UNIT 10 STATISTICAL INFERENCE IN SIMPLE LINEAR REGRESSION

Structure

10.1 INTRODUCTION

In Unit 9, you have studied the simple linear regression model for one dependent variable Y and one independent variable X. In linear regression analysis, the parameters a and b are assumed to be unknown and for the given data on Y and X, we obtain the least squares estimates of these parameters. The estimates of parameters a and b are linear functions of observations Y. Since Y is a random variable, the least squares estimates $\hat{a}$ and $\hat{b}$ are also random variables. In Unit 9, we have also descrived the method of detecting model deficiencies using residual analysis.

In this unit, we discuss the inferential aspect of simple linear regression. We examine whether there is a significant linear relationship between Y and X –we test whether regression coefficient b is zero or different from zero in Sec. 10.2. If it is zero, then there is no significant dependence of Y on X. Hence, the application of regression will not be useful. If b is not zero, it may be of interest to explore how much variability in Y is accounted for by this linear relationship, in other words, how good the fit is.

The coefficient of determination, R^2, gives a measure of goodness of fit of the model and is discussed in Sec. 10.3. It is simply the square of the correlation coefficient between X and Y and lies between 0 and 1. When the linear relationship between Y and X is good, most observations lie very close to the fitted line $\hat{Y} = \hat{a} + \hat{b}X$. If the relationship between Y and X is poor, the value of R^2 would be close to zero, and most values of Y lie far away from the fitted line. We also obtain the confidence interval of b and the predicted value of E(Y) corresponding to a given value of X in Sec. 10.4.

In the next unit, we shall discuss the multiple linear regression model and the estimation of model parameters. We shall also discuss a method for calculating the coefficient of determination (R^2) and Adjusted R^2.

Objectives

After studying this unit, you should be able to:

- test the significance of regression coefficients;

- describe and determine the coefficient of determination, R^2; and
- form the confidence interval of regression coefficient b and calculate the predicted value of E(Y) corresponding to a given value of X.

10.2 TEST OF SIGNIFICANCE OF REGRESSION COEFFICIENTS

For testing the significance of regression coefficients, we need to follow the assumptions, which have been discussed in Sec. 9.2.1 in Unit 9. For convenience, we state them again:

i) We always assume that the independent variable or predictor variable X is not a random variable and it takes fixed values. In some cases, these may be random variables. But here we make inferences by putting the condition on them that these are fixed (not random) variables.

ii) We assume that for a given (fixed) X, Y has the following mean and variance:

$$\text{Mean} = E(Y|X) = a + b\,X$$

and $$\text{Variance} = V(Y|X) = \sigma^2$$

Further, the error e is distributed normally with mean zero and variance σ^2, i.e.,

$$E(e) = 0; \qquad V(e) = \sigma^2$$

iii) We further assume that Y_i, i =1, 2,…, n, are independent. Thus, they are normally distributed random variables with mean (a+bX) and variance σ^2.

10.2.1 Distribution of Regression Coefficients

In Sec. 9.3.1 of Unit 9, we have discussed the estimation of regression coefficients a and b using the method of least squares. We have denoted these least squares estimates by $\hat{a}$ and $\hat{b}$. These are given by

$$\hat{a} = \bar{Y} - \hat{b}\,\bar{X} \quad \text{and} \quad \hat{b} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \qquad \ldots (1a)$$

The parameters $\hat{a}$ and $\hat{b}$ are unbiased estimates of regression parameters a and b, respectively, and their variances are given as

$$V(\hat{a}) = s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right) \text{ and } V(\hat{b}) = \frac{\sigma^2}{SS_X} \qquad \ldots (1b)$$

where $SS_X = \sum (X_i - \bar{X})^2$

Note that $\hat{a}$ and $\hat{b}$ are linear functions of Y_i s. Here we use a simple property of normally distributed variables that a new variable, which is a linear combination of normally distributed random variables, also follows the normal distribution with mean and variance of the linear combination. Thus, $\hat{a}$ and $\hat{b}$ are normally distributed with mean and variance as given above. Therefore, we write

$$\hat{a} \sim N\left(a, s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right)\right) \quad \hat{b} \sim N\left(b, \frac{s^2}{SS_X}\right) \qquad \ldots (2)$$

We cannot determine variances of $\hat{a}$ and $\hat{b}$ unless we know the value of σ^2. Generally, it is not known and has to be estimated from the data. Its estimation has already been discussed in Unit 9 and an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \sum (r_i - \bar{r})^2 / (n-k) = SS_{Res}$$

In the case of simple linear regression model, we have two parameters a and b. Thus, k = 2, and an unbiased estimate of σ^2 is given by

$$SS_{Res} = \sum (r_i - \bar{r})^2 / (n-2) = \sum r_i^2 / (n-2) \qquad \ldots(3)$$

Let us understand these concepts with the help of an example.

Example 1: Using the data of Example 1 of Unit 9, calculate SS_{Res} as an estimate of σ^2. Use this estimate for calculating the variances of $\hat{a}$ and $\hat{b}$.

