# UNIT 6 STATISTICAL INFERENCE IN SIMPLE LINEAR REGRESSION

## Structure



## 6.1 INTRODUCTION

In Unit 5, you have learnt how to fit a simple linear regression model using the method of least squares. As we have discussed in “MST-002: Descriptive Statistics”, descriptive statistic(s) such as mean, variance, correlation coefficient and regression coefficient calculated from a given data, in fact, provide sample specific values as the data in hand is mostly a sample selected from a population. Hence, these do not convey much information about the population. In most studies, our aim is to draw inference about the population characteristics on the basis of information available from the sample, as you have learnt in Block 3 of “MST-004: Statistical Inference”. In fact, the dependence or independence between two (or more) variables in the sample does not necessarily imply the same for the population as well. For example, if the value of sample regression coefficient $(\hat{\beta}_1)$ is not zero, it does not necessarily imply that the population regression coefficient $(\beta_1)$ will also not be zero. In such a situation, we may have a population of several data points where X and Y are independent. In that case, the line of best fit would be parallel to the X-axis implying that there is no slope, i.e., $\beta_1 = 0$. However, a random sample selected from the same population may not have zero slope. In such cases, it is desirable to explore the inferential aspects for analysing regression coefficients.

This unit deals with the inferential aspects of simple linear regression. In Sec. 6.2, we explain how to draw the residual and normal probability plots to check the adequacy of the fitted regression model. We consider testing the significance of the individual regression coefficients, the intercept and the slope in Sec. 6.3 whereas in Sec. 6.4, we discuss the significance of the overall fitted simple regression model. In Sec. 6.5, we describe computation of the confidence intervals of intercept and slope which contain the value of that coefficient at $(1-\alpha)100\%$ confidence level. We discuss the coefficient of determination in Sec. 6.6.

## Objectives

After studying this unit, you should be able to:

- construct the residual and normal probability plots;
- check the adequacy of the fitted model with the help of residual analysis;
- conduct hypotheses testing of significance of the intercept and the slope;
- test significance of the overall fitted simple linear regression model;
- determine interval estimation of the intercept and the slope; and
- compute and interpret the coefficient of determination.

## 6.2 RESIDUAL ANALYSIS

In Sec. 5.2, we have discussed some assumptions of the simple and multiple linear regression models. Recall these assumptions from Sec. 5.2.1, wherein we described the linearity, homoscedasticity and normality assumptions required for regression models.

Ideally, the sum of residuals for all given observations should always be equal to zero, i.e., $\sum_{i=1}^{n} r_i = 0$. But sometime, rounding errors may change the result when we approximate the decimal places.

It is necessary to check the validity of these assumptions for applying a valid regression analysis. In this section, you will learn the residual analysis technique for checking the validity of some basic assumptions to ensure the adequacy of the regression model.

Residuals play an important role in investigating the accuracy of the fitted regression model and in detecting departures from the underlying assumptions of the regression model.

You have learnt in Sec. 5.3 of Unit 5 that the residual is a difference between the observed value and the corresponding predicted value of the response variable. Let $r_i$ be the $i^{th}$ residual defined as:

$$r_i = y_i - \hat{y}_i; \qquad i = 1, 2, ..., n \qquad \ldots (1)$$

From the properties of residuals, you know that it has zero mean and variance $\sigma^2$. The value of $\sigma^2$ is unknown and we estimate it using equation (25) of Unit 5 as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \qquad \ldots (2)$$

When we have outliers or extreme values in the data, it is useful to determine the scaled residuals. We have different methods for computing scaled residuals. Here, we shall discuss only the standardised residuals as the remaining methods are beyond the scope of this course.

We can compute the $i^{th}$ standardised residual as:

$$s_i = \frac{r_i}{\hat{\sigma}}; \qquad i = 1, 2, ..., n \qquad \ldots (3)$$

The standardised residuals have zero mean and their variance is approximately equal to one. We can determine the values of outliers with the help of standardised residuals. Generally, we consider the values of standardised residuals greater than three, i.e., $|s_i| > 3$, say, as an indication of an outlier.

It is always desirable to check the behaviour of residuals. Residuals possibly explain any irregularity of the fitted regression model that might have occurred and misled us.

Let us consider the following example to explain this point.

**Example 1:** For the data on systolic blood pressure and age given in Example 1 of Unit 5,

i) obtain the standardised residuals, and

ii) determine the outlier through standardised residuals.

**Solution:** We have fitted the regression model for the given data in Example 2 of Unit 5. We have obtained the best fitted regression model using the method of least squares as:

$$\hat{Y} = 90.0427 + 1.1189\,X$$

We have determined the predicted values of the response variable and the values of residuals in Columns 6 and 7 of Table 3 in Unit 5, respectively.

From the solution of Example 3 of Unit 5, we have

$$\hat{\sigma} = \sqrt{2.5258} = 1.5893$$

It is to be noted that all calculations were performed up to 15 fixed decimal places for showing accurate results in this block. For the sake of simplicity, we are showing results up to 4 decimal places only. The results may vary if we carry out the calculations by fixing values at various decimal places.

We now compute the standardised residuals using equation (3) for all observations given in the data and arrange them in Column (6) of Table 1 as follows:

**Table 1: Computation of Standardised Residuals k**

| S. No. | $y_i$ | Predicted Values $(\hat{y}_i)$ | Residuals $r_i = (y_i - \hat{y}_i)$ | $r_i^2 = (y_i - \hat{y}_i)^2$ | Standardised Residuals $s_i = \frac{r_i}{\hat{\sigma}}$ | $(y_i - \bar{y})$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| 1 | 124 | 123.6098 | 0.3902 | 0.1523 | 0.2455 | – 1.4 | 1.96 |
| 2 | 134 | 132.5610 | 1.4390 | 2.0708 | 0.9055 | 8.6 | 73.96 |
| 3 | 135 | 133.6799 | 1.3201 | 1.7427 | 0.8306 | 9.6 | 92.16 |
| 4 | 121 | 119.1341 | 1.8659 | 3.4814 | 1.1740 | – 4.4 | 19.36 |
| 5 | 122 | 122.4909 | – 0.4909 | 0.2409 | – 0.3089 | – 3.4 | 11.56 |
| 6 | 119 | 120.2530 | – 1.2530 | 1.5701 | – 0.7884 | – 6.4 | 40.96 |
| 7 | 128 | 125.8476 | 2.1524 | 4.6330 | 1.3544 | 2.6 | 6.76 |
| 8 | 118 | 118.0152 | – 0.0152 | 0.0002 | – 0.0096 | – 7.4 | 54.76 |
| 9 | 120 | 119.1341 | 0.8659 | 0.7497 | 0.5448 | – 5.4 | 29.16 |
| 10 | 123 | 124.7287 | – 1.7287 | 2.9883 | – 1.0877 | – 2.4 | 5.76 |
| 11 | 129 | 131.4421 | – 2.4421 | 5.9637 | – 1.5366 | 3.6 | 12.96 |
| 12 | 117 | 118.0152 | – 1.0152 | 1.0307 | – 0.6388 | – 8.4 | 70.56 |
| 13 | 131 | 129.2043 | 1.7957 | 3.2247 | 1.1299 | 5.6 | 31.36 |
| 14 | 126 | 128.0854 | – 2.0854 | 4.3488 | – 1.3121 | 0.6 | 0.36 |
| 15 | 134 | 134.7988 | – 0.7988 | 0.6381 | – 0.5026 | 8.6 | 73.96 |
| Total | | | | 32.8354 | | | 525.60 |

Since the values of standardised residuals for all observations are less than 3, there is no indication of outlier in the given data.

After we have obtained the residuals and standardised residuals, we should be able to decide whether some underlying assumptions hold or not. We use the graphical method, which is helpful in checking the validity of these assumptions. Let us check how well the regression model fits the given data. For this purpose, we draw the residual and normal probability plots. We shall discuss these plots in the following sub-sections, one at a time.

### 6.2.1 Residual Plot

In a **residual plot**, we take the residuals or standardised residuals on the vertical (Y) axis and the predicted values of the response variable or the regressor variable, X in case of simple regression and $X_i$'s $(i = 1, 2, ...k)$ in case of multiple regression on the horizontal (X) axis. If the points in a residual plot are randomly dispersed around the horizontal axis, we say that a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate. Some common types of residuals plots considering residuals versus predicted variable $(\hat{Y})$ are as follows:

The residual plot can be used to detect non-linearity and/or unequal variance. A normal plot of the residuals can be used to detect non-normality.

1. A satisfactory residual plot should be more or less within the horizontal band (random pattern) of points as shown in Fig. 6.1.

