Unit 4 :- Multiple Correlation and Regression

Partial correlation coefficient

It is the statistical technique of studying rel^{n}ship between a dependent variable and an independent variable after controlling the effect of other independent variables.

Types:

i) Zero Order Partial Correlation Coefficient.
ii) First Order Partial Correlation Coefficient
(iii) Higher Order Partial Correlation Coefficient

(1) Zero order Partial correlation coefficient

It studies the relationship betw two variables but without controlling other variables on both correlating variables.

(2) First order partial correlation coefficient

→ It studies the rel^{n}ship between two variable after removing the association of third variable with both of those correlating variables.

(3) Higher order partial correlation coefficient

→ It studies the degree of association beth two variables after controlling for the effect of two or more variables.

Imp

(2) First order partial correlation coeff.

Consider the study of 3 variables; Y, X_1 & X_2. where Y being a dependent variable & X_1 & X_2 being independent variables. We may be interested to know correlation beth Y & X_1, Y & X_2 & X_1 & X_2. But correlation beth any two variable may be partly due to the correlation ~~beth~~ of 3rd variables with both correlating variables.

In such situation, we control the effect of 3rd variable on both correlating variable.

Y = B.P of person
X_1 = Age of person
X_2 = Weight " "

$r_{Y1.2}$ = Partial correlation coefficient beth Y and X_1 after controlling for the effect of X_2

$$= \frac{r_{Y1} - r_{Y2} \cdot r_{12}}{\sqrt{1 - r^2_{Y2}}\sqrt{1 - r^2_{12}}}$$

$r_{Y2.1}$ = ---------- Y and X_2 ---- X_1

$$= \frac{r_{Y2} - r_{Y1} \cdot r_{12}}{\sqrt{1 - r^2_{Y1}}\sqrt{1 - r^2_{12}}}$$

$r_{12.Y}$ = ------ X_1 and X_2 after ---- Y

$$= \frac{r_{12} - r_{Y1} \cdot r_{Y2}}{\sqrt{1 - r^2_{Y1}}\sqrt{1 - r^2_{Y2}}}$$

$r_{Y1} = r_{1Y}$
$r_{Y2} = r_{2Y}$
$r_{12} = r_{21}$

$-1 \leq r_{Y1.2} \leq +1$
$-1 \leq r_{Y2.1} \leq +1$
$-1 \leq r_{12.Y} \leq +1$

$$r_{y1} = \frac{n\Sigma yx_1 - \Sigma y \cdot \Sigma x_1}{\sqrt{n\Sigma y^2 - (\Sigma y)^2}\sqrt{n\Sigma x_1^2 - (\Sigma x_1)^2}}$$

$$r_{y2} = \frac{n\Sigma yx_2 - \Sigma y \cdot \Sigma x_2}{\sqrt{n\Sigma y^2 - (\Sigma y)^2}\sqrt{n\Sigma x_2^2 - (\Sigma x_2)^2}}$$

$$r_{12} = \frac{n\Sigma x_1 x_2 - \Sigma x_1 \cdot \Sigma x_2}{\sqrt{n\Sigma x_1^2 - (\Sigma x_1)^2}\sqrt{n\Sigma x_2^2 - (\Sigma x_2)^2}}$$

Numerical

The salaries of IT workers are expected to be dependent on years of schooling and yrs. of work experience. The following table gives information on the annual salary (in thousands. of dollar) for 12 persons.

S.N	Ann. Salary (in thds. of dollars)	Years of schooling	Years of experience
1.	52	16	6
2.	44	12	10
3.	48	13	15
4.	77	20	3
5.	68	18	11
6.	48	16	2
7.	59	14	12
8.	83	18	4
9.	28	12	6
10.	61	16	9
11.	27	12	2

12. 69 16 18

(a) Compute simple correlation coefficient.
(b) Compute partial correlation coefficient & interpret the results.

Soln:-

Here, Y = Annual Salary
X_1 = Yrs. of schooling
X_2 = Yrs. of experience

Computational Result

$\Sigma Y_i = 634$	$\Sigma Y^2 = 36086$	$\Sigma Y X_1 = 10036$
$\Sigma X_1 = 183$	$\Sigma X_1^2 = 2869$	$\Sigma Y X_2 = 5865$
$\Sigma X_2 = 103$	$\Sigma X_2^2 = 1155$	$\Sigma X_1 X_2 = 1569$
		$n = 12$