Solution: We have the following values (from the solution of Example 1, Sec. 9.4 of Unit 9):

$$\bar{X} = 6.97, \; SS_X = \sum (X_i - \bar{X})^2 = 6.40$$

$$SS_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 61.81$$

$$SS_Y = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2 = 2801 - 10 \times (14)^2 = 640.1$$

$$\sum r_i^2 = \left(SS_Y - \frac{(SS_{XY})^2}{SS_X} \right) = \left(640.1 - \frac{(61.81)^2}{6.40} \right)$$

$$= 640.10 - 596.95 = 43.51$$

Putting the above values in equation (3), we get

$$\hat{\sigma}^2 = \sum \frac{r_i^2}{8} = 5.42$$

Using equation (1a) for $\hat{a}$ and $\hat{b}$ and substituting $\hat{\sigma}^2$ for σ^2 in equation (1b), we get

$$V(\hat{a}) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X} \right) = 5.42 \left\{ \frac{1}{10} + \frac{(6.97)^2}{6.40} \right\}$$

$$= 5.42 \; (0.10 + 7.59) = 5.42 \times 7.69 = 41.68$$

$$V(\hat{b}) = \frac{\hat{\sigma}^2}{SS_X} = \frac{5.42}{6.40} = 0.85$$

10.2.2 Hypothesis Testing of Regression Coefficients

The regression coefficients, which we have stimated in Unit 9, are subject to random variation as they depend on Y. It means that while working with different samples, we obtain different estimates. Hence, the sample estimates

of a and b may differ from the population values. As population values of a and b are not known, one may hypothesise that the sample has come from a population whose true values are $a = a_0$ and $b = b_0$. On the basis of a sample (Y_i, X_i), i = 1, 2, …., n, one may like to test whether these hypotheses are acceptable or not.

This is particularly important when one is interested in knowing whether Y and X are linearly related or not. If Y and X are not linearly related, b = 0 and the population relationsip is given as:

$$Y = a + e$$

rather than $Y = a + bX + e$

Hence, we may be interested in testing a null hypothesis

$$H_0: b = 0$$

against the alternative hypothesis

$$H_1 : b \neq 0$$

Or, in general, we may like to test two hypotheses

i) $H_0: a = a_0$ against $H_1: a \neq a_0$

ii) $H_0: b = b_0$ against $H_1: b \neq b_0$

These are two hypotheses and we shall derive the tests for these hypotheses separately.

i) Test of $H_0: a = a_0$ against $H_1: a \neq a_0$

We have seen that $\hat{a}$ is an unbiased estimate of a. From equation (2), it is normally distributed as follows:

$$\hat{a} \sim N\left\{a, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right)\right\}$$

Thus, if we consider deviation of $\hat{a}$ from a and divide it by its standard error, we obtain a variable which follows standard normal N (0, 1).

If the null hypothesis $H_0: a = a_0$ is correct, then we write the z-statistic as

$$Z = \frac{(\hat{a} - a_0)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right)}} \sim N(0,1)$$

As we do not know σ^2, we cannot use this as a test statistic. Therefore, we replace σ^2 by $\hat{\sigma}^2$ and get a t-statistic as

$$t = \frac{(\hat{a} - a_0)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right)}} \quad \ldots (4)$$

which is distributed as a Student's 't'-statistic with (n – 2) d.f.

This is an appropriate test statistic to test the null hypothesis:

$H_0: a = a_0$ against $H_1: a \neq a_0$

If we want to test at α level of significance, this will be a two-sided test and one can compare the calculated value of |t| with tabulated value at α level of significance, i.e., $t_{\alpha/2}$. We accept H_0, if $|t| < t_{\alpha/2}$. Otherwise, we reject the null hypothesis. One-sided test can also be constructed in a similar way.

You have learnt about testing of hypotheses in the course MST-004 entitled Statistical Inference.

ii) Test of $H_0: b = b_0$ against $H_1: b \neq b_0$

From equation (2), $\hat{b}$ is distributed as

$$\hat{b} \sim N\left\{b, \frac{\sigma^2}{SS_X}\right\}$$

If the null hypothesis $H_0: b = b_0$ is correct, then we write the z-statistic as

$$Z = \frac{(\hat{b} - b)}{\sqrt{\sigma^2 / SS_X}} \sim N(0,1)$$

Since σ^2 is not known, we replace σ^2 by $\hat{\sigma}^2$ and get a t-statistic as

$$t = \frac{(\hat{b} - b)}{\sqrt{\hat{\sigma}^2 / SS_X}} \qquad \ldots (5)$$

where the statistic 't' follows Student's t-distribution with $(n - 2)$ d.f., under H_0. This is an appropriate test for the null hypothesis. The value |t| is calculated from equation (5) and is compared with the tabulated value at α level of significance, i.e., $t_{\alpha/2}$ (for two-sided hypothesis) and if $|t| < t_{\alpha/2}$, we accept the hypothesis. Otherwise, we reject H_0. One-sided null hypothesis can also be tested in a similar way. Let us explain this further taking an example.