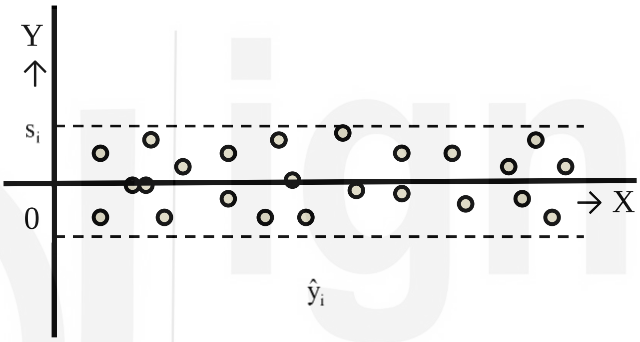



**Fig. 6.1: Horizontal band.**

2. When we have the heteroscedastic data (non-constant variance), we obtain the residual plot shown in Figs. 6.2 (a) and (b) show an outward or inward opening funnel pattern, respectively. This indicates that the variance of the error terms is not constant but the error variance increases or decreases with the response variable. These types of models may have very poor ability to account for the variability.

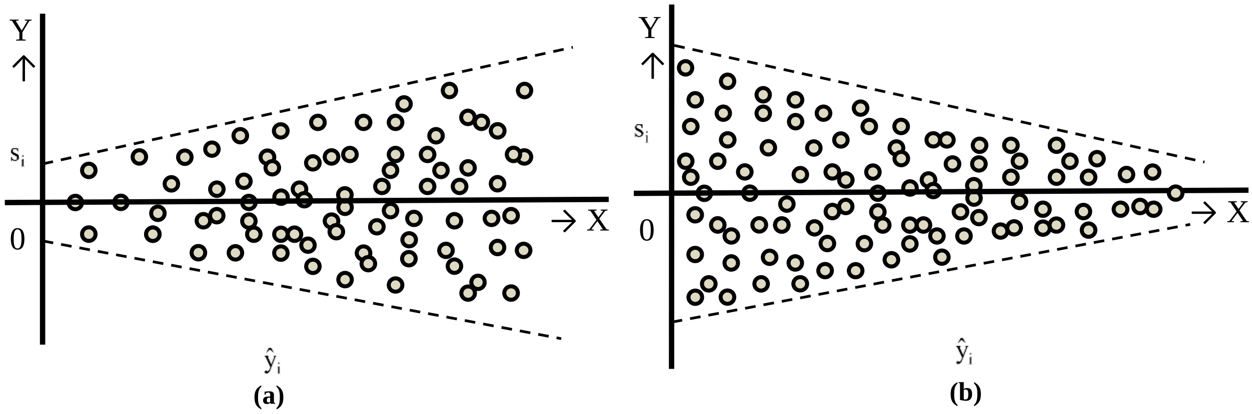
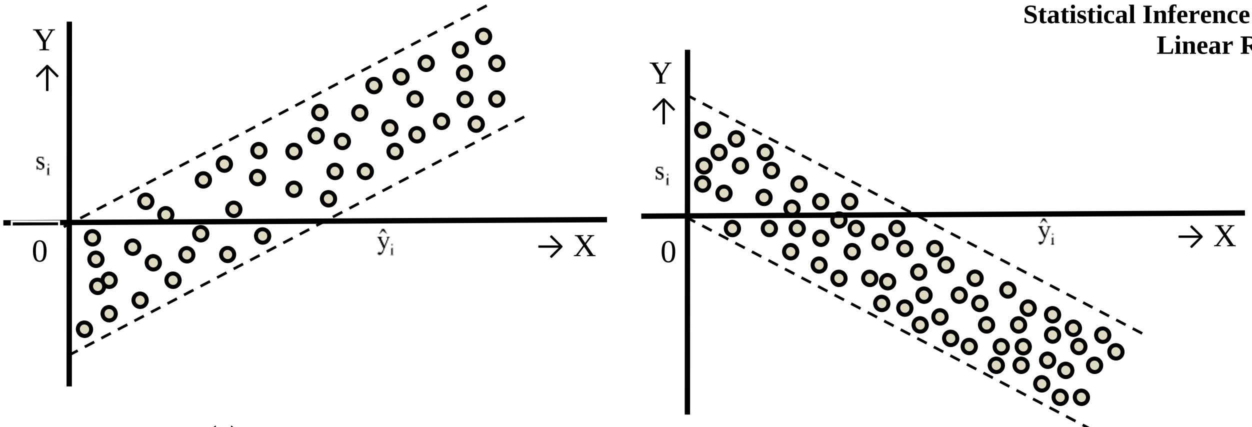



**Fig 6.2: (a) Outward funnel shape; (b) Inward funnel shape.**

3. A trend may be exhibited in the residual plot as shown in Fig. 6.3 for which there is likely to be an error in the computation of regression model or some additional regression analysis may be needed in the model formulation.




**(a)** **(b)**

**Fig. 6.3: (a) Upward trend shape; (b) Downward trend shape.**

4. If the relationship between Y and X is non-linear, the pattern of the residuals shown in Fig. 6.4 will be observed. In this situation, a curvilinear relationship is suggested to fit an appropriate regression model.

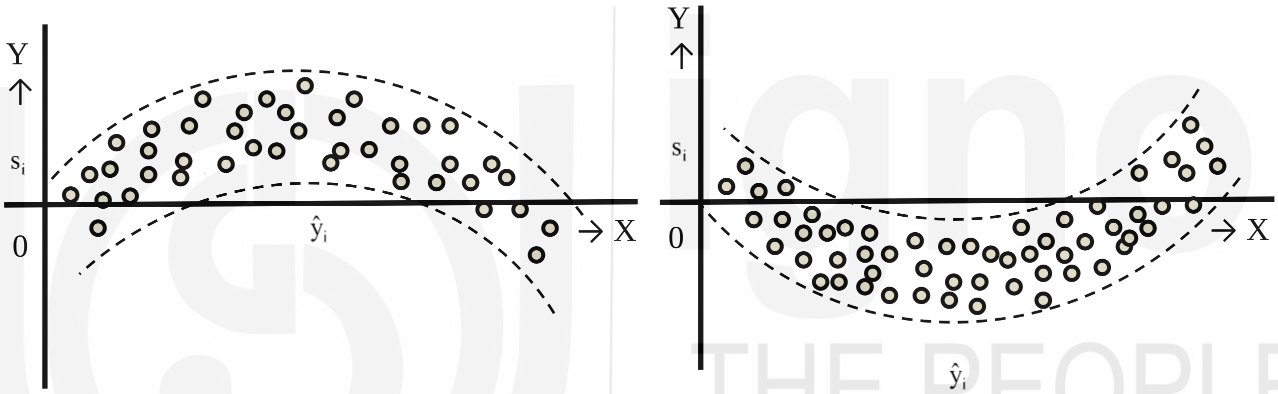



**Fig. 6.4: Non-linear patterns.**

5. If the residuals indicate a pattern which is accommodated inside a double bow as shown in Fig. 6.5, the variance of error terms is considered as non-constant. In this situation, the response variable (Y) may be a proportion between 0 and 1.

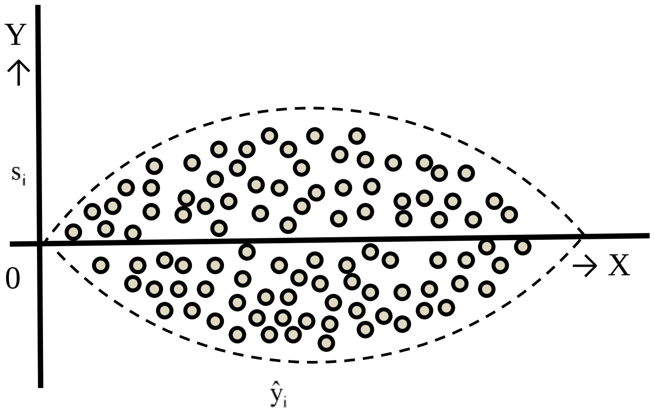



**Fig. 6.5: Double bow pattern.**

We can use transformation either to the regressor variable (X) or to the response variable (Y) for dealing with such inequalities of variance or we can also use weighted least squares method to deal with such situations. We are not going to discuss them as these are beyond the scope of this course.

Note that by plotting standardised residuals versus predicted values in residual plots, we also use them for identifying large values of the residuals which are known as outliers.

You may like to solve the following example for a better understanding of the concept of a residual plot.

**Example 2:** Construct the residual plot for the data on systolic blood pressure given in Example 1 and interpret it.

**Solution:** To obtain a residual plot, we consider the predicted values $(\hat{y}_i)$ and standardised residuals ($s_i$) on the horizontal and vertical axes, respectively. We now mark the points corresponding to the standardised residuals against the fitted values of Y obtained in Table 1. In this way, we obtain the residual plot shown in Fig. 6.6.