Computational sample correlation coefficient

$$r_{Y_i} = \frac{n\Sigma Y X_1 - \Sigma Y \Sigma X_1}{\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}}$$

$$= \frac{12 \times 10036 - 634 \times 183}{\sqrt{12 \times 36086 - (634)^2}\sqrt{12 \times 2869 - (183)^2}}$$

$$= 0.8164$$

$$r_{Y_2} = \frac{n\Sigma Y X_2 - \Sigma Y \Sigma X_2}{\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}\sqrt{n\Sigma X_2 - (\Sigma X_2)^2}}$$

$$= 0.505$$

$$r_{12} = \frac{n\Sigma X_1 X_2 - \Sigma X_1 \Sigma X_2}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}\sqrt{n\Sigma X_2 - (\Sigma X_2)^2}}$$

$$= -0.0120$$

$$r_{y_1 \cdot 2} = \frac{r_{y_1} - r_{y_2} \cdot r_{12}}{\sqrt{1 - r_{y_2}^2}\sqrt{1 - r_{12}^2}}$$

$$= \frac{0.816 - 0.505 \times (-0.0120)}{\sqrt{1-(0.505)^2}\sqrt{1-(-0.0120)^2}} = 0.952$$

$$r_{y_2 \cdot 1} = \frac{r_{y_2} - r_{y_1} \cdot r_{12}}{\sqrt{1 - r_{y_1}^2}\sqrt{1 - r_{12}^2}}$$

$$= \frac{0.505 - 0.8164(-0.0120)}{\sqrt{1-(0.8164)^2}\sqrt{1-(0.0120)^2}} = 0.891$$

$$r_{12 \cdot y} = \frac{r_{12} - r_{y_1} \cdot r_{y_2}}{\sqrt{1 - r_{y_1}^2}\sqrt{1 - r_{y_2}^2}}$$

$$= \frac{-0.0120 - 0.8164 \times 0.505}{\sqrt{1-(0.8164)^2}\sqrt{1-(0.505)^2}} = -0.849$$

Simple Correlation	Partial correlation
$r_{y_1} = 0.816$	$r_{y_1 \cdot 2} = 0.952$
$r_{y_2} = 0.505$	$r_{y_2 \cdot 1} = 0.891$
$r_{12} = -0.0120$	$r_{12 \cdot y} = -0.849$

Interpretation :-

1. There is a very strong positive correlation betn annual salary and years of education after controlling for years of experience i.e. among the people with same yrs. of experience.

2. There is a very strong positive correlation betn annual salary and yrs. of experience after controlling for the effect of yrs. of education i.e. among the people with same yrs. of education.

3. There is a strong negative correlation betn yrs. of education and yrs. of schooling after controlling for the effect of Annual salary. It means that people with less education has to work for longer period in order to get the same salary as the people with higher degree or more yrs. of education.

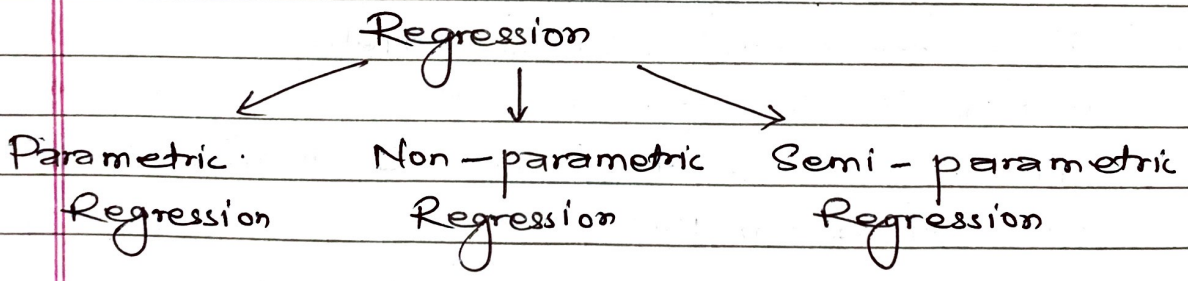
Multiple linear regression

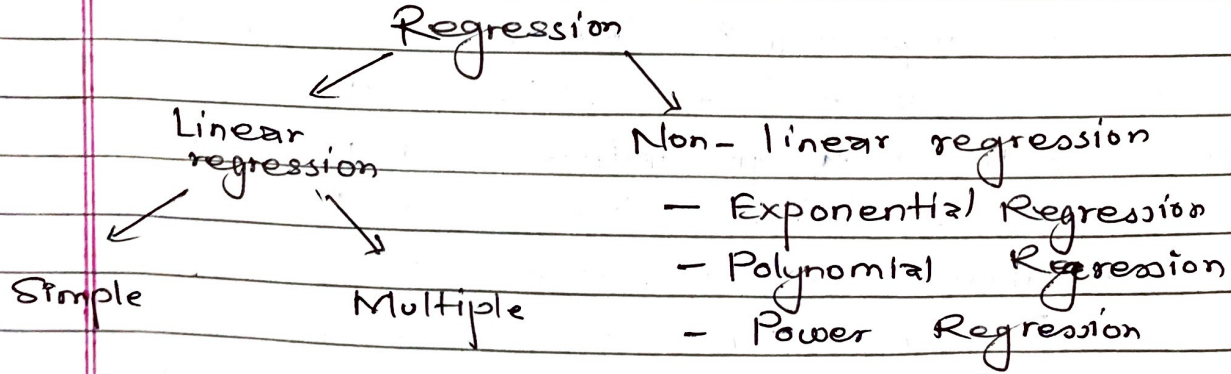
In multiple linear regression, we assume that there is a linear relationship between dependent variable Y and set of 'p' independent variables $X_1, X_2, \ldots\ldots, X_p$

Multiple – More than one IV.