Example 2: Using the data of Example 1 given in Sec. 9.4 of Unit 9, test the following hypotheses, taking $\alpha = 0.05$:

i) $H_0: b = 8$ against $H_1: b \neq 8$

ii) $H_0: b = 8$ against $H_1: b > 8$

iii) $H_0: b = 8$ against $H_1: b < 8$

Solution: From the solution of Example 1, we have

$\bar{X} = 6.97$, $\bar{Y} = 14.7$, $V(\hat{a})$ 41.68, $V(\hat{b}) = 0.85$ and $\hat{\sigma}^2 = 5.48$;

From equation (12) of Unit 9, we get

$$\hat{b} = SS_{XY} / SS_X = 61.81 / 6.40 = 9.66$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} = 14.7 - 9.66 \times 6.97 = -52.63$$

i) $H_0: b = 8$ against $H_1: b \neq 8$

Here $b_0 = 8$ and from equation (5), we have

$$t = \frac{(\hat{b} - b)}{\sqrt{\hat{\sigma}^2 / SS_X}} = \frac{(9.66 - 8)}{\sqrt{5.48/6.40}}$$

$$= \frac{1.66}{0.925} = 1.79 \qquad \ldots (i)$$

The Statistics tables are provided in the last pages in most books on Statistics. However, these are also given in the Appendix in Block 3 of the course MST-004.

From the Statistics table, the value $t_{\alpha/2}$ at 8 d.f. is 2.31. Since the calculated value $|t| = 1.79 < 2.31$, we accept the null hypothesis, i.e., there is no evidence against the null hypothesis H_0.

ii) H_0: b = 8 against H_1: b > 8

We have calculated in (i) that t = 1.79. It is a right tail test and we reject the hypothesis H_0 if $t > t_\alpha$ at 8 d.f.

In this case, $t_{0.05}$ at 8 d.f. = 1.86 (from Statistics table). Since $t < t_\alpha$, i.e., $1.79 < 1.86$, we accept the null hypothesis and conclude that there is no evidence against the null hypothesis.

iii) H_0: b = 8; against H_1: b < 8

It is a left tail test and we reject the hypothesis when $t \le t_{(1-\alpha)}$ or $t \le -t_\alpha$.

From the Statistics table, $t_{0.95} = -t_{0.05} = -1.86$ at 8 d.f. and t = 1.79. Since $t > -t_{0.05}$, we accept H_0 and conclude that there is no evidence against the null hypothesis.

From all the three tests we conclude that the null hypothesis H_0: b = 8 cannot be rejected.

10.3 COEFFICIENT OF DETERMINATION

In the previous section, we have tested for the difference of coefficient b from b_0. If b is zero, it means that Y is not influenced by X at least in a linear fashion. If b is not zero, it may be of interest to explore how much variability in Y is accounted for by this linear relationship. In other words, how good the fit is. When the fit is good, most points must lie on the fitted line

$$\hat{Y} = \hat{a} + \hat{b}X$$

or very close to it, This means that the deviation of the fitted values from Y must be small for all the observations, i.e., the absolute value of the **residual**

$$r_i = Y_i - \hat{Y}_i$$

must be small for all observations. If this is so, then

$$\sum_{i=1}^{n} r_i^2$$

should be small. Since the values of r_i are affected by the unit in which observations are measured, we should have a measure of goodness of fit, which is not affected by the measurement unit. This is what we shall determine now.

The sum of squares of residuals denoted by SS_{Res} is obtained as follows:

$$SS_{Res} = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left[Y_i - \bar{Y} - \hat{b}\left(X_i - \bar{X}\right) \right]^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 - 2\hat{b} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) + \hat{b}^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

$$= SS_Y - 2\left(SS_{XY}\right)^2 / SS_X + \left(SS_{XY}\right)^2 / SS_X$$

or $$SS_{Res} = \left(SS_Y - \left(SS_{XY}\right)^2 / SS_X\right) \quad \ldots (6)$$

This is the sum of squares which involves variability due to random errors. It does not arise due to the regressor variable. However, if we subtract it from the total variability of Y around its mean, i.e, SS_Y, the sum of squares will have the contribution of the regressor variable in the variability, i.e.,

$$SS_Y - (SS_Y - \frac{\left(SS_{XY}\right)^2}{SS_X}) = \frac{\left(SS_{XY}\right)^2}{SS_X}$$

This is always positive and is known as **variability accounted for by X**. Now, this is not unit free. To make it unit free, we divide it by total variability in Y around mean SS_Y. The ratio of variability in Y accounted for by X to the total variability in Y around the mean is called the **coefficient of determination** and is denoted by R^2

$$R^2 = \frac{\left(SS_{XY}\right)^2}{SS_X \times SS_Y} \quad \ldots (7)$$

You may recognise that it is the square of the correlation coefficient between X and Y, i.e., r^2. Since r lies between −1 and 1, R^2 lies between 0 and 1. It is closer to 1 when the linear relationship between Y and X is very good. In this case most observations lie very close to the fitted line of $Y = \hat{a} + \hat{b}X$. When R^2 is closer to zero, the relationship between Y and X is very poor, and most observations lie far away from the fitted line. Hence, the value of R^2 gives a good measure of the fitness of the line. Let us take up an example to illustrate this concept.