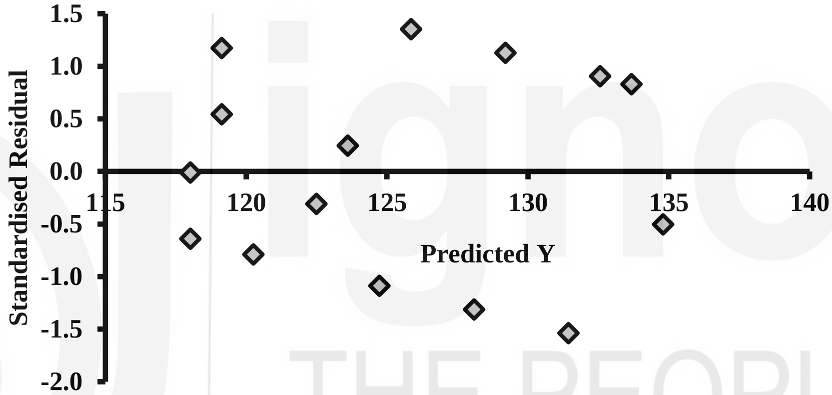



**Fig: 6.6: Residual plot.**

The standardised residuals shown in Fig. 6.6 appear to be approximately randomly scattered throughout a horizontal band around 0.0 on Y-axis. Hence, the assumption of linear regression is valid or we can say that the simple linear regression model fits the given data well. Fig. 6.6 does not show any irregularity in variance of the error terms and indication of possible outliers.

Now that you are familiar with the residual plot, we discuss the normal probability plot used to check the normality assumption.

## 6.2.2 Normal Probability Plot

You have learnt in "MST-004: Statistical Inference" that most statistical tests rely upon certain assumptions about the variables, which are used in the analysis. When these assumptions are not satisfied, the results may not be reliable. Of these, the assumption of normality is the most important since the t-test, F-test and confidence intervals depend on it. So we construct normal probability plot to check the normality assumption. The steps involved in the construction of a normal probability plot are as follows:

1. We arrange the standardised residuals computed using equation (3) in increasing order. Let $s_{(1)}, s_{(2)}, s_{(3)}, \ldots, s_{(n)}$ be the standardised residuals in increasing order or ordered standardised residuals.

2. We assign ranks to the ordered standardised residuals. Let $R_1$, $R_2$, $R_3$, …, $R_n$ be the ranks assigned to the ordered standardised residuals $s_{(1)}$, $s_{(2)}$, $s_{(3)}$, …, $s_{(n)}$, respectively.

3. After ranking the standardised residuals, we compute the cumulative probability ($p_i$) for each corresponding ranked standardised residual:

$$p_i = \frac{(R_i - 0.5)}{n}; \qquad i=1, 2, 3, \ldots, n \qquad \ldots (4)$$

4. We can also express the cumulative probability in terms of percentage by multiplying $p_i$ with 100. It is known as percentile cumulative probability ($P_i$) or percentiles. Thus,

$$P_i = p_i \times 100 = \frac{(R_i - 0.5)}{n} \times 100; \qquad i=1, 2, 3, \ldots, n \qquad \ldots (5)$$

5. Then we plot either cumulative probability ($p_i$) or percentile cumulative probabilities ($P_i$) on the vertical (Y) axis and the standardised residuals on the horizontal (X) axis to obtain the normal probability plot.

6. We can also take the expected normal values instead of the cumulative probabilities on the vertical axis. The expected normal value can be computed as follows:

$$z_i = \Phi^{-1}(p_i) = \Phi^{-1}\left[\frac{(i-0.5)}{n}\right] \qquad \ldots (6)$$

where $\Phi$ denotes the standard normal cumulative distribution.

The resulting points should lie approximately along a straight line if it satisfies the assumption of normality. In the normal probability plot, we determine the straight line visually which passes through the central points of all values rather than the extreme points at both ends. While dealing with normal probability plot, you can come across different patterns of the plot. We consider and interpret some of the patterns as follows:

1. If all points lie approximately along a straight line as shown in Fig. 6.7, the normal probability plot is called ideal and it will satisfy the normality assumption. A substantial amount of departure from the straight line indicates non-normality.

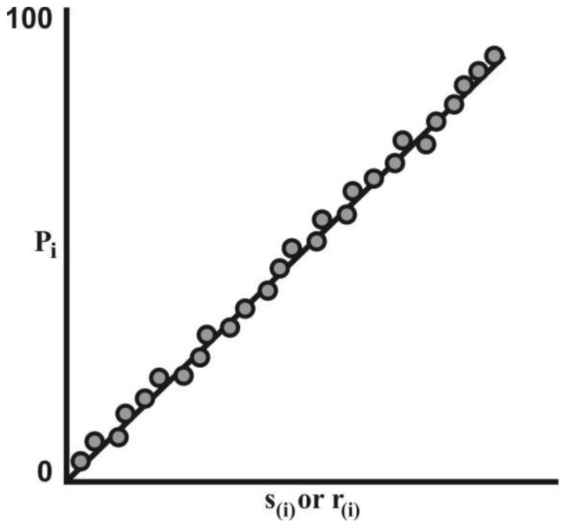
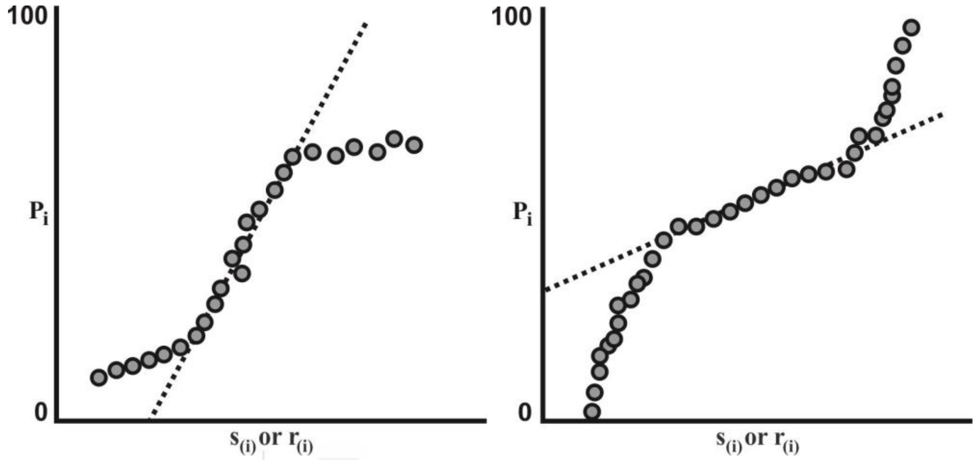



**Fig. 6.7**

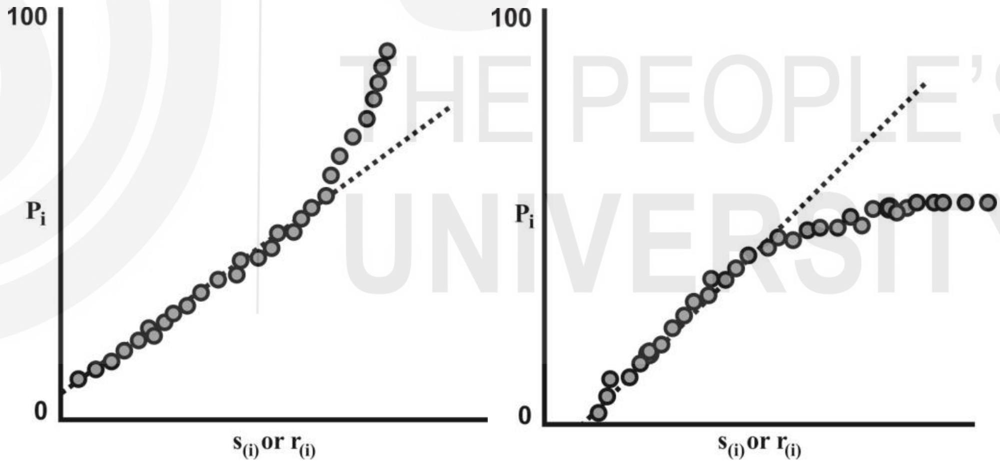
2. The normal probability plots shown in Figs. 6.8 (a) and (b) do not lie along a straight line. This indicates that there is some problem with the normality assumption. A sharp upward and downward curve at both ends of the plot indicates that the end points of this distribution have heavier tails than the normal probability plot shown in Fig. 6.7. The plots shown in Fig. 6.8 cannot be considered as normal.




**(a)** **(b)**

**Fig. 6.8**

3. In Figs 6.9 (a) and (b), the normal probability plots show an upward and downward trend, which indicate positive and negative skewness at the ends of the curves, respectively.




**(a)** **(b)**

**Fig. 6.9**

Let us now solve an example to explain the construction of normal probability plot.

