Or $\beta = \hat{\beta} \pm t_{0.025}\frac{\sigma^*}{\sqrt{\Sigma x_i^2}}$

Which gives confidence level of β.

For hypothesis testing concerning regression parameters we hypothesise that there is no relationship between the explanatory variable X and the dependent variable Y in the regression model $Y = \alpha + \beta X$.

Here our, $H_0 : \beta = 0$ and

$H_A : \beta \neq 0$

To test we take t-statistics and determine the acceptance and critical region.

If we assume $H_0 : \beta = 0$ is true, then

$$T = \frac{\hat{\beta}\sqrt{\Sigma x_i^2}}{\sigma^*} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

$$\because \quad \sigma_{\hat{\beta}}^2 = \frac{\sigma^{*2}}{\Sigma x_i^2}$$

This has t-distribution with (n-2) degree of freedom and the boundary between acceptance and critical region can be determined from t-distribution table for any given level of significance and degree of freedom.

The acceptance region for 2-tailed test at (n-2) degree of freedom will be

$$-t_{0.025}\ S.E(\hat{\beta}) \leq \hat{\beta} \leq +t_{0.025}\ S.E(\hat{\beta})$$

2.12 Coefficient of Determination or Goodness of Fit :

The measure of goodness of fit is the square of correlation coefficient or r^2. It is the percentage of total variation in the dependent variable which can be explained by the independent variable, Therefore r^2 is known as coefficient of determination. For example, if $r^2 = .70$, it means that the estimated regression line is able to explain 70% of the total variation of dependent variables around mean.

In a simple linear regression model

$Y_i = \alpha + \beta x_i + u_i$

Total variation in $Y = \sum_{i=1}^{n} y_i^2$

$$= \sum_{i=1}^{n}(Y - \bar{Y})^2 \quad \text{..................(i)}$$

The explained variable is

$$\Sigma \hat{y}_i^2 = \Sigma(\hat{Y}_i - \bar{Y}_i)^2 \quad \text{....................(ii)}$$

We know variation $e_i = Y_i - \hat{Y}_i$ is not explained by the regression line. Thus, the sum of the squared residuals gives the total unexplained variation of dependent variable around mean.

Unexplained variation = $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ .... (iii)

Thus, total variation in Y = Explained variation + unexplained variation

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 \quad \text{.......... (iv)}$$

Where $e_i = Y_i - \hat{Y}$ = deviation of Y_i from the regression line

$y_i = Y_i - \overline{Y}$ = deviation Y_i from mean

$\hat{y}_i = \hat{Y}_i - \overline{Y}$ = deviation of the regression value $\hat{Y}_i$ from the mean

From residuals we have

$$ei = Yi - \hat{Y}_i \quad \text{..........(v)}$$

Putting the values

$$e_i = (y_i + \overline{Y}) - (\sum \hat{y}_i + \overline{Y})$$

$$e_i = y_i - \hat{y}_i$$

$$y_i = \hat{y}_i + e_i \quad \text{..........(vi)}$$

This equation shows that each deviation of Y from its mean consists components – (i) explained variation and (ii) unexplained variation.

Now, squaring equation (vi) and taking summation

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}(\hat{y}_i + e_i)^2$$

$$= \sum \hat{y}_i^2 + \sum e_i^2 + 2\sum \hat{y}_i e_i$$

$$= \sum y_i^2 + \sum e_i^2 \qquad [\because \sum_{i=1}^{n} \hat{y}_i e_i = o]$$

Since, $\hat{y}_i = \hat{y}_i - \overline{Y}$

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

And $\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X}$

Therefore, $y_i = (\hat{\alpha} + \hat{\beta}X_i) - (\hat{\alpha} + \hat{\beta}\overline{X})$

$$= \hat{\beta}(X_i - \overline{X})$$

$\hat{y}_i = \hat{\beta} x_i$ where, $x_i = X_i - \overline{X}$

We know $e_i = y_i - \hat{y}_i = y_i - \hat{\beta} x_i$

$$\therefore \sum_{i=1}^{n} y_i e_i = \sum_{i=1}^{n} \hat{\beta} x_i (y_i - \hat{\beta} x_i)$$

Or, $\sum_{i=1}^{n} y_i e_i = \hat{\beta}(\sum x_i y_i - \hat{\beta} \sum x_i^2)$

But $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$

$$\therefore \sum_{i=1}^{n} y_i e_i = \hat{\beta}\left(\sum x_i y_i - \frac{\sum x_i y_i}{\sum x_i^2} \sum x_i^2\right) = 0$$

$\therefore \quad \sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \quad$ ....(vii)

Or $\begin{bmatrix} \text{Total} \\ \text{variation} \end{bmatrix} = \begin{bmatrix} \text{Explained} \\ \text{variation} \end{bmatrix} + \begin{bmatrix} \text{Unexplained} \\ \text{variation} \end{bmatrix}$

The percentage of total variation and explained variation can be determined as –

$$\frac{\sum \hat{y}_i^2}{\sum y_i^2}$$

But , $\quad \hat{y}_i = \hat{\beta} x_i$

$$\therefore \quad \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{\beta} x_i)^2}{\sum y_i^2} = \hat{\beta} \frac{\sum x_i^2}{\sum y_i^2}$$

Substituting for $\hat{\beta}$ we have

$$\frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2}{(\sum x_i^2)^2} = \frac{\sum x_i^2}{\sum y_i^2}$$

$$\text{Or} \quad \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2} = r^2$$

$$\text{Since} \quad r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

Since, r^2 determines the proportion of variation in Y which is explained by variation in Y, called Co-efficient of determination.

Illustration 2.1

Obtain the usual regression results from the following data 20 pairs of observation on X and Y.

$\sum X_i = 228, \ \sum Y_i = 3121, \ \sum X_i Y_i = 38297, \sum X^2 = 3204$

$\sum x_i y_i = 3347.60, = \sum x_i^2 = 604.80$ and $\sum y_i^2 = 19837$

Solution :

We are asked to fit a linear regression line.

i) Estimation of $\hat{\alpha}$ and $\hat{\beta}$

Given $\sum X_i = 228$, n =20 $\therefore \bar{X} = \frac{\sum X_i}{n} = \frac{228}{20} = 11.4$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{3347.60}{604.80} = 5.54$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Or $\quad \hat{\alpha} = 156.05 - (5.54)(11.40) = 92.95$

Therefore our estimated regression line is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

Or $\quad \hat{Y}_i = 92.95 + 5.54 X_i$