Example 3: Using the data of Example 1, calculate the coefficient of determination, R^2, and comment on the goodness of fit of the regression line. Also calculate the correlation coefficient r.

Solution: For the data of Example 1, we have

$$SS_X = 6.40, \quad SS_{XY} = 61.81, \; SS_Y = 640.1$$

On putting these values in equation (7), we get

$$R^2 = \frac{\left(SS_{XY}\right)^2}{SS_X \times SS_Y} = \frac{\left(61.81\right)^2}{6.40 \times 640.1} = 0.933$$

Since R^2 is close to 1 and 93.3% variability in variable Y is due to X, the regression line is a good fit.

Now $r = \sqrt{R^2}$ and we have

$$r = \sqrt{0.933} = 0.966$$

Alternatively,

$$r = \frac{SS_{XY}}{\sqrt{SS_X \times SS_Y}} = \frac{61.81}{\sqrt{6.40 \times 640.1}} = 0.966$$

Adjusted R^2

The coefficient of determination, R^2, has a weakness. In equation (7), the denominator is fixed and the numerator increases if additional independent variables are included in the analysis. This is due to the contribution of the regressor variables to the variability. This results in a higher R^2, even when the new variable causes the equation to become less efficient (worse).

The **Adjusted R^2** value is an effort to correct this drawback by adjusting both the numerator and the denominator by their respective degrees of freedom. It is given by

$$\tilde{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k}\right) \quad \ldots (8)$$

where R^2 is the coefficient of determination, n, the number of observations and k, the number of independent variables.

Note that unlike R^2, the Adjusted R^2 contains the d.f. term. So even if the contribution of the additional variable is less, the value of $\tilde{R}^2$ may reduce. This implies that the equation with the smallest standard error of the estimate may have the highest Adjusted R^2.

Example 4: Determine Adjusted R^2 for Example 3.

Solution: From the solution of Example 3, we have

$R^2 = 0.933$, $n = 10$ and $k = 2$

Therefore, $\tilde{R}^2 = 1 - \left(1 - (0.933)^2\right)\left(\frac{9}{8}\right)$

$$\tilde{R}^2 = 1 - (0.13)(1.125) = 0.925$$

You may now like to solve the following exercises to check your understanding.

E1) What is the statistic t for testing a hypothesis? How is it used?

E2) Define the Coefficient of Determination.

E3) A study was conducted on the effect of temperature (X) on the yield of a chemical process (Y). The following data (in coded form) were collected:

X	–5	–4	–3	–2	–1	0	1	2	3	4	5
Y	–1	5	4	7	10	8	9	13	14	13	18

i) Assuming a model Y = a+ bX+ e, what are the least squares estimates of a and b ?

ii) Calculate the variances of the estimated regression coefficients.

iii) Test the hypothesis that the temperature (X) has no effect on the yield (Y) of the chemical process, ie., H_0: b = 0; H_1: b ≠ 0 at α = 0.05.

10.4 CONFIDENCE INTERVAL

In Sec. 10.2, we have derived the variances of the estimated coefficients. We have also discussed how to test their significance. If the regression line is a good fit, which is indicated by a high value of R^2, it may be of interest to predict the mean value of Y for some given value of $X = X_0$. It may be of interest to calculate its variance and confidence interval to get an idea of its accuracy.

Suppose a regression line $Y = a + b X + e$ is fitted and the fitted line is given by

$$\hat{Y} = \hat{a} + \hat{b} X$$

Since $E(\hat{a}) = a$, and $E(\hat{b}) = b$, we have

$$E(\hat{Y}) = a + b X = E(Y)$$

Thus, if we want to estimate E(Y) at a given value of X, say X_0, an unbiased estimate is given by

$$\begin{aligned}\hat{Y} &= \hat{a} + \hat{b}X_0 = \bar{Y} - \hat{b}\bar{X} + \hat{b}X_0 \\ &= \bar{Y} + \hat{b}(X_0 - \bar{X})\end{aligned}$$

$$\begin{aligned}V(\hat{Y}) &= V\left[\bar{Y} + \hat{b}(X_0 - \bar{X})\right] \\ &= V(\bar{Y}) + V(\hat{b})(X_0 - \bar{X})^2 + 2\,\mathrm{Cov}(\bar{Y}, \hat{b})(X_0 - \bar{X}) \\ &= \left[\frac{\sigma^2}{n} + \frac{\sigma^2 (X_0 - \bar{X})^2}{SS_X}\right] \qquad \ldots (9)\end{aligned}$$

since $\mathrm{Cov}(\bar{Y}, \hat{b}) = 0$. We note that $V(\hat{Y})$ is minimum when $X_0 = \bar{X}$ and is given by $\frac{\sigma^2}{n}$. It increases rapidly in proportion to $(X_0 - \bar{X})^2$. Equation (9) provides the variance of predicted value of Y, which depends on σ^2.