**Example 3:** Construct the normal probability plot for the data on systolic blood pressure and age considered in Example 1.

**Solution:** In Example 2, you have already computed the residuals and standardised residuals for the data given in Example 1 of Unit 5. To draw the normal probability plot, we compute the ordered standardised residuals and percentiles. We arrange the standardised residuals computed in Column 6 of Table 1 in increasing order as shown in Column 2 of Table 2. You should note that we can also assign rank to the standardised residuals themselves, without arranging them in increasing order. After arranging the standardised residuals, we calculate the cumulative probabilities and percentile cumulative probabilities using equations (4) and (5) for the corresponding ordered standardised residuals and enter them in Columns 4 and 5 of Table 2, respectively.

**Table 2: Computation of Percentile Cumulative Probability**

| S. No. | Ordered Standardised Residual ($s_{(i)}$) | Ranks ($R_i$) | Cumulative Probability ($p_i$) | Percentile Cumulative Probability ($P_i$) |
|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** |
| 1 | – 1.5366 | 1 | 0.0333 | 3.3333 |
| 2 | – 1.3121 | 2 | 0.1000 | 10.0000 |
| 3 | – 1.0877 | 3 | 0.1667 | 16.6667 |
| 4 | – 0.7884 | 4 | 0.2333 | 23.3333 |
| 5 | – 0.6388 | 5 | 0.3000 | 30.0000 |
| 6 | – 0.5026 | 6 | 0.3667 | 36.6667 |
| 7 | – 0.3089 | 7 | 0.4333 | 43.3333 |
| 8 | – 0.0096 | 8 | 0.5000 | 50.0000 |
| 9 | 0.2455 | 9 | 0.5667 | 56.6667 |
| 10 | 0.5448 | 10 | 0.6333 | 63.3333 |
| 11 | 0.8306 | 11 | 0.7000 | 70.0000 |
| 12 | 0.9055 | 12 | 0.7667 | 76.6667 |
| 13 | 1.1299 | 13 | 0.8333 | 83.3333 |
| 14 | 1.1740 | 14 | 0.9000 | 90.0000 |
| 15 | 1.3544 | 15 | 0.9667 | 96.6667 |

Then we plot the ordered or ranked standardised residuals against the percentiles and obtain a normal probability plot as shown in Fig. 6.10.

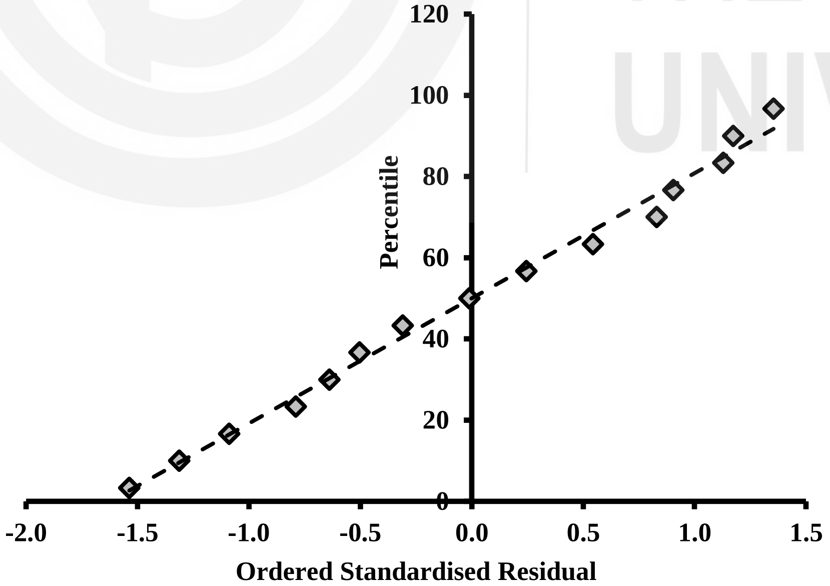



**Fig. 6.10: Normal Probability Plot.**

Note that the resulting points would lie approximately along a straight line as shown in Fig. 6.10. Notice from Fig. 6.10 that some points of the distribution deviate slightly from the straight line, but do not lie very far from the central points. It indicates that the distribution of error terms (residuals) is approximately normally distributed.

You may like to pause here and check your understanding about residual analysis by answering the following exercises.

**E1)** For the exercise given in **E4** of Unit 5,

(i) determine the standardised residuals, and

(ii) construct the residual and normal probability plots, and interpret the results.

**E2)** For the exercise given in **E5** of Unit 5,

(i) obtain the standardised residuals,

(ii) construct the residual and normal probability plots, and interpret the results.

## 6.3 TEST OF SIGNIFICANCE OF INDIVIDUAL REGRESSION COEFFICIENTS

In Sec. 5.3 of Unit 5, you have learnt how to estimate the parameters of a simple linear regression model, i.e., $\beta_0$ and $\beta_1$ using the method of least squares. We have denoted the least squares estimators of regression coefficients, $\beta_0$ and $\beta_1$ by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. We now discuss the procedure for testing significance of the estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$. Since we estimate $\beta_0$ and $\beta_1$ with the help of a given random sample using the method of least squares, if we consider other samples, we shall get the different estimates of $\beta_0$ and $\beta_1$. So we use the usual hypothesis testing procedure to draw inference about the population parameters using the sampled data. In such a situation, we test null hypothesis that the sample has come from the population for which the values of $\beta_0$ and $\beta_1$ are equal to some specific known values, say, $\beta_0^*$ and $\beta_1^*$, respectively.

We now discuss the procedure of hypothesis testing for intercept and slope which is the same as the standard procedure of hypothesis testing you have learnt in MST-004.

### 6.3.1 Hypothesis Testing for the Intercept

For testing the hypothesis that the sample comes from the population for which the value of $\beta_0$ is equal to $\beta_0^*$, we follow the steps given below.

**Step 1: Set up the hypotheses**

We specify the null and alternative hypotheses as:

$$\left.\begin{array}{ll}\text{Null hypothesis} & H_0 : \beta_0 = \beta_0^* \\ \text{Alternative hypothesis} & H_1 : \beta_0 \neq \beta_0^*\end{array}\right\} \quad \ldots (7)$$

**Step 2: Compute the test statistic**

Under the null hypothesis $H_0 : \beta_0 = \beta_0^*$, we define the test statistic as follows:

$$z = \frac{\hat{\beta}_0 - E(\hat{\beta}_0)}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\left[\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)\right]}} \quad \text{... (8)}$$

where z-statistic follows a standard normal distribution, i.e., $z \sim N(0, 1)$. It is to be noted that we use the z-statistic given in equation (8) only when $\sigma^2$ is known. But, generally, the value of $\sigma^2$ is unknown. In such a situation, we consider an unbiased estimator of $\sigma^2$, i.e., $\hat{\sigma}^2$. You have learnt in MST-004 that when we use the estimated value of the variance, we apply t-test instead of z-test.

We use $\hat{\sigma}^2$ instead of $\sigma^2$ in equation (8) and obtain a t-statistic as:

$$t = \frac{\hat{\beta}_0 - E(\hat{\beta}_0)}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\left[\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)\right]}} \quad \text{... (9)}$$

where t-statistic follows t-distribution with (n – 2) degrees of freedom.

**Step 3: Determine the critical value**

We determine the critical or tabulated value $t_{(n-2),\alpha/2}$ for t-statistic at α% level of significance and (n – 2) degrees of freedom. This value has been tabulated for various degrees of freedom in Table I given at the end of this block as explained in Unit 4 of MST-004.

**Step 4: Take decision about the null hypothesis**

If $|t| \geq t_{(n-2),\alpha/2}$, we reject the null hypothesis at $\alpha\%$ level of significance (Fig. 6.11). Otherwise, we do not reject it. In other words, we can say that there are enough evidences against $H_0$ to conclude that the estimated regression line does not pass through the point $\beta_0^*$.

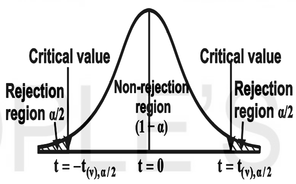



**Fig. 6.11**

## 6.3.2. Hypothesis testing for the slope

For testing the hypothesis that the sample comes from the population for which the slope $(\beta_1)$ is equal to $\beta_1^*$, we follow the same steps as used for the intercept:

**Step 1: Set up the hypotheses**

We formulate the null and alternative hypotheses as:

$$\left.\begin{array}{ll} \text{Null hypothesis} & H_0 : \beta_1 = \beta_1^* \\ \text{Alternative hypothesis} & H_1 : \beta_1 \neq \beta_1^* \end{array}\right\} \quad \text{... (10)}$$

**Step 2: Compute the test statistic**

If the null hypothesis $H_0 : \beta_1 = \beta_1^*$ is true, we define a test statistic as:

$$z = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2 / SS_x}} \qquad \ldots (11)$$

where z follows a standard normal distribution, i.e., $z \sim N(0, 1)$. When $\sigma^2$ is unknown, we consider $\hat{\sigma}^2$ instead of $\sigma^2$ and apply t-test instead of z-test. We define a t-statistic as:

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}^2 / SS_x}} \qquad \ldots (12)$$

This t-statistic follows a t-distribution with (n – 2) degrees of freedom.