Linear – Linear relnship

The main advantage of multiple linear regression is that it allows us to estimate or predict that values of dependent with more informations from the several independent variables.





Model

The multiple linear regression model with 'p' independent variable is given by,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p + e$$

The multiple linear linear regression model with two independent variable is given by,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

where,

Y = Dependent Variable

X_1 = First Independent variable

X_2 = Second Independent variable

β_0 = Popⁿ y-intercept

β_1 = Popⁿ slope of Y with X_1 holding effect of X_2 constant

β_2 = Popⁿ slope of Y with X_2 holding effect of X_1 constant

e = Error of prediction

Observed y Estimated

$$Y = Y - \hat{Y}$$

Deterministic Model — No error

Statistical Model — with error

Equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta} X_1 + \hat{\beta}_2 X_2 \quad , e = Y - \hat{Y}$$

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2$$

$\downarrow \quad \downarrow \quad \downarrow$

$\beta_0 \quad \beta_1 \quad \beta_2$

$$E(\hat{\beta}_0) = E(b_0) = \beta_0$$
$$E(\hat{\beta}_1) = E(b_1) = \beta_1$$
$$E(\hat{\beta}_2) = E(b_2) = \beta_2$$

where,

$\hat{Y}_i$ = Estimated value of dependent variable

X_{i1} = first independent variable

X_{i2} = Second independent variable

b_0 = Sample Y-intercept

b_1 = Sample slope of Y on X_1, holding the effect of X_2 constant

b_2 = " " " Y on X_2 " " "
" X_1 "

Interpretation of b_0

b_0 measures the average value of Y for $X_1 = 0$, $X_2 = 0$, it can be either +ve or -ve.

Interpretation of b_1

b_1 measure the average value of y for unit change in X_1 keeping the effect of X_2 constant. It can be either positive or negative.

Interpretation of b_2

The b_2 measures the average values of Y for unit change in X_2 keeping the effect of X_2 constant. It can be either +ve or -ve.

Estimating the regression coefficients

The values of regression coefficient $\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_1$ $\hat{\beta}_2 = b_2$ are not known; and they are estimated from sample data.

The two common methods for estimating the regression coefficients are :-

(1) Ordinary least square method (OLS method)

(2) Maximum likelihood estimation method (MLE method)

Ordinary least - square method

The principle behind least square method is to find the optimum value of b_0, b_1 and b_2 so that the sum of residuals or error is minimum.

Residual or Error$(e_i) = Y_i - \hat{Y}_i$, $i = 1, 2, \ldots, n$

Let $S = \sum_{i=1}^{n} e_i^2$

$$= \sum_{i=1}^{n} \{Y_i - \hat{Y}_i\}^2$$

$$= \sum_{i=1}^{n} \{Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2}\}^2$$

The least square method minimum s function partially differentiating function wrt b_0, we get,

$$\frac{\partial s}{\partial b_0} = 0 = \frac{\partial}{\partial b_0}\left\{\sum_{i=1}^{n} y_i - b_0 - b_1 x_{i1} - b_2 x_{i2}\right\}^2 = 0$$

$$\Rightarrow \Sigma y_i = n b_0 + b_1 \Sigma i_1 + b_2 \Sigma x_{i2}$$

Partially diffn s wrt b_1, we get,

$$\frac{\partial s}{\partial b_1} = 0 \Rightarrow \sum_{i=1}^{n} y_i x_{i1} = b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} i_1 x_{i2}$$

Finally, partially diffn s wrt b_2, we get,

$$\frac{\partial s}{\partial b_2} = 0 \Rightarrow \sum_{i=1}^{n} y_i x_{i2} = b_0 \sum_{i=1}^{n} x_{i2} + b_1 \sum_{i=1}^{n} x_{i1} x_{i2} + b_2 \sum_{i=1}^{n} x_{i2}^2$$

∴ The normal eqns of estimating b_0, b_1 & b_2 are,

$$\Sigma y_i = n b_0 + b_1 \Sigma x_{i1} + b_2 \Sigma x_{i2}$$
$$\Sigma y_i x_{i1} = b_0 \Sigma x_{i1} + b_1 \Sigma x_{i1}^2 + b_2 \Sigma x_{i1} x_{i2}$$
$$\Sigma y_i x_{i2} = b_0 \Sigma x_{i2} + b_1 \Sigma x_{i1} x_{i2} + b_2 \Sigma x_{i2}^2$$