10.4.1 Confidence Interval for b

We have discussed the tests of hypotheses about regression coefficients a and b in Sec. 10.2.2. Once we have tested the t-statistic, we can construct the confidence intervals for b. From equation (5), we have

$$t = \frac{(\hat{b} - b)}{\sqrt{\hat{\sigma}^2 / SS_X}} \sim \text{Student's } t - \text{distribution on } (n-2) \text{ d.f.}$$

Since the variable t follows the Student's t-distribution, we may write

$$P\{|t| \le t_{\alpha/2}\} = 1-\alpha$$

or $$P\{-t_{\alpha/2} \le t \le t_{\alpha/2}\} = 1-\alpha$$

or $$P\left\{-t_{\alpha/2} \le \frac{(\hat{b}-b)}{\sqrt{\hat{\sigma}^2/SS_X}} \le t_{\alpha/2}\right\} = 1-\alpha$$

or $$P\left\{\hat{b} - t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}} \le b \le \hat{b} + t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}}\right\} = 1-\alpha \quad \ldots (10)$$

Let us denote the lower confidence limit and upper confidence limit by b_L and b_U.Then, we have

$$b_L = \hat{b} - t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}} \quad \text{and} \quad b_U = \hat{b} + t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}} \quad \ldots (11)$$

Hence, the probability that b lies between b_L and b_U is $(1-\alpha)$ and the interval (b_L, b_U) is known as $(1-\alpha)$% confidence interval. You must note that the parameter b is not a random variable, but the interval (b_L, b_U) is random. This means that in repeated sampling, the intervals will be different in each case and $100(1-\alpha)$% of interval will contain parameter b and 100 α% intervals will not contain parameter b. In one sample case, the interval (b_L, b_U) may contain b or may not contain b. However, it certainly tells us that if the interval is big, the estimate is not very accurate. If it is small, the estimate is better.

10.4.2 Confidence Interval for Predicted Value at $X = X_0$

We have to find the confidence interval for the mean value of Y at $X = X_0$, i.e.,

$$E(Y \mid X_0) = a + bX_0$$

We have seen that an unbiased estimate of $E(Y \mid X_0)$ is

$$\hat{Y} = \hat{a} + \hat{b}X_0$$

We have already seen that the variance of estimate of $E(Y \mid X_0)$ is

$$V(\hat{Y}) = \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)$$

As we do not know σ^2, we substitute $\hat{\sigma}^2$ for it. Thus, we get

$$V(\hat{Y}) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)$$

We have discussed in Block-3 of MST-004 that

$$t = \frac{\hat{Y} - E(Y)}{\sqrt{V(\hat{Y})}} \sim \text{Student's t - distribution on } (n-2)\text{d.f.}$$

Thus, we have

$$P\{|t| \le t_{\alpha/2}\} = 1-\alpha$$

or $$P\left\{-t_{\alpha/2} \le t \le t_{\alpha/2}\right\} = 1-\alpha$$

or $$P\left\{\hat{Y} - t_{\alpha/2}\sqrt{V(\hat{Y})} \le E(Y) \le \hat{Y} + t_{\alpha/2}\sqrt{V(\hat{Y})}\right\} = 1-\alpha$$

Let us denote the limits by $\hat{Y}_L$ and $\hat{Y}_U$, respectively, such that

$$\hat{Y}_L = \hat{Y} - t_{\alpha/2}\sqrt{V(\hat{Y})} = \hat{Y} - t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)}$$

$$\hat{Y}_U = \hat{Y} + t_{\alpha/2}\sqrt{V(\hat{Y})} = \hat{Y} + t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)} \quad \ldots (12)$$

Then the confidence interval for

$$E(Y|\, X = X_0) = a + bX_0$$

is given by $(\hat{Y}_L, \hat{Y}_U)$. Note that the interval is of minimum length at $X_0 = \bar{X}$ and increases rapidly at the rate of $(X_0 - \bar{X})^2$.

An important application of the regression analysis is to predict the new optimum value Y_0 corresponding to a specific value of the regressor variable X, i.e., $X = X_0$. The confidence interval on the mean response at $X = X_0$ given above is inappropriate for this prediction because it is an interval estimate based on the mean of Y. Therefore, the variance of the estimate of Y_0 becomes

$$V\left(\hat{Y}_0\right) = V\left(Y_0 - \hat{Y}_0\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)$$

as the variance of the error term is also to be included.