**Step 3: Determine the critical value**

We read out the critical or tabulated value $t_{(n-2),\alpha/2}$ for t-statistic at α% level of significance and (n – 2) degrees of freedom. This value has been tabulated for various degrees of freedom in Table I given at the end of this block.

**Step 4: Take decision about the null hypothesis**

If $|t| \geq t_{(n-2),\alpha/2}$, we may reject the null hypothesis at α% level of significance and say that we have enough evidences against $H_0$ (Fig. 6.12). Otherwise, we do not reject $H_0$. If we consider the value of $\beta_1^* = 0$ in testing of slope, we may infer that there are sufficient evidences to conclude that there exists a linear relationship between Y and X for rejection of the null hypothesis $H_0 : \beta_1 = 0$.

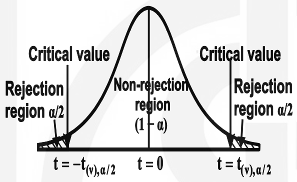



**Fig. 6.12**

It is to be noted that we can also use p-value approach to make a decision. If p-value is less than the significance level α, we may reject the null hypothesis. Otherwise, we do not reject it at α% level of significance.

It is also important to note that we consider the values of $\beta_0^*$ and $\beta_1^*$ equal to zero for testing the significance of individual regression coefficients, i.e., for testing significance of intercept ($H_0 : \beta_0 = 0$) and slope ($H_0 : \beta_1 = 0$), respectively.

Note that here we have considered a two-sided alternative hypothesis for testing the intercept and slope. You can also consider one sided alterative hypothesis according to your requirement as we have explained in MST-004.

So far in this section, we have discussed various steps involved in testing the significance of intercept and slope. Let us now take up an example to illustrate the method.

**Example 4:** For the SBP data given in Example 2 of Unit 5, test the following null hypothesis for testing the significance of the regression coefficients at 5% level of significance:

i) $H_0: \beta_0 = 90$ against $H_1: \beta_0 \neq 90$

ii) $H_0: \beta_1 = 2$ against $H_1: \beta_1 \neq 2$

**Solution:** From the solution of Example 2 of Unit 5, we have

$$\hat{\beta}_0 = 90.0427 \text{ and } \hat{\beta}_1 = 1.1189$$

From the solution of Example 3 of Unit 5, we have

$SE(\hat{\beta}_0) = 2.5644$ and $SE(\hat{\beta}_1) = 0.0801$

(i) $H_0 : \beta_0 = 90$ against $H_1 : \beta_0 \neq 90$

Here we have been given $\beta_0^* = 90$

We now compute the value of t-statistic using equation (9) as:

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{SE(\hat{\beta}_0)} = \frac{90.0427 - 90}{2.5644} = \frac{0.0427}{2.5644} = 0.0166$$

$$|t| = 0.0166$$

Since $\alpha = 0.05$ and $\alpha/2 = 0.025$, the tabulated t value at 5% level of significance with 13 degrees of freedom will be $t_{13,0.025} = 2.160$.

Since the calculated t value, i.e., $|t| = 0.01664420 < 2.160$, we do not reject the null hypothesis at 5% level of significance. Hence, we may conclude that there is insufficient evidence against $H_0$ and we may consider the value of intercept to be equal to 90mm/Hg.

It is to be noted that all calculations were performed up to 15 fixed decimal places for showing accurate results in this block. For the sake of simplicity, we are showing results up to 4 decimal places only. The results may vary if we carry out the calculations by fixing values at various decimal places.

(ii) $H_0 : \beta_1 = 2$ against $H_1 : \beta_1 \neq 2$

Here we have been given $\beta_1^* = 2$

We determine the value of t-statistic using equation (12) as:

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)} = \frac{1.1189 - 2}{0.0801} = \frac{-0.8811}{0.0801} = -10.9990$$

$$|t| = 10.9990$$

From Table I given at the end of this block, the tabulated value of t at 5% level of significance with 13 degrees of freedom is 2.160. Since the calculated value |t|=10.99897691> 2.160, we may reject the null hypothesis at 5% level of significance. Since there are evidences against the null hypothesis $H_0$, we may conclude that the value of $\beta_1$ is not equal to 2 mm/Hg per year.

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

---

**E3)** Explain the rejection/acceptance criteria for testing the significance of intercept and slope.

**E4)** For the exercise given in **E2**, test the following null hypothesis for testing the significance of the regression coefficients at 1% level of significance:

i) $H_0$: $\beta_0 = 470$ against $H_1$: $\beta_0 \neq 470$

ii) $H_0$: $\beta_1 = 0$ against $H_1$: $\beta_1 \neq 0$

**E5)** For the exercise given in **E3**, test the following null hypothesis for testing the significance of the regression coefficients at 5% level of significance:

i) $H_0$: $\beta_0 = 0$ against $H_1$: $\beta_0 \neq 0$

ii) $H_0$: $\beta_1 = 1.5$ against $H_1$: $\beta_1 \neq 1.5$

---

## 6.4 TEST OF SIGNIFICANCE OF OVERALL SIMPLE LINEAR REGRESSION MODEL

In Sec. 6.3, we have explained how to test the significance of the individual regression coefficients, namely, intercept and slope. The testing significance of the overall regression model is the same as testing the significance of the slope taking $\beta_1^* = 0$ in the case of simple regression model. We set-up the null and alternative hypotheses to test the significance of the simple linear regression model as:

$$\left.\begin{array}{ll}\text{Null hypothesis} & H_0 : \beta_1 = 0 \\ \text{Alternative hypothesis} & H_1 : \beta_1 \neq 0\end{array}\right\} \quad \ldots (13)$$

We can use two approaches to test the significance of the regression model: (i) t-test and (ii) ANOVA.

Both ANOVA and t-test methods provide the same results and serve the same purpose when we deal with the simple regression model, i.e., when we have only one regressor variable in the model. In this unit, we are considering only t-test approach which is the same as described in Sec. 6.2.2. We shall explain the ANOVA approach in Unit 7 in the case of multiple regression model, which can be applied for simple regression model as well. If we are able to reject the null hypothesis on the basis of the results obtained, we can say that there is sufficient evidence to conclude that there exists a linear relationship between Y and X.

Let us solve the following example to understand the procedure for testing the significance of the fitted simple regression model numerically.

**Example 5:** Test the significance of the regression model fitted in Example 2 given in Unit 5.

**Solution:** For testing the significance of the fitted model

$$Y = 90.0427 + 1.1189\ X\,,$$

we consider the null hypotheses $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

Here we have $\beta_1^* = 0$

We determine the value of t-statistic using equation (12) as:

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)} = \frac{1.1189 - 0}{0.0801} = \frac{1.1189}{0.0801} = 13.9676$$

$$|t| = 13.9676$$

From Table I given at the end of this block, the tabulated value of t at 5% level of significance with 13 degrees of freedom is 2.160. Since the calculated value |t|=13.96755891> 2.160, we may reject the null hypothesis at 5% level of significance. Since there is enough evidence against the null hypothesis $H_0$, we may conclude that the value of $\beta_1$ is not equal to zero and the fitted regression model is significant.

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

**E6)** For the exercise given in **E4**, test the significance of the fitted regression model at 1% level of significance.

**E7)** For the exercise given in **E5**, test the significance of the fitted regression model at 5% level of significance.

# 6.5 CONFIDENCE INTERVAL OF REGRESSION COEFFICIENTS

You have learnt in Block 3 of MST-004 that we usually compute the confidence interval (CI) to determine the range, which contains the unknown value of the parameters, e.g., regression coefficients ($\beta_0$ and $\beta_1$) in case of simple regression analysis. We can also test the significance of the regression coefficients, which we did through the hypothesis testing method discussed in Sec. 6.3 with the help of confidence interval. For example, if the confidence interval of $\beta_1$ does not contain the value of the regression coefficients under null hypotheses, we conclude that we have enough evidence against null hypothesis and may reject the null hypothesis. On the other hand, if it contains the value of the regression coefficient under $H_0$, we conclude that we do not have enough evidence against $H_0$ and do not reject $H_0$. Note that the confidence interval provides a range of possible values of regression coefficients instead of a fixed estimated value as we obtained using the least squares method in Unit 5 as a point estimation.