Matrix form

$$\begin{bmatrix} \Sigma y_i \\ \Sigma y_i x_{i1} \\ \Sigma y_i x_{i2} \end{bmatrix} = \begin{bmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i1} x_{i2} \\ \Sigma x_{i2} & \Sigma x_{i1} x_{i2} & \Sigma x_{i2}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

#Regression modelling of annual salary date

→ y = Annual salary (x 1000 $)
x_1 = years of schooling
x_2 = years of education

The regression eqn of salary on yrs of schooling and yrs. of experience is,

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} \longrightarrow (I)$$

Computational Result

$\Sigma y = 634$ $\quad \Sigma Y^2 = 36086$ $\quad \Sigma Y X_1 = 10036$

$\Sigma X_i = 183$ $\quad \Sigma X_1^2 = 2869$ $\quad \Sigma X X_2 = 5865$

$\Sigma X_2 = 102$ $\quad \Sigma X_2^2 = 1155$ $\quad \Sigma X_1 X_2 = 1569$

$n = 12$

∴ The normal eqns for the estimating b_0, b_1 and b_2 are,

$$\Sigma Y_i = n b_0 + b_1 \Sigma X_{i1} + b_2 \Sigma X_{i2} \quad \text{---- (II)}$$

$$\Sigma Y_i X_{i1} = b_0 \Sigma X_{i1} + b_1 \Sigma X_{i1}^2 + b_2 \Sigma X_{i1} X_{i2} \quad \text{--- (III)}$$

$$\Sigma Y_i X_{i2} = b_0 \Sigma X_{i2} + b_1 \Sigma X_{i1} X_{i2} + b_2 \Sigma i_2^2 \quad \text{--- (IV)}$$

Substituting the value in the above eqn we get,

$$634 = 12 b_0 + 183 b_1 + 109 b_2 \longrightarrow (V)$$

$$10036 = 183 b_0 + 2869 b_1 + 1569 b_2 \longrightarrow (VI)$$

$$5865 = 103 b_0 + 1569 b_1 + 1159 b_2 \longrightarrow (VII)$$

Solving (V), (VI) & (VII), we get,

$b_0 = -33.005$

$b_1 = 4.7321$

$b_2 = 1.5925$

The fitted eqn is,

$$\hat{Y} = -33.005 + 4.7321 X_1 + 1.5923 X_2$$

Estimated annual salary = −33.0005 + 4.7321 x year of schooling + 1.5925 x year of experience

Interpretation $b_0 = -33.005$

When $X_1 = 0, X_2 = 0$, $\hat{y} = -33.0005$ (not interpretable)

Interpretations of $b_1 = 4.7321$

For every increment of yrs. of schooling by an amount of 1 yrs, the annual salary increases in average by an amount of 4.7321 (x 1000 $) = $4732.1, provided that the effect of yrs of experience is kept constant.

Interpretations of $b_2 = 1.5925$

For every increment of yrs. of schooling by an amount of 1 yrs., the annual salary increases in average by an amount of 1.5925 (x 1000 $) = $1592.5, provided that the effect of yrs. of experience is kept constant.

Prediction

What is the likely annual salary for the person who has completed 14 yrs. of education and has 6 yrs. of experience?

Soln:-

$$\hat{y} = -33.0005 + 4.7321 X_1 + 1.5925 X_2$$

Put,

$X_1 = 14$, $X_2 = 6$

$$\hat{y} = -33.0005 + 4.7321 \times 14 + 1.5925 \times 6$$

$$= 42.8029 \text{ (x1000 \$)}$$

$$= \$ 42802.9$$

Alternate methods for estimating regression coefficient

1. Let,

$$W = \begin{vmatrix} 1 & r_{y1} & r_{y2} \\ r_{y1} & 1 & r_{12} \\ r_{y2} & r_{y1} & 1 \end{vmatrix}$$

$$b_1 = \frac{S_y}{S_1} \cdot \frac{\begin{vmatrix} r_{y1} & r_{12} \\ r_{12} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{12} \\ r_{12} & 1 \end{vmatrix}} \qquad b_2 = \frac{S_y}{S_2} \cdot \frac{\begin{vmatrix} 1 & r_{y1} \\ r_{12} & r_{y2} \end{vmatrix}}{\begin{vmatrix} 1 & r_{12} \\ r_{12} & 1 \end{vmatrix}}$$

S_y = S.D of Y
S_1 = S.D of X_1
S_2 = S.D of X_2

2. Matrix Approach :-

Let $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$ $\quad \beta = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}_{3 \times 1}$

$$\beta = (X'X)^{-1}(X'Y)$$

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix}_{n \times 3}$$

$$\beta = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}_{3\times 1}$$

Evaluation of filled equation:-

The filled equation of multiple linear regression must be estimated before it is used for prediction purposes.

