Linear Regression

Regression Analysis is a Statistical technique is very useful

Best line has pass through avg value of x and y

Least Square Method.

$$y = b_0 + b_1 x$$

Slop for Estimated Regression equation.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Sum of Squares and Sum of cross-products.

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Slope } (m) = \frac{S_{xy}}{S_{xx}}$$

SSE = Error Sum of Squares

$$SSE = S_{yy} - \frac{S_{xy}}{S_{xx}}$$

Simple Linear Regression

We have n pairs of observations (x_1, y_1), (x_2, y_2), P (x_n, y_n).

The estimates of β_0 and β_1 should result in a line that is a "best fit" to the data.

German Scientist Karl Gauss (1777-1855) proposed estimating the parameters in equation

$$\hat{y} = b_0 + b_1 x + \epsilon$$

to minimize the sum of squares of the vertical deviations.

y-intercept for the Estimated Regression equation $b_0 = \bar{y} - b_1 \bar{x}$

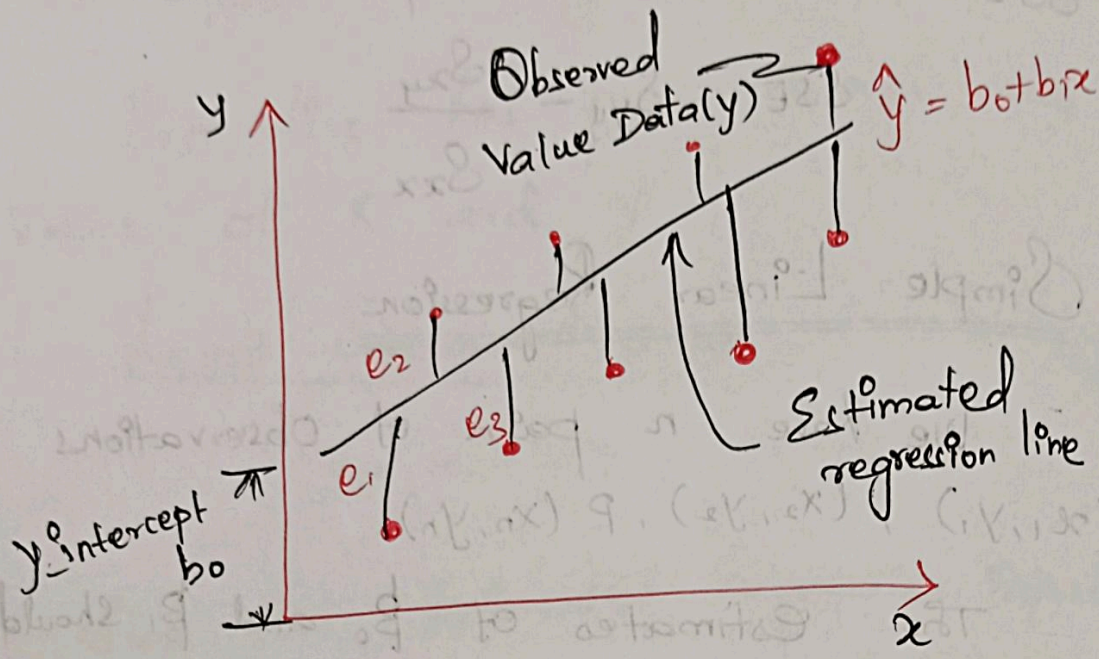
x_i - value of independent variable

y_i - value of dependent variable

$\bar{x}$ - mean value independent variable

$\bar{y}$ - mean value dependent variable

n - total number of observations.



Ex:

An Auto Company periodically has a special week-long sale. As part of the advertising campaign runs one or more television commercials during the weekend proceeding the sale

Data from 5 samples.