You have learnt the concept of confidence interval in Block 3 of MST-004 in detail. You know that we can define the (1 – α) 100% confidence interval for an estimator as follows:

(i) When σ is known

$$\text{Confidence Interval} = \text{Estimator} \pm z_{\alpha/2} \times \text{SE(Estimator)}$$

(ii) When σ is unknown

$$\text{Confidence Interval} = \text{Estimator} \pm t_{(n-2),\alpha/2} \times \text{SE(Estimator)}$$

Therefore, we can define the lower and upper limits of $(1-\alpha)100\%$ confidence interval of $\beta_0$ as follows:

$$\text{Lower Limit}: \beta_{0L} = \hat{\beta}_0 - t_{(n-2),\alpha/2}\,SE(\hat{\beta}_0)$$
$$\text{Upper Limit}: \beta_{0U} = \hat{\beta}_0 + t_{(n-2),\alpha/2}\,SE(\hat{\beta}_0) \quad \ldots (14)$$

In the same way, we can determine the $(1-\alpha)100\%$ confidence limits for $\beta_1$ as:

$$\text{Lower Limit}: \beta_{1L} = \hat{\beta}_1 - t_{(n-2),\alpha/2}\,SE(\hat{\beta}_1)$$
$$\text{Upper Limit}: \beta_{1U} = \hat{\beta}_1 + t_{(n-2),\alpha/2}\,SE(\hat{\beta}_1) \quad \ldots (15)$$

Thus, the confidence interval for $\beta_0$ and $\beta_1$ will be $(\beta_{0L}, \beta_{0U})$ and $(\beta_{1L}, \beta_{1U})$, respectively.

Let us solve an example to get a better idea about computation of the confidence interval of the regression coefficients ($\beta_0$ and $\beta_1$).

**Example 6:** For the SBP and age data given in Example 1 of Unit 5, obtain the 95% confidence interval for the regression coefficients $\beta_0$ and $\beta_1$.

**Solution:** From the solution of Examples 2 and 3, we have

$$\hat{\beta}_0 = 90.0427 \text{ and } \hat{\beta}_1 = 1.1189$$

$$SE(\hat{\beta}_0) = 2.5644, \quad SE(\hat{\beta}_1) = 0.0801 \text{ and } t_{13,0.025} = 2.160$$

**Interpretation:** This confidence interval indicates that the values of $\beta_0$ for different samples taken from the same population of SBP patients. The lowest value is around 85 and the highest value being 96. For the given data, it is found to be approximately 90 which is included in the computed confidence interval.

The lower and upper confidence limits of $\beta_0$ can be determined using equation (14) as:

$$\beta_{0L} = \hat{\beta}_0 - t_{(n-2),\alpha/2} SE(\hat{\beta}_0)$$

$$= 90.0427 - 2.160 \times 2.5644 = 90.0427 - 5.5392 = 84.5035$$

$$\beta_{0U} = \hat{\beta}_0 + t_{(n-2),\alpha/2} SE(\hat{\beta}_0)$$

$$= 90.0427 + 2.160 \times 2.5644 = 90.0427 + 5.5392 = 95.5819$$

Thus, the confidence interval of $\beta_0$ is (84.5035, 95.5819).

**Interpretation:** This confidence interval indicates that the values of $\beta_1$ for different samples taken from the same population of SBP patients. The lowest value is around 0.95 and the highest value being 1.29. For the given data, it is found to be approximately 1.11 which is included in the computed confidence interval.

From equation (15), the lower and upper confidence limits of $\beta_1$ can be computed as:

$$\beta_{1L} = \hat{\beta}_1 - t_{(n-2),\alpha/2} SE(\hat{\beta}_1)$$

$$= 1.1189 - 2.160 \times 0.0801 = 1.1189 - 0.1730 = 0.9459$$

$$\beta_{1U} = \hat{\beta}_1 + t_{(n-2),\alpha/2} SE(\hat{\beta}_1)$$

$$= 1.1189 + 2.160 \times 0.0801 = 1.1189 + 0.1730 = 1.2919$$

Hence, the confidence interval of $\beta_1$ is (0.9459, 1.2919)

Before studying the next section, you may like to solve the following exercises for practice.

**E8)** For the exercise given in **E5**, determine the 99% confidence intervals of the intercept and slope.

**E9)** For the exercise given in **E6**, compute the 95% confidence intervals of the regression coefficients.

## 6.6 COEFFICIENT OF DETERMINATION

So far, you have learnt how to test the significance of individual regression coefficients and overall fitted regression model. You have also learnt how to compute confidence intervals for the regression coefficients. In this section, we discuss how to check the goodness of the fitted regression model using the coefficient of determination. After fitting the simple linear regression model, we determine the percentage of total variation in Y, which is explained by the fitted regression model. This percentage is measured by a quantity known as the coefficient of determination. We generally denote it by $R^2$. It also describes the strength of linear relationship between the response variable (Y) and the regressor variable (X).

The coefficient of determination ($R^2$) measures the proportion of variability around the mean response, i.e., $\bar{y}$, which is explained by the regression model. In other words, we can say that it is the proportion of variation in Y explained by the fitted regression model. Therefore, we define the coefficient of determination as:

$$R^2 = \frac{\text{Variation explained by regression model}}{\text{Total variation in Y}}$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{... (16)}$$

We can also redefine $R^2$ as:

$$R^2 = 1 - \frac{\text{Variation not explained by regression model}}{\text{Total Variation}}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{... (17)}$$

where $(y_i - \hat{y}_i) = r_i$ is the value of the $i^{th}$ residual.

We can also compute the value of $R^2$ with the help of the value calculated in ANOVA table which will be discussed in Unit 8.

The quantity in the **numerator** of equation (16) measures the remaining variability in Y after considering the regressor variable (X), while the **denominator** measures the variability in Y without considering the regressor variable (X). The values of $R^2$ lie between 0 and 1, i.e., $0 \le R^2 \le 1$. The quantity $(1 - R^2)$ is known as the coefficient of alienation (non-determination). It measures the proportion of variability around the mean response which is not explained by the regression model.

The following example will help you understand how to compute and interpret the coefficient of determination of the fitted simple linear regression model:

**Example 7:** Using the data given Example 1, calculate the coefficient of determination and comment on the goodness of fit of the regression model.

**Solution:** From Table 1 given in Example 1, we obtain

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 32.8354 \quad \text{and} \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = 525.60$$

We calculate the coefficient of determination using equation (17) as:

$$R^2 = 1 - \frac{32.8354}{525.60}$$

$$= 1 - 0.0625 = 0.9375$$

$$\text{or, } R^2 \cong 93.75\%$$

Coefficient of alienation:

$$\left(1-R^2\right)=1-0.9375=0.0625$$

It implies that the fitted regression model explains 93.75% of variation in Y due to regressor variable X and only approximately 6.25% variation in Y is not explained by the fitted regression model. Since the explained variation is quite high, we can consider the fitted regression model as a good model.

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

---

**E10)** Differentiate between coefficient of determination and coefficient of alienation.

**E11)** For the exercise given in **E5,** determine the coefficient of determination and coefficient of alienation. Also interpret the results.

**E12)** Obtain the coefficient of determination and coefficient of alienation for the exercise given in **E6**. Also interpret the results.

---

Let us now summarise what you have learnt in this unit.

## 6.7 SUMMARY

1. We use residual analysis to check the validity of some basic assumptions and ensure the adequacy of the regression model. For detecting non-linearity and/or unequal variance of error terms, we use the residual plot, while the normal probability plot is used for detecting non-normality in the error terms.
2. The **residual plot** considers the residuals on the vertical (Y) axis and the predicted values or the regressor variable on the horizontal (X) axis. If the points on a residual plot are randomly dispersed around a horizontal band, a linear regression model will be appropriate for the given data; otherwise, a non-linear model will be more appropriate.
3. The normality assumption is important because the t-test, F-test and confidence intervals depend on it. We construct the normal probability plot to check the validity of the normality assumption.
4. In the **normal probability plot**, we plot cumulative probabilities ($p_i$) or percentile cumulative probabilities ($P_i$) on the vertical (Y) axis and the standardised residuals on the horizontal (X) axis. To satisfy the assumption of normality, the resulting points should lie approximately along a straight line. We determine the straight line visually, which passes through the central values of all the points rather than the extreme points at both ends in the normal probability plot.
5. For testing the significance of the intercept and slope when the value of $\sigma^2$ is unknown, we apply t-test instead of z-test.
6. The t-statistic for testing the $H_0:\beta_0=\beta_0^*$ against $H_1:\beta_0\neq\beta_0^*$ is given as:

$$t=\frac{\hat{\beta}_0-\beta_0^*}{SE(\hat{\beta}_0)}=\frac{\hat{\beta}_0-\beta_0^*}{\sqrt{\left[\hat{\sigma}^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SS_x}\right)\right]}}$$