Then the confidence interval for a new predicted value

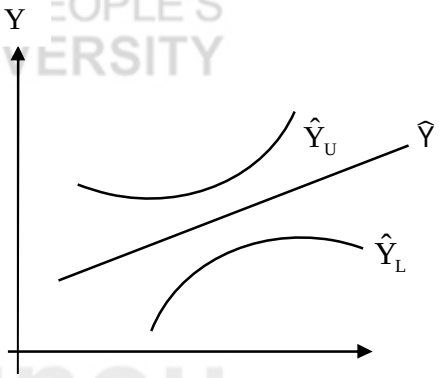
$$E(Y|\, X = X_0) = a + bX_0$$

is given by $(\hat{Y}_L, \hat{Y}_U)$ as follows:

$$\hat{Y}_L = \hat{Y}_0 - t_{\alpha/2}\sqrt{V(\hat{Y}_0)} = \hat{Y}_0 - t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)}$$

$$\hat{Y}_U = \hat{Y}_0 + t_{\alpha/2}\sqrt{V(\hat{Y}_0)} = \hat{Y}_0 + t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)}$$

Fig. 10.1 shows the predicted values of E (Y | X_0) and their confidence intervals.



X

Fig. 10.1: Predicted values of E (Y | X_0) and their confidence intervals.

This shows that the confidence interval is of minimum length when $X_0 = \overline{X}$ and it increases rapidly at a rate in proportion to $(X_0 - \overline{X})^2$. This shows that the prediction of E(Y) is most accurate at $X_0 = \overline{X}$ and the accuracy reduces rapidly as X_0 departs from $\overline{X}$.

Example 4: For the data given in E3, find the 95% confidence intervals for b and E (Y) at $X_0 = -5, -3, -1, 0, 1, 3, 5$.

Solution: From the solution of Example 3, we have

$$\hat{a} = 9.27, \quad \hat{b} = 1.44, \quad \hat{\sigma}^2 = 2.36, \quad \overline{X} = 0, \quad SS_X = 110$$

$$V(\hat{a}) = 0.236, \quad V(\hat{b}) = 0.021, \quad t_{0.025,9} = 2.262$$

Hence from equation (11), the lower confidence limit and upper confidence limit for b are given by

$$\hat{b}_L = \hat{b} - t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}} = 1.44 - 2.262 \times \sqrt{0.021} = 1.112$$

$$\hat{b}_U = \hat{b} + t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}} = 1.44 + 2.262 \times \sqrt{0.021} = 1.768$$

Therefore, the fitted straight line equation is given by

$$\hat{Y} = 9.27 + 1.44 X_0$$

Thus, from equation (12), the upper and lower confidence limits for E(Y) are given by

$$\hat{Y}_L = 9.27 + 1.44X_0 - 2.262\sqrt{2.36\left\{\frac{1}{11} + \frac{(X_0 - \overline{X})^2}{SS_X}\right\}}$$

$$\hat{Y}_U = 9.27 + 1.44X_0 + 2.262\sqrt{2.36\left\{\frac{1}{11} + \frac{(X_0 - \overline{X})^2}{SS_X}\right\}}$$

The following table gives the predicted values of E (Y | X_0) and their confidence intervals:

S. No.	X_0	$\hat{Y}$	$\hat{Y}_L$	$\hat{Y}_U$
1	−5	2.07	0.41	3.73
2	−3	5.07	3.63	6.51
3	−1	7.83	6.74	8.92
4	0	9.27	8.22	10.32
5	1	10.71	9.62	11.80
6	3	13.59	12.15	15.03

7	5	16.47	14.81	18.13

You may now like to solve the following problems to assess your understanding.

E4) A firm wants to know whether there is any linear relationship between the size of its sales force (X) and its yearly sales revenue (Y). The records of 10 years were examined and the data are listed in the following table:

Year	**1980**	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**
X	15	18	24	22	25	29	30	32	35	38
Y	1.35	1.62	2.33	2.41	2.63	2.93	3.41	3.26	3.63	4.15

where X and Y are the number of salespersons and sales in hundred thousand rupees, respectively. The following results are given:

$\Sigma X = 265$, $\Sigma Y = 27.73$, $SS_X = 485.6$, $SS_Y = 6.978$, $SS_{XY} = 57.456$

i) Fit a regression line taking Y as the dependent variable and X as the independent variable.

ii) Test whether the sales force has any effect on Y, by testing the significance of regression coefficient b, at 5% significance level.

iii) Find the coefficient of determination R^2 and comment on the goodness of fit of the regression line.

iv) Estimate the sales value (Y) for the year 1990 for which X = 40.

v) Find the 95% confidence interval for b and also for the estimated value of E(Y|1990).

E5) Derive $(1-\alpha)$% confidence interval for the regression coefficient a.

E6) The following table gives the data concerning the monthly usage of electricity Y (in KWA) and the size of the home X (in square feet) for a sample of ten homes:

Y:	11820	11720	12640	14930	15710	17110	18040	18400	19560	19540
X:	12900	13500	14700	16000	17100	18400	19800	22300	24000	29300

i) Fit the regression line.

ii) Test whether the monthly usage of electricity is influenced by the size of the home.

iii) Find out what percentage of variability in Y is accounted for by variation in X, by calculating the coefficient of determination R^2.