The following are the methods of calculating filled equations:-

(1) Multiple correlation of coefficient
(2) Coefficient of multiple determination
(3) Standard Error of estimate
(4) F-test for whole world
(5) t-test for individual slope
(6) Residual Analysis.

Multiple Correlation Coefficient

The correlation between observed Y values with its estimated values $\hat{Y}$ as given by multiple linear regression eqn $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$ is called multiple correlation coefficient. It is denoted by $R_{y.12}$

$$R_{y.12} = \frac{Cov(y,\hat{y})}{\sqrt{V(x)}\sqrt{V(\hat{y})}} = \frac{n\Sigma y\hat{y} - \Sigma y \cdot \Sigma \hat{y}}{\sqrt{n\Sigma y^2 - (\Sigma y)^2}\sqrt{n\Sigma \hat{y}^2 - (\Sigma \hat{y})^2}}$$

Alternatively,

$$R_{y.12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}\cdot r_{y2}\cdot r_{12}}{1 - r_{12}^2}}$$

$0 \leq R_{y.12} \leq 1$ | High value of $R_{y.12}$ indicates good fit

1. $R_{y.12} = 1$ — 100% fit
2. $R_{y.12} = 0$ — No fit
3. $R_{y.12} \to 1$ — Good fit
4. $R_{y.12} \to 0$ — Poor fit

2. Coefficient of multiple determination

then,

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \text{Total sum of squares}$$

The SST can be split into two components :-

1. SSR
2. SSE

$SST = SSR + SSE$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \text{Sum of squares due to regression}$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \text{Sum of squares due to errors}$$

If regression fits well, large portion SST is due to SSR. Then, ratio of SSR to the SST is a good measure of aptness of the fitted eqn.

Coefficient of determination is given by,

$$R^2_{Y.12} = \frac{SSR}{SST} \quad \Big| \quad 0 < R^2_{Y.12} \leq 1$$

SST, SSR, SSE — Given in exam

Again,

$$SST = SSR + SSE$$

$$\therefore R^2_{Y.12} = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$\boxed{R^2_{Y.12} = 1 - \frac{SSE}{SST}}$$

Alternatively,

$$R^2_{Y.12} = \frac{r^2_{Y1} + r^2_{Y2} - r_{Y1} \cdot r_{Y2} \cdot r_{12}}{1 - r^2_{12}}$$

Y = Annual salary
X_1 = Yrs. of schooling
X_2 = Yrs. of experience

$$\hat{Y} = -33.0085 + 4.7321 X_1 + 1.5925 X_2$$

$r_{Y1} = 0.816$ $\quad r_{Y2} = 0.505$ $\quad r_{12} = -0.012$

$$R^2_{Y.12} = 0.926 = 92.6\%$$

About 93% variation in annual salary explained by variation in yrs. of schooling and yrs of experiences. ie:

the linear relⁿ betⁿ annual salary, yrs. of schooling and yrs. of experience. Remaining 7% variation in annual salary given by the factors/variables other than the yrs. of schooling and yrs. of experience.

3. Standard Error of Estimate

The standard error of estimate is the deviation of the residuals and it is denoted by,

$$S_e = \sqrt{\frac{SSE}{n-p-1}}$$

$$= \sqrt{\frac{SSE}{n-2-1}}$$

$$= \sqrt{\frac{SSE}{n-3}}$$

$p = 2$ In our study (X_1, X_2)

SSE = Error sum of squares

$$= \sum_{i=1}^{n} (X_i - \hat{X}_i)$$

Low value of S_e indicates good fit and high value of S_e indicates poor fit.

$$0 \leq S_e \leq \infty$$

$S_e = 0$ perfect fit 100% fit

S_e measures the variability of points around the regression plane.

4. Overall or global test of model accuracy :-

The test evaluate the fitted eqn as whole i.e. it determines whether there is linear rel^{n}ship beth dependent variable Y and independent X_1 and X_2.

$H_0: \beta_1 = \beta_2 = 0$ (There is no rel^n beth Y, X_1 and X_2)

$H_1: \beta_1 \neq 0$ (for at least one $j = 1, 2$)

Test Statistics

The test statistic is given by,

$$F = \frac{MSR}{MSE}$$

MSR = Variable due to regression

$$= \frac{SSR}{P}$$

$$= \frac{SSR}{2}$$

p = No. of independent variable in the model

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

MSE = Variance due to error

$$= \frac{SSE}{n-p-1}$$

$$= \frac{SSE}{n-3}$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ANOVA Table

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F - ratio (Cal F)
Regression	$p = 2$	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$MSR = SSR/2$	$F = \frac{MSR}{MSE}$
Error	$n-p-1$ $= n-2-1$ $= n-3$	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-3}$	
Total	$n-1$	$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$\overline{MST}$	

Decision Rule:-

Reject H_0 if $cal \cdot F > F_\alpha \{ 2, n-3 \}$

2 → numerator d.f

$n-3$ → Denominator d.f

classmate
Date
Page

Q.

