intercept term. Thus, for a simple linear regression, it is $[n - (2 - 1)] = (n - 1)$. The *F*-distribution is used for testing the significance of the whole model. It has two parameters. The d.f. for a F test, in general is $(k - 1)$ and $(n - k)$. K includes the intercept term. Hence, in a simple linear regression, the d.f. for F is: $(2 - 1)$ and $(n - 2)$ or 1

and ($n - 2$) Note that in a simple linear regression, the t test and the F test are equivalent because the number of independent variable is only one.

Check Your Progress 3

1) The Gauss-Markov theorem states that the Ordinary Least Squares (OLS) estimators are also the best linear unbiased estimator (BLUE). The presence of BLUE property implies that the estimator obtained by the OLS method retains the following properties: (i) it is linear, i.e., the estimator is a linear function of a random variable such as the dependent variable Y in the regression model; (ii) it is unbiased, i.e., its average or expected value is equal to the true value in the sense that $E(b_2) = \beta_2$; (iii) it has minimum variance in the class of all such linear unbiased estimators. Such an estimator with the least variance is also known as an efficient estimator.

2) Prediction implies predicting two types of values: prediction of conditional mean, i.e., $E(Y \mid X_0)$ → a point on the population regression line. This is called as the Mean Prediction. Prediction of individual Y value, corresponding f (X_0) is called the Individual Prediction.