7. We define a t-statistic for testing the $H_0:\beta_1=\beta_1^*$ against $H_1:\beta_1\neq\beta_1^*$ as

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}^2 / SS_x}}$$

8. Testing the significance of the fitted simple linear regression model is the same as testing the significance of the slope when $\beta_1^* = 0.$ Most often we are interested in testing whether slope $(\beta_1)$ is equal to 0 or not, so that we can draw inference about the independence or dependence of the response variable on the regressor variable in the population.
9. We define the lower and upper limits of $(1-\alpha)100\%$ confidence interval $(\beta_{0L,}\ \beta_{0U})$ of $\beta_0$ when σ is unknown as:

$$\text{Lower Upper}: \beta_{0L} = \hat{\beta}_0 - t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_0\right)$$

$$\text{Upper Lower}: \beta_{0U} = \hat{\beta}_0 + t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_0\right)$$

10. We determine the lower and upper limits of $(1-\alpha)100\%$ confidence interval $(\beta_{1L,}\ \beta_{1U})$ of $\beta_1$ when σ is unknown as:

$$\text{Lower Limit}: \beta_{1L} = \hat{\beta}_1 - t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_1\right)$$

$$\text{Upper Limit}: \beta_{1U} = \hat{\beta}_1 + t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_1\right)$$

11. We measure the proportion of variation in Y explained by the fitted regression model by the **coefficient of determination** ($R^2$) as:

$$R^2 = 1 - \frac{\text{Variation not explained by regression model}}{\text{Total Variation in Y}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

12. The quantity $(1-R^2)$ is known as the **coefficient of alienation** (non-determination).

## 6.8 SOLUTIONS / ANSWERS

**E1)** We have, n – 2 = 15 − 2 = 13 and $\sum_{i=1}^{15} r_i^2 = 3387.5969$

The variance of the residuals can be estimated as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{15} r_i^2}{13} = \frac{3387.5969}{13} = 260.5844$$

| S. No. | Predicted Random Blood Sugar ($\hat{y}_i$) | Residuals ($r_i$) | $r_i^2$ | Standard Residuals ($s_i$) |
|---|---|---|---|---|
| 1 | 430.4264 | −0.4264 | 0.1818 | −0.0264 |
| 2 | 419.4186 | 0.5814 | 0.3380 | 0.0360 |
| 3 | 408.4109 | 1.5891 | 2.5254 | 0.0984 |
| 4 | 408.4109 | −8.4109 | 70.7424 | −0.5210 |
| 5 | 375.3876 | 14.6124 | 213.5223 | 0.9052 |
| 6 | 397.4031 | −2.4031 | 5.7749 | −0.1489 |

| | | | | |
|---|---|---|---|---|
| 7 | 408.4109 | 11.5891 | 134.3083 | 0.7179 |
| 8 | 397.4031 | 12.5969 | 158.6819 | 0.7804 |
| 9 | 397.4031 | 2.5969 | 6.7439 | 0.1609 |
| 10 | 386.3953 | 3.6047 | 12.9935 | 0.2233 |
| 11 | 375.3876 | 4.6124 | 21.2743 | 0.2857 |
| 12 | 364.3798 | 5.6202 | 31.5861 | 0.3482 |
| 13 | 397.4031 | −7.4031 | 54.8059 | −0.4586 |
| 14 | 353.3721 | 11.6279 | 135.2082 | 0.7203 |
| 15 | 375.3876 | −50.3876 | 2538.9099 | −3.1214 |
| Total | 5895 | 0 | 3387.5969 | 0 |

The value of standardised residual for the 15th observation is –3.0 which is an indication of outlier (Fig. 6.13).

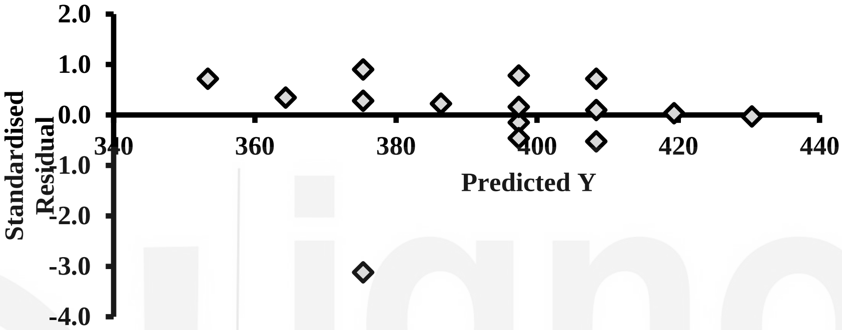



**Fig. 6.13**

The standardised residuals shown in Fig. 6.13 appear to have downward trend. Hence, the assumption of linear regression does not seem to be valid. It also indicates the presence of an outlier in the data.

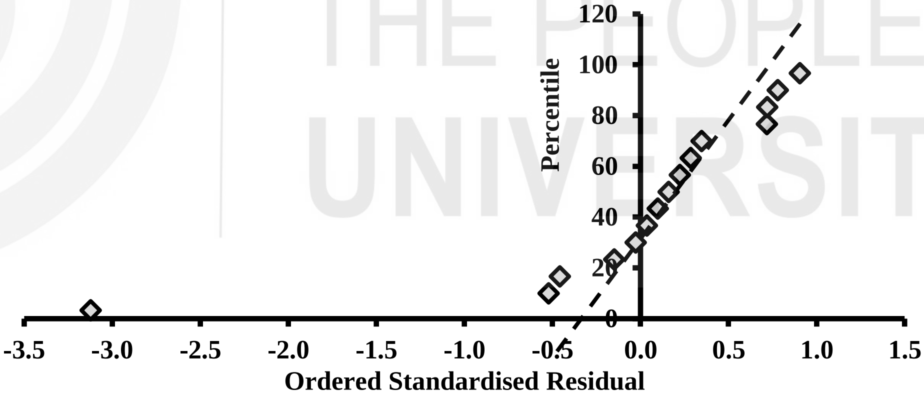



**Fig. 6.14**

The resulting points are not lying approximately on a straight line as shown in Fig. 6.14. It indicates that the distribution of error terms is not normally distributed.

**E2)** We obtained

$$\sum_{i=1}^{15}\left(y_i - \hat{y}_i\right)^2 = 0.6314$$

From equation (2), we have

$$\hat{\sigma}^2 = \frac{1}{10}\sum_{i=1}^{12}\left(y_i - \hat{y}_i\right)^2 = \frac{0.6314}{10} = 0.0631$$

| S. No. | Predicted Birth Weight | Residuals | Standard Residuals |
|---|---|---|---|
| 1 | 2.4303 | − 0.0303 | − 0.1207 |
| 2 | 2.5300 | − 0.2300 | − 0.9154 |
| 3 | 1.8322 | 0.1678 | 0.6680 |
| 4 | 3.0285 | − 0.1285 | − 0.5113 |
| 5 | 2.7294 | − 0.0294 | − 0.1170 |
| 6 | 3.6836 | − 0.4836 | − 1.9246 |
| 7 | 3.3988 | 0.0012 | 0.0049 |
| 8 | 2.9715 | − 0.1715 | −0.6826 |
| 9 | 3.0570 | 0.1430 | 0.5692 |
| 10 | 3.4700 | 0.2300 | 0.9154 |
| 11 | 3.5697 | 0.4303 | 1.7125 |
| 12 | 3.2991 | 0.1009 | 0.4016 |

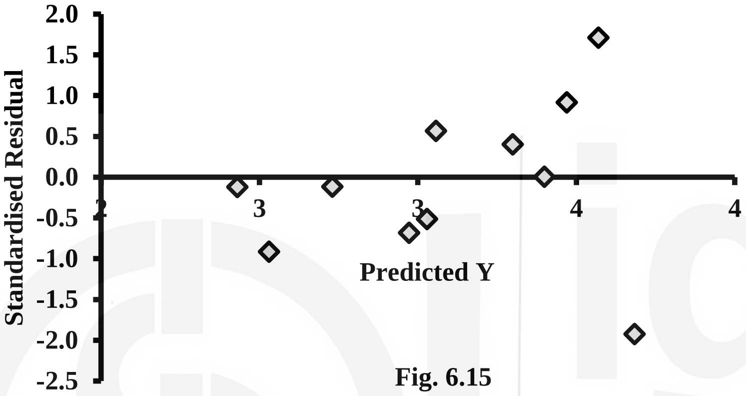



**Fig. 6.15**

The standardised residuals shown in Fig. 6.15 appear to have a curved pattern. Hence, the assumption of linear regression does not seem to be valid.