Y	X_1	X_2	$\hat{Y} = -33 + 4.7621 X_1 + 1.5925 X_2$
52	16	6	52.2686
44	12	10	39.7102
48	13	15	52.4048
77	20	8	74.3820
68	18	11	69.6953
48	16	2	45.8986
59	14	12	52.3594
53	18	4	58.5478
28	12	6	33.3402
61	16	9	57.0461
27	12	2	26.1702
69	16	18	71.3786

$(Y-\bar{Y})^2$	$(\hat{Y}-Y)^2$
0.0714596	0.3189
18.402384 04	172.2158
19.402263 04	0.1836
6.853924	464.3465
2.874042 09	284.3270
4.415881 96	48.0901
44.097568 36	0.2246
80.778084 84	32.6885
28.517736 04	379.9809
15.633325 22	17.7477
0.000888 04	668.8999
5.657737 96	343.9282

SSE = 176.7059

SSR = 2412.9187

SSR = 2412.9187
+ SSE = 176.7059
SST = 2589.6246

$$\bar{Y} = \frac{\Sigma Y}{n} = 52.833$$

$$R^2 = \frac{SSR}{SST}$$

$$= \frac{2412.9187}{2589.6246}$$

$$= 0.9317$$

$$= 93.17\%$$

F - test for whole model
(Overall test of model accuracy)

Step 1:- Null and Alternative Hypothesis

$H_0 : \beta_1 = \beta_2 = 0$ (There is no linear rel^n bet^n Y, X_1 & X_2)
(X_1 and X_2 are not significant prediction for Y)

$H_1 : \beta_j \neq 0$ for at least one $j = 1, 2$
(At least one indep. variable is significant prediction for Y)

Step 2 : Choice of α for the test

$\alpha = 5\%$

Step 3: Test Statistic

$$F = \frac{MSR}{MSE}$$

, where $MSR = SSR/p \quad p = 2$

$MSE = SSE/n-p-1$

The test statistic follows F distn with $p = 2$ d.f in numerator and $n-p-1 = n-3$ d.f in denominator.

$$F \sim F_{\alpha}\{2, n-3\}$$

Step 4: Calculated F

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	Cal F
Regression	$p = 2$	SSR = 2,412.9182	$MSR = \frac{2412.91352}{2}$ =	$F = \frac{MSR}{MSE}$ = 1206.459 / 19.6339 = 61.4478
Error	$n-p-1$ = 12−2−1 = 9	SSE = 176.7059	$MSE = \frac{176.7059}{2}$	
Total	$n-1 = 12-1$ = 11	SST = 2589.6216		

∴ Cal F = 61.4478

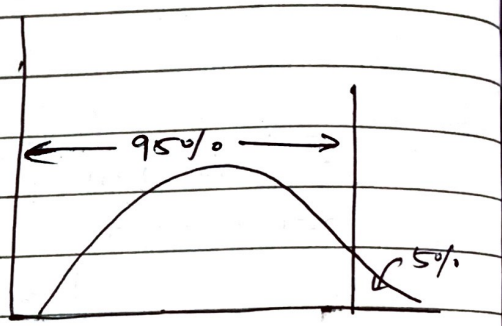
Step 5: Tab F

Numerator d.f = 2

Denominator d.f = 9

$\alpha = 5\%$

$F_{0.05}(2,9) = 4.26$



AR:- $F < 4.26$

RR:- $F \geq 4.26$

Step 6:- Statistical Decision

Since, cal F = 61.4478 $\gg$ Tab F = 4.26, we strongly rejected H_0 at 5% level of significance.

Step 7:- Conclusion

At least one slope is significant ie. at least one IV is significant prediction of Y.

t- test for individual slope

If f-test test of overall goodness of model reject H_0 then at least one IV is significant for Y. To know which I.V is significant for Y, we have to perform t-test for individual slope.

Hypothesis

$H_0: \beta_j = 0$ (X_j is not significant for Y)

$H_1: \beta_j \neq 0$ (X_j is significant for Y)

$j = 1, 2$

$j = 1 \quad X_1$

$j = 2 \quad X_2$

Test Statistic

$$t = \frac{b_j}{SE(b_j)} \quad \left| \quad \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right.$$

where,

b_j = Sample slope of Y on X_j holding other

$H_1: \beta_1 \neq 0$ (There is linear ~~~~
~~~~~~~~ i.e. x_1 is significant of y)

Step 2:- Choice of α

$\alpha = 5\%$

Step 3:- Test Statistic

$$t = \frac{b_1}{SE(b_1)} \qquad SE(b_1) = \frac{\sqrt{MSE}}{\sqrt{SSx_1}}$$

$$SSx_1 = \Sigma(x_1 - \bar{x_1})^2$$

Given

The test statistic has t dist^n with $n-p-1 = n-3$ d.f

Step 4:- Calculated F

$MSE = 19.6339$

$SSx_1 = \Sigma(x_1 - \bar{x_1})^2$

$\Sigma x_1 = 183$

$\bar{x_1} = 15.25$

$(x - \bar{x})^2$

$E(x_1 - \bar{x_1})^2 = 78.25$

$$S.E(b_1) = \frac{\sqrt{MSE}}{\sqrt{SSx_1}} = \frac{\sqrt{19.6339}}{\sqrt{78.25}}$$

$= 0.5009\ 11629$
~~~~~~~~