Number of TV Ads	Number of Cars Sold
1	14
3	24
2	18
1	17
3	27

y = Number of Cars Sold

x = Number of TV ads

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$
1	14	−1	−6	1
3	24	1	4	1
2	18	0	−2	0
1	17	−1	−3	1
3	27	1	7	1
$\bar{x}$ 2	$\bar{y}$ 20			4

Slope

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{(-1 \times -6) + (1 \times 4) + (0 \times -2) + (-1 \times -3) + (1 \times 7)}{4}$$

$$= \frac{6 + 4 + 0 + 3 + 7}{4}$$

$$b_1 = 20/4 = 5$$

Y-intercept

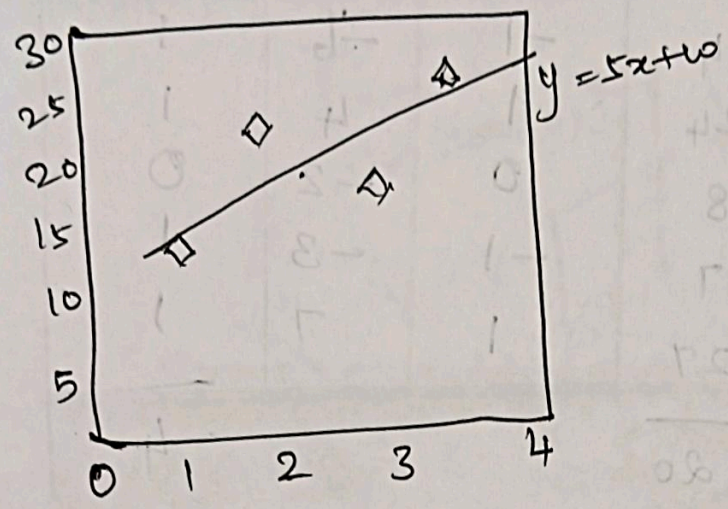
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$= 20 - 5(2)$$

$$b_0 = 10$$

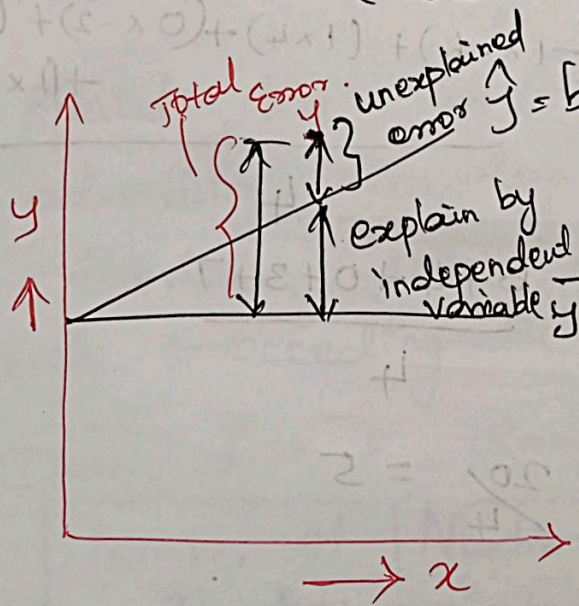
Estimated Regression Equation.

$$\hat{y} = 10 + 5x$$



Score = R^2 = Coefficient of Determination.

$$R^2 = \frac{SSR \text{ (Regression Sum of Square)}}{SST \text{ (Total Sum of Square)}}$$



Assume no independent Variable.

find mean of previous data

$$SST = SSR + SSE$$

$$\Sigma(y-\bar{y})^2 = \Sigma(\hat{y}-\bar{y})^2 + \Sigma(y-\hat{y})^2$$

$$R^2 = \frac{\text{Explained Error}}{\text{Total Error}}$$

$$= \frac{SSR}{SST}$$

R^2 interval 0 to 1

$$R^2 = \frac{SSR}{SST} = \frac{100}{114} = 0.8772.$$

The regression relationship is very strong; 88% of the variability in the number of cars sold can be explained by linear relationship between the number of TV ads and no. cars sold.

Sample Correlation Coefficient.

$$r_{xy} = (\text{Sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{Sign of } b_1)\sqrt{r^2}$$

$$\hat{y} = b_0 + b_1 x.$$

Correlation coefficient $r_{xy} = +\sqrt{.8772}$

$$r_{xy} = +.9366$$

Assumptions about Error Term e

- Error e is a random variable with mean of zero
- variance of e denoted by e^2, same for all values of independent variable
- values of e are independent
- error e is a normally distributed random variable

Testing for Significance

Conduct a hypothesis test to determine whether the value of β_1 is zero.

t Test and F Test.

both require an estimate of σ^2, the variance of e in regression model.