E7) An economist suspects that a firm's profitability, as measured by the percentage rate of return on investment, is largely a function of its production efficiency. She has gathered information on the returns and the production efficiency index (expressed in percentage; the higher its value, the more efficient the firm) for 8 local companies. The sample data are shown in the following table:

Firm	1	2	3	4	5	6	7	8
Profitability (Y)	12.3	25.8	23.3	16.9	15.4	6.9	8.9	10.2
Efficiency Index (X)	40.8	53.2	60.1	38.4	33.7	20.8	22.9	20.7

i) Calculate a and b and find the regression equation for the data.

ii) Comment on the goodness of fit of the regression line after calculating the coefficient of determination R^2.

iii) Does the regression line seem to provide a reasonable fit?

We now summarise the concepts that we have discussed in this unit.

10.5 SUMMARY

1. The parameters $\hat{a}$ and $\hat{b}$ are unbiased estimates of regression parameters a and b, respectively, and their variances are given as

$$V(\hat{a}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right), \qquad V(\hat{b}) = \frac{s^2}{SS_X}$$

where $\hat{a}$ and $\hat{b}$ are linear functions of Y_i .

2. A variable, which is a linear combination of normally distributed random variables, also follows the normal distribution with mean and variance of the linear combination. Thus, $\hat{a}$ and $\hat{b}$ are normally distributed with mean and variance as given above.

3. In simple linear regression model, we have two parameters a and b. Thus, k = 2, and an unbiased estimate of σ^2 is given by

$$SS_{Res} = \sum(r_i - \bar{r})^2/(n-2) = \sum r_i^2/(n-2)$$

4. Since population values of a and b are not known, we may hypothesise that the sample has come from a population whose true values are $a = a_0$ and $b = b_0$. On the basis of a sample (Y_i, X_i), i = 1, 2, …., n, we test whether these hypotheses are acceptable or not. This is particularly important when we are interested in knowing whether Y and X are linearly related or not.

5. The ratio of variability in Y accounted for by X to the total variability in Y around the mean is called the **coefficient of determination** and is denoted by R^2

$$R^2 = \frac{(SS_{XY})^2}{SS_X \times SS_Y}$$

The coefficient of determination, R^2, gives a measure of the goodness of fit of the regression line.

6. Confidence interval for regression coefficient b is

$$b_L = \hat{b} - t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}}$$

and $$b_U = \hat{b} + t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}}$$

and of predicted value of $E(Y|X=X_0)$ is

$$\hat{Y}_L = \hat{Y} - t_{\alpha/2}\sqrt{V(\hat{Y})} = \hat{Y} - t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)}$$

$$\hat{Y}_u = \hat{Y} + t_{\alpha/2}\sqrt{V(\hat{Y})} = \hat{Y} + t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_X}\right)}$$

7. The parameter b is not a random variable, but the interval (b_L, b_U) is random. This means that in repeated sampling, the intervals will be different in each case and $100(1-\alpha)\%$ of interval will contain parameter b and $100\,\alpha\%$ intervals will not contain parameter b. In one sample case, the interval (b_L, b_U) may contain b or may not contain b.

10.6 SOLUTIONS/ANSWERS

E1) Refer to Section 10.2.

E2) Refer to Section 10.3.

E3) From the given data, we have the following:

$$\sum X = 0, \quad \sum X^2 = 110 \qquad \sum Y = 102, \quad \sum Y^2 = 1194$$

$$\sum XY = 158 \quad SS_X = 110, \quad SS_Y = 248.18, \qquad SS_{XY} = 158$$

i) Therefore, from equation (1a) for $\hat{a}$ and $\hat{b}$, we get

$$\hat{b} = SS_{XY}/SS_X = 1.44, \text{ and } \quad \hat{a} = \bar{Y} - \hat{b}\bar{X} = 9.27 - 0 = 9.27$$

ii) On putting the values of SS_X, SS_Y and SS_{XY} in equation (1b), we get

$$SS_{Res} = \left(SS_Y - (SS_{XY})^2/SS_X\right)/(n-2)$$

$$= \left\{248.18 - (158)^2/110\right\}/9 = 21.23/9 = 2.36$$

Thus $$V(\hat{a}) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right) = 2.36\left[\frac{1}{11} + 0\right] = 0.236$$

and $$V(\hat{b}) = \frac{\hat{\sigma}^2}{SS_X} = 2.36/110 = 0.021$$

iii) To test the null hypothesis H_0: b = 0; against H_1: $b \neq 0$

we have $$t = (\hat{b} - 0)/\sqrt{\frac{\hat{\sigma}^2}{SS_X}} = 1.44/\sqrt{0.021} = 9.93$$

It is a two-tailed test, hence from table $t_{0.025} = 2.262$ at 9 d.f. Therefore, the calculated value is $|t| > 2.262$. Thus, we reject the null hypothesis and conclude that the temperature affects the yield of chemical process.