Now,

$$t = \frac{b_1}{SE(b_1)}$$

$$= \frac{4.7321}{0.5009}$$

$$= 9.447$$

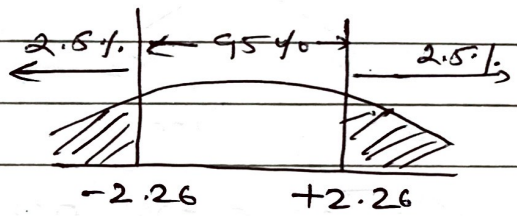
Step 5 :- Tabulated t

The test is two sided, and $\alpha = 5\%$

$$D.f = n - 3$$

$$= 12 - 3$$

$$= 9$$



AR :- $-2.26 < t < +2.26$

RR :- $t \geq 2.26$ or $t \leq 2.26$

Step 6 :- Statistical Decision

Since, cal $t = 9.447 >$ upper critical $t = 2.26$, we reject H_0.

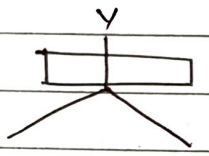
Step 7 :- Conclusion

The variable X_1 is significant prediction for Y.

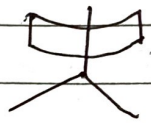
Assumption of multiple linear regression:-

1. Variables have non-zero variance.
2. Variables are measured without error.
3. There must be a linear relnship between dependent variable and independent variable.

Y, X_1, X_2



linear



Non-linear

4. The errors are normally and independently distributed.
5. Equal variance Assumption (Homosdestacity)
   This assumption states that the variance of error terms are similar across the values of the independent variables.
6. Non multicollinearity states the independent variables are not highly correlated with each other.