Estimate of σ

The mean Square error (MSE) provides the estimate of σ^2

$$s^2 = MSE = SSE/(n-2)$$

$$SSE = \Sigma (y_i - \hat{y}_i)^2$$

$$= \Sigma (y_i - b_0 - b_1 x_i)^2$$

Square root of s^2 resulting s is called Standard error of the estimate.

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

s_e = Standard error of the estimate (σ^2)

$$= \frac{SSE}{n-2} = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2}$$

t Test

Hypotheses

No relationship between x and y

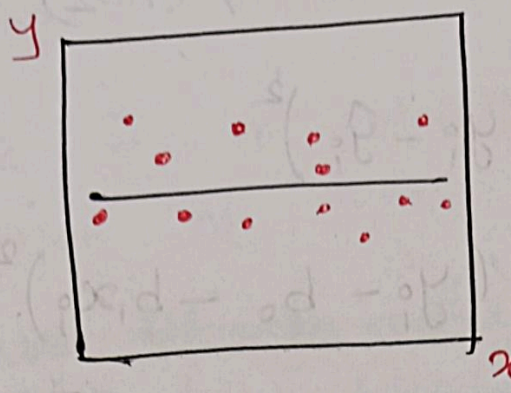
$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Test Statistic $t = \frac{b_1}{s_{b_1}}$

Case 1

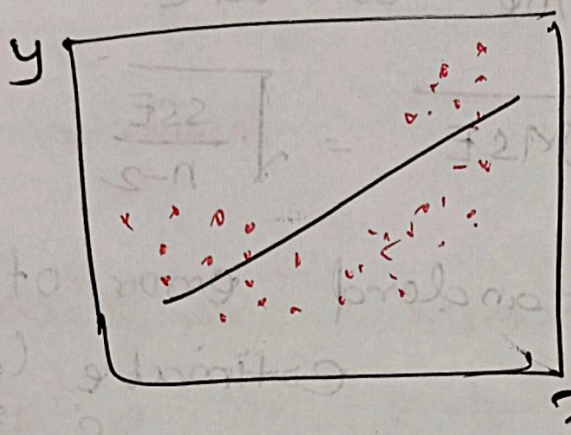
$\beta_1 = 0$.



hypothesis is not rejected

Case 2

$\beta_1 \neq 0$



hypothesis is rejected.

The Standard error of the regression slope coefficient (b_1) is

$$S_{b_1} = \frac{S_\epsilon}{\sqrt{(x-\bar{x})^2}} = \frac{S_\epsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

S_{b_1} = Estimate of Standard error of least Squares slope.

$$S_E = \sqrt{\frac{SSE}{n-2}}$$ Sample Standard error of estimate

Rejection Rule

Reject H_0 if P-value $\leq \alpha$

(or) $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

where

$t_{\alpha/2}$ based on t distribution.

$n-2$ degree of freedom.

Testing for significance : t Test

1. Determine the hypotheses

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

2. Specify the level of significance.

$\alpha = .05$

3. Select the test statistics

$$t = \frac{b_1}{s_{b_1}}$$

4. State the rejection rule.

Reject H_0 if P-value $\leq .05$

(or) $|t| > 3.182$ (with 3 degree of freedom).

5. Compute the value of test statistic

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{5}{1.08} = 4.63$$

6. Determine whether to reject H_0.

$t = 4.541$ provides an area of .01 in the upper tail. Hence, the p-value is less than .02. (Also $t = 4.63 > 3.182$)

We can reject H_0.

Hypothesis Tests for the Slope of the Regression model

$$t = \frac{b_1 - \beta_1}{S_b}$$

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

$H_0 : \beta_1 \leq 0$
$H_1 : \beta_1 > 0$

$H_0 : \beta_1 \geq 0$
$H_1 : \beta_1 < 0$

where $S_b = \frac{S_e}{\sqrt{SS_{xx}}}$

$S_e = \sqrt{\frac{SSE}{n-2}}$

$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$

β_1 = hypothesized slope

$df = n - 2$