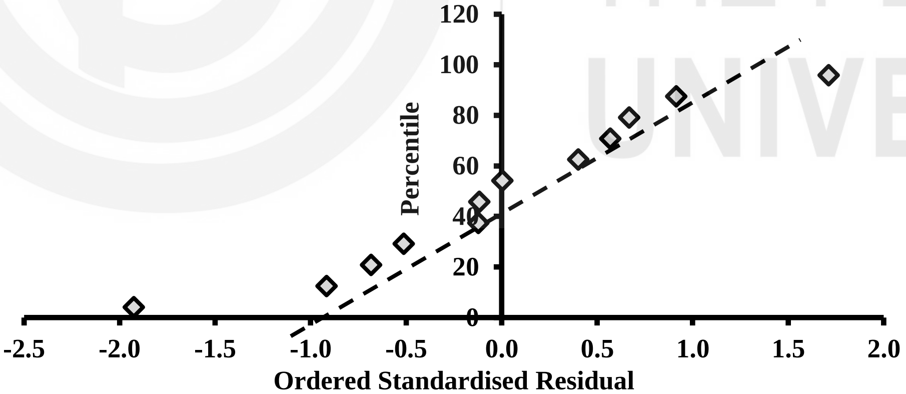



**Fig. 6.16**

The resulting points do not lie approximately on a straight line as shown in Fig. 6.16. It indicates that the error terms are not normally distributed.

**E3)** Refer to **Sec 6.3**.

**E4)** From the solution of **E4** and **E7** of Unit 5, we have

$$\hat{\beta}_0 = 474.4574 \text{ and } \hat{\beta}_1 = -2.2016$$

$$SE(\hat{\beta}_0) = 17.1439 \text{ and } SE(\hat{\beta}_1) = 0.4494$$

(i) $H_0: \beta_0 = 470$ against $H_1: \beta_0 \neq 470$

Here we have $\beta_0^* = 470$

We now compute the value of t-statistic using equation (9) as:

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{SE\left(\hat{\beta}_0\right)} = \frac{474.4574 - 470}{13.4169} = \frac{4.4574}{17.1439} = 0.25999$$

$$|t| = 0.25999$$

Since $\alpha = 0.01$ and, $\alpha / 2 = 0.005$

From Table I given at the end of this block, we determine the tabulated t value at 1% level of significance with 13 degrees of freedom as:

$$t_{13,0.005} = 3.012$$

Since $|t| = 0.25999 < 3.012$, we do not reject the null hypothesis at 1% level of significance. Hence, we may conclude that there is no evidence against the $H_0$.

(ii) $H_0$: $\beta_1 = 0$ against $H_1$: $\beta_1 \neq 0$

Here we have $\beta_1^* = 0$

We determine the value of t-statistic using equation (12) as:

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)} = \frac{-2.2016 - 0}{0.4494} = -4.8983$$

$$|t| = 4.8983$$

From Table I given at the end of this block, the tabulated value of t at 1% level of significance with 13 degrees of freedom is 3.012. Since $|t| = 4.8983 > 3.012$, we may reject the null hypothesis at 1% level of significance. We may now conclude that the value of $\beta_1$ is not equal to 0.

**E5)** From the solution of **E5** and **E8** of Unit 5, we have

$$\hat{\beta}_0 = -2.4120 \quad \text{and} \quad \hat{\beta}_1 = 0.1424$$

$$SE(\hat{\beta}_0) = 0.7580 \quad \text{and} \quad SE(\hat{\beta}_1) = 0.0199$$

(i) $H_0$: $\beta_0 = 0$ against $H_1$: $\beta_0 \neq 0$

Here we have $\beta_0^* = 470$

We now compute the value of t-statistic using equation (9) as:

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{SE\left(\hat{\beta}_0\right)} = \frac{-2.4120 - 0}{0.7580} = -3.1821$$

$$|t| = 3.1821$$

Since $\alpha = 0.05$ and, $\alpha / 2 = 0.025$

From Table I given at the end of this block, we determine the tabulated t value at 5% level of significance with 10 degrees of freedom as:

$$t_{10,0.025} = 2.228$$

Since $|t| = 3.1821 > 2.228$ we may reject the null hypothesis at 5% level of significance.

(ii) $H_0$: $\beta_1 = 0.15$ against $H_1$: $\beta_1 \neq 0.15$

Here we have $\beta_1^* = 0.15$

We determine the value of t-statistic using equation (12) as:

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)} = \frac{0.1424 - 0.15}{0.0199} = -0.3818$$

$$|t| = 0.3818$$

From Table I, the value of t at 5% level of significance with 10 degrees of freedom is 2.228. Since |t| = 0.3818 < 2.228, we do not reject the null hypothesis at 5% level of significance. We may now conclude that the value of $\beta_1$ may be considered as equal to 0.15.

**E6)** Same as **E4 (ii)**

**E7)** $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

Here we have $\beta_1^* = 0$

We determine the value of t-statistic using equation (12) as:

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE\left(\hat{\beta}_1\right)} = \frac{0.1424 - 0}{0.0199} = 7.1729$$

$$|t| = 7.1729$$

The tabulated value of t at 5% level of significance with 10 degrees of freedom is 2.228. Since $|t| = 7.1729 > 2.228$, we may reject the null hypothesis at 5% level of significance. We may conclude that the fitted regression model is significant.

**E8)** From the solution of **E4**, we have

$$\hat{\beta}_0 = 474.4574 \text{ and } \hat{\beta}_1 = -2.2016$$

$$SE(\hat{\beta}_0) = 17.1439,\ SE(\hat{\beta}_1) = 0.4494 \text{ and}$$

$$t_{13,0.005} = 3.012$$

The lower and upper confidence limits of $\beta_0$ are:

$$\beta_{0L} = \hat{\beta}_0 - t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_0\right)$$

$$= 474.4574 - 3.012 \times 17.1439 = 422.8198$$

$$\beta_{0U} = \hat{\beta}_0 + t_{(n-2),\alpha/2} SE\left(\hat{\beta}_0\right)$$

$$= 474.4574 + 3.012 \times 17.1439 = 526.0949$$

The lower and upper confidence limits of $\beta_1$ are:

$$\beta_{1L} = \hat{\beta}_1 - t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_1\right)$$

$$= -\ 2.2016 - 3.012 \times 0.4494 = -3.5553$$

$$\beta_{1U} = \hat{\beta}_1 + t_{(n-2),\alpha/2} SE\left(\hat{\beta}_1\right)$$

$$= -\ 2.2016 + 3.012 \times 0.4494 = -0.8478$$

**E9)** From the solution of **E5**, we have

$$\hat{\beta}_0 = -2.4120 \text{ and } \hat{\beta}_1 = 0.1424$$

$$SE(\hat{\beta}_0) = 0.7580,\ SE(\hat{\beta}_1) = 0.0199 \text{ and } t_{10,0.025} = 2.228$$

Hence, from equation (14), the lower and upper confidence limits of $\beta_0$ are:

$$\beta_{0L} = \hat{\beta}_0 - t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_0\right)$$

$$= -2.4120 - 2.228 \times 0.7580 = -4.1007$$

$$\beta_{0U} = \hat{\beta}_0 + t_{(n-2),\alpha/2} SE\left(\hat{\beta}_0\right)$$

$$= -2.4120 + 2.228 \times 0.7580 = -0.7232$$

From equation (15), the lower and upper confidence limits of $\beta_1$ are:

$$\beta_{1L} = \hat{\beta}_1 - t_{(n-2),\alpha/2}\, SE\left(\hat{\beta}_1\right)$$

$$= 0.1424 - 2.228 \times 0.0199 = 0.0982$$

$$\beta_{1U} = \hat{\beta}_1 + t_{(n-2),\alpha/2} SE\left(\hat{\beta}_1\right)$$

$$= 0.1424 + 2.228 \times 0.0199 = 0.1867$$

**E10)** Refer to **Sec. 6.4**.

**E11)** From the solution of **E2**, we have

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = 3387.5969 \text{ and } \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 9640$$

Coefficient of determination:

$$R^2 = 1 - \frac{3387.5969}{9640} = 1 - 0.3514 = 0.6486$$

$$\cong 64.86\%$$

Coefficient of alienation:

$$\left(1 - R^2\right) = 1 - 0.6486 = 0.3514 \cong 35.14\%$$

The fitted regression model explains only 64.86% of variation in Y due to regressor variable X. Approximately 35.14% variation in Y is not explained by the fitted regression model.

**E12)** From the solution of **E3**, we have

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = 0.6314 \quad \text{and} \quad \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 3.88$$

Coefficient of determination:

$$R^2 = 1 - \frac{0.6314}{3.88} = 1 - 0.1627 = 0.8373 \cong 83.73\%$$

Coefficient of alienation:

$$\left(1 - R^2\right) = 1 - 0.8373 = 0.1627 \cong 16.27\%$$

The fitted regression model explains 83.73% of variation in Y due to regressor variable X. Approximately 16.27% variation in Y is not explained by the fitted regression model.