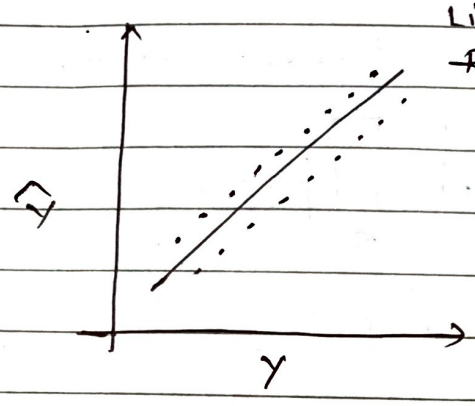
Residual Analysis:-

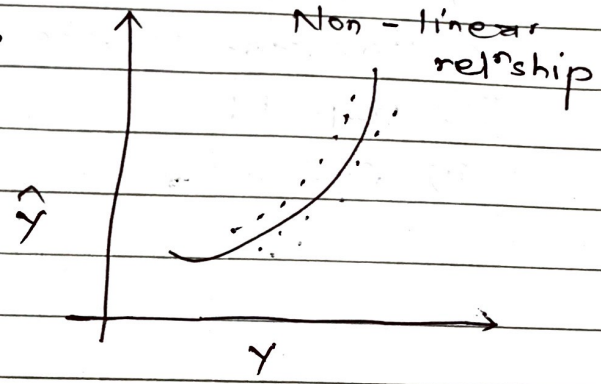
Linearity Assumption:-

This assumption is test by using following two graphs:-

(a) A graph of Y vs $\hat{Y}$

(b) A graph of e vs $\hat{Y}$



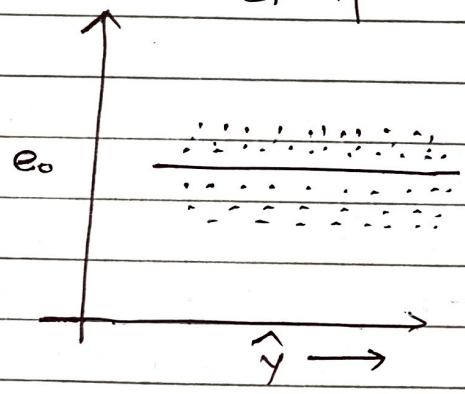


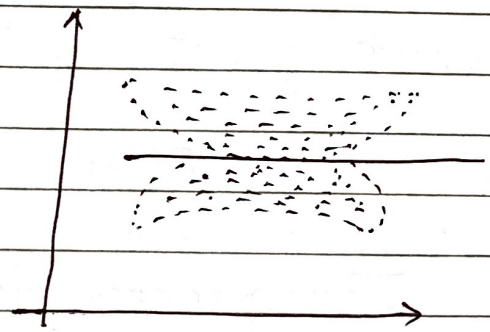
Homosedasticity

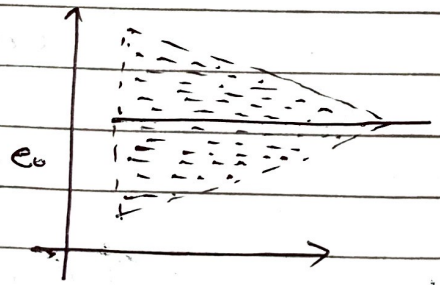
Linear Relnship

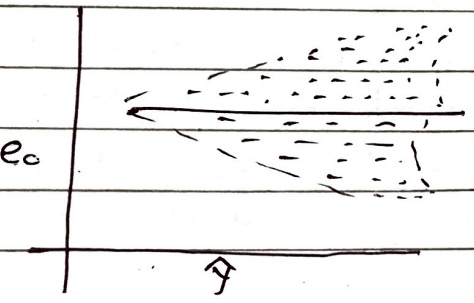
Non linear relnship

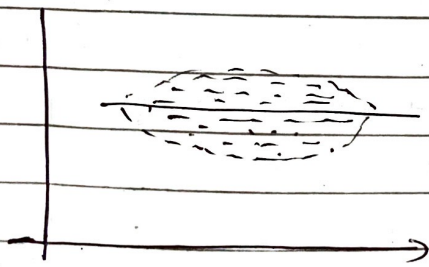
~~Heterodestacity~~ Se







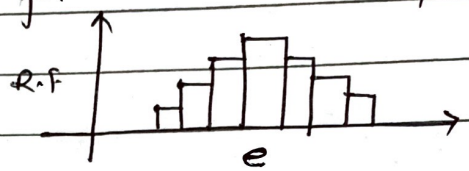


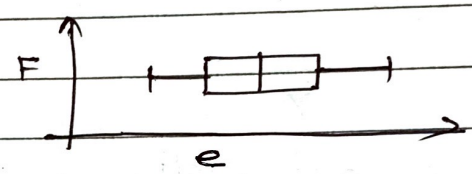


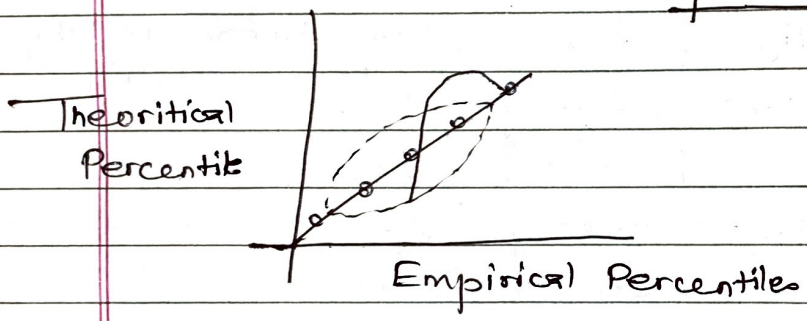
Normality of error:-

(1) Histogram
(2) Box and whisker plot
(3) p-p plot
(4) Q-Q plot

Normality of error

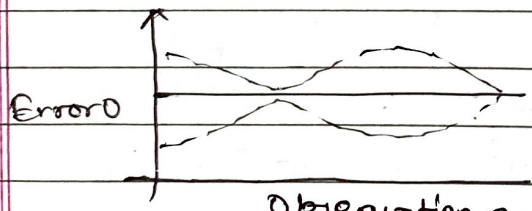


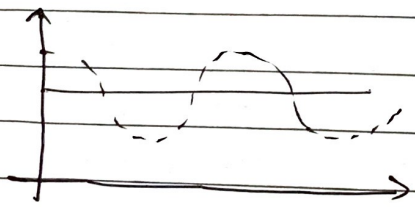




Independence of Errors:-

y-axis → Error
x-axis → Observation order





Autocorrelation or Serial correlation } Dependence of error

Non- multi collinearity

$Y = b_0 + b_1 X_1 + b_2 X_2$

SE

SE = high

Multicollinearity

It is desirable that is a strong correlation betn dependent variable and independent variable. Multicollinearity is the state in which there is a moderate to high inter correlations or interassociations amount the independent variables (predictions). If multi collinearity exists two or more predictors the realiability of the regression coefficients is reduced and therefore prediction and statistical inferences are not reliable.

Effects of multi collinearity

1. The standard error of regression coefficients are likely to be high. There will be change in signs as well magnitude of regression coefficients from sample to sample.

2. The estimated regression coefficients of any one variable depends on which other independent variables are included in the model.