E4) From the given data, we get the following results:

$\sum X = 268, \sum Y = 27.73$, $SS_X = 485.6$, $SS_Y = 6.978$, $SS_{XY} = 57.456$ and n = 10

i) On putting these values in equation (1a) we get

$$\hat{b} = SS_{XY} / SS_X = 0.118,$$

$$SS_{Res} = (SS_Y - (SS_{XY})^2 / SS_X) / (n - 2) = 0.022$$

and $$\hat{a} = \overline{Y} - \hat{b}\overline{X} = 2.773 - 0.118 \times 26.8 = -0.389$$

Thus, the fitted line is

$$\hat{Y} = -0.389 + 0.118\,X$$

ii) We have to test

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

From equation (5)

$$t = \frac{\hat{b} - 0}{\sqrt{\hat{\sigma}^2 / SS_X}} = \frac{0.118}{\sqrt{0.022 / 485.6}}$$

$$= \frac{0.118}{\sqrt{0.00045}} = \frac{0.118}{0.021} = 5.62$$

where the tabulated value of $t_{0.025}$ on 8 d.f. is 2.306.

Since $|t| > 2.306$, the null hypothesis is rejected. This means that the sales force X has a significant effect on sales revenue Y.

iii) Correlation Coefficient r

$$r = \frac{SS_{XY}}{\sqrt{SS_X \times SS_Y}} = \frac{57.456}{\sqrt{485.6 \times 6.978}}$$

$$= \frac{57.456}{58.210} = 0.987$$

Hence the coefficient of determination, $R^2 = r^2 = 0.974$. Thus 97% of the variability is accounted for by X. Hence, the regression model gives a very good fit.

iv) The value of estimated sales for the year 1990 is obtained by substituting X = 40 in the formula for $\hat{Y}$

$$\hat{Y} = \hat{a} + \hat{b}X = -0.389 + 0.118 \times 40 = 4.33$$

v) On putting the values in the equations

$$\hat{b}_L = \hat{b} - t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}} \text{ and } \hat{b}_U = \hat{b} + t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{SS_X}}$$

We have

$$\hat{b}_U = 0.118 + 2.306 \times 0.0067 = 0.133$$

and $$\hat{b}_L = 0.118 - 2.306 \times 0.0067 = 0.102$$

where tabulated value of $t_{0.025}$ on 8 d.f. is 2.306.

E5) Test statistic for testing hypothesis

H_0: $a = a_0$ against H_1: $a \neq a_0$

given in equation (3) as

$$t = \frac{(\hat{a} - a_0)}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{X}^2}{SS_X})}}$$

has Students-t distribution on (n−2) d.f. We replace a_0 by a and construct confidence intervals as before, using

$$P\{|t| \leq t_{\alpha/2}\} = 1-\alpha$$

giving (1−α)% confidence limits. Thus,

$$\hat{a}_L = \hat{a} - t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(\bar{X})^2}{SS_X}\right)}$$

and $$\hat{a}_U = \hat{a} + t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(\bar{X})^2}{SS_X}\right)}$$

E6) We have calculated the following values from the given data:

ΣX=1,88,000; ΣY=1,59,470;

SS_X=2,41,14,000; SS_Y=84,64,020; SS_{XY}=1,30,28,900

i) On putting the above values in equation (1a), we get

$$\hat{b} = SS_{XY}/SS_X = 0.5403, \text{ and } \hat{a} = \bar{Y} - \hat{b}\bar{X} = 5789.36$$

The fitted line is

$$\hat{Y} = 5789.4 + 0.5403X$$

$$SS_{Res} = \left(SS_Y - (SS_{XY})^2/SS_X\right)/8 = 178056.02$$

ii) Here we have to test the hypothesis

H_0: b = 0 against H_1: b ≠0

We consider the test statistic

$$t = \frac{\hat{b} - 0}{\sqrt{\hat{\sigma}^2/SS_X}} = \frac{0.5403}{0.0859} = 6.289$$

where $|t| > t_{0.025}$ at 8 d.f. = 2.306.

Hence, the size of the house affects the electricity use considerably.

iii) Similarly, from equation (4), we have

$$r = \frac{SS_{XY}}{\sqrt{SS_X \times SS_Y}} = \frac{13028900}{14286405.3} = 0.9119$$

Therefore, the coefficient of determination, $R^2 = r^2 = 0.83156$.

Hence, 83.156% of variability in Y is accounted for by X.

E7) We have calculated the following values from the given data:

$\Sigma Y = 119.7$; $\Sigma X = 290.6$; $\Sigma X^2 = 12102.68$; $\Sigma Y^2 = 2113.45$,

$\Sigma XY = 5001.14$; $SS_X = 1546.64$; $SS_Y = 322.439$, $SS_{XY} = 653.037$,

$\hat{\sigma}^2 = 7.785$.

Therefore, we have

i) $\hat{b} = SS_{XY}/SS_X = 0.422$ and $\hat{a} = \overline{Y} - \hat{b}\overline{X} = -0.367$

$$\therefore \hat{Y} = -0.367 + 0.422X$$

ii) Similarly, from equation (4), we have

$$r = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}} = \frac{653.037}{706.185} = 0.925$$

iii) Coefficient of determination, $R^2 = r^2 = 0.855$

Thus 85.5% of variability in Y is accounted for by X. Hence, the regression line is a good fit.