INTRODUCTION TO ECONOMETRICS

Econometrics can be defined as measurement in economics. It therefore deals with modelling of the economy factors or elements e.g modelling demand and supply

Econometrics models can be classified into 2 depending on the nature of the data.

1. Regression models
2. Time series models

In this course we shall deal with regression models

Regression model

It is a mathematical equation that is used in forecasting such that the sum of the squared deviations of the estimated values from the actual values is minimised

Classification of Regression Models

a) Simple regression models. — This a regression model that contains only one independent variable e.g $y = a + bx$, $y = e^x$.

b) Multiple regression model — This a regression model with more than one independent variable eg $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots \beta_n x_n$

$y = e^{\beta_1 x_1, \beta_2 x_2}$.

c) Linear regression model - This is a regression model whose independent variable (s) power is one ie the degree of the independent variable (s) is one.

eg $y = \alpha + \beta x_1$ - simple linear regression

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n$ - multiple linear regression

d) Non-linear regression models - these includes all other regression models that are not linear ie logarithimic, exponential, quadratic etc.

In regression analysis we deal with;

a) Estimation of regression parameters.

The parameters are estimated such that the sum of the squared error term is minimised

b) Evaluation of the regression model

This involves analysing how well the estimated regression model fits the data from Eh it was obtained

Simple linear regression model

This is a regression model of the form $y = \alpha + \beta x$

$y = a + bx$

$$a = \frac{\Sigma y - b\Sigma x}{n}$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

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

$$\Sigma e^2 = \Sigma [y - (\alpha + \beta x)]^2$$

$$\Sigma e^2 = \Sigma [y - \alpha - \beta x]^2$$

Estimation of α

$$\frac{\delta \Sigma e^2}{\delta \alpha} = 2\Sigma (y - \alpha - \beta x)(-1)$$

$$\frac{0}{-2} = \frac{-2\Sigma (y - \alpha - \beta x)}{-2}$$

$$\Sigma (y - \alpha - \beta x) = 0$$

$$\Sigma y - n\alpha - \beta \Sigma x = 0$$

$$n\alpha = \Sigma y - \beta \Sigma x$$

$$\alpha = \frac{\Sigma y - \beta \Sigma x}{n}$$

We can estimate the parameters α and β such that we minimise on the sum of the square error where $\Sigma e^2 = \Sigma (y - \hat{y})^2$

We differentiate partially the sum of the squared term with respect to each of the parameters

Estimation of α is as shown above.

Estimation of β.

$$\Sigma e^2 = \Sigma (y - \alpha - \beta x)^2$$

$$\frac{\partial \Sigma e^2}{\partial \beta} = 2\Sigma (y - \alpha - \beta x)(-x)$$

$$\frac{0}{2} = \frac{-2}{2}\Sigma (y - \alpha - \beta x)(-x)$$

$$0 = \Sigma (y - \alpha - \beta x)(-x)$$

~~$0 = (\Sigma y - n\alpha - \beta \Sigma x)(\Sigma x)$~~

$$0 = \Sigma (-xy + x\alpha + \beta x^2)$$

$$0 = \Sigma xy + \alpha \Sigma x + \beta \Sigma x^2$$

but $\alpha = \frac{\Sigma y - \beta \Sigma x}{n}$

$$0 = -\Sigma xy + \left(\frac{\Sigma y - \beta \Sigma x}{n}\right)\Sigma x + \beta \Sigma x^2$$

$$0 = -n\Sigma xy + \Sigma x \Sigma y - \beta (\Sigma x)^2 + \beta n \Sigma x^2$$

$$\beta (\Sigma x)^2 - \beta n \Sigma x^2 = \Sigma x \Sigma y - n \Sigma xy$$

$$\beta (\Sigma x^2 -$$

+ - +

- + -

+ - +

$$0 = \Sigma xy - \alpha \Sigma x - \beta \Sigma x^2$$

but

$$\alpha = \frac{\Sigma y - \beta \Sigma x}{n}.$$

$$0 = \Sigma xy - \left(\frac{\Sigma y - \beta \Sigma x}{n}\right)\Sigma x - \beta \Sigma x^2$$

$$n\Sigma xy - \Sigma x \Sigma y + \beta(\Sigma x)^2 - n\beta \Sigma x^2 = 0.$$

$$n\beta \Sigma x^2 - \beta(\Sigma x)^2 = n\Sigma xy - \Sigma x \Sigma y$$

$$\beta(n\Sigma x^2 - (\Sigma x)^2 = n\Sigma xy - \Sigma x \Sigma y$$

$$\beta = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}.$$

Illustration.

The following data has been provided by the mgt accountant of ABC Ltd

Production Units x	Pdctn Cost. Sh'000' y	xy	x^2
14	376	5264	196
18	398	7164	324
22	409	8998	484
27	458	12366	729
28	437	~~12 288~~ 12 236	784
32	459	~~14 658~~ 14 688	1024
36	471	~~16 956~~ 16 956	1296
39	482	~~18 798~~ 18 798	1521
42	491	~~20 622~~ 20622	1764
45	519	23,355	2025
303	4500	140,447	10,147

Required

(i) A regression equation of the pdctn cost on production units.

(ii) Estimate the production cost for producing 65 units.

$$\beta = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{10 \times 140447 - 303 \times 4500}{10 \times 10147 - (303)^2}$$

$$= \frac{40970}{9661}$$

$$\beta = 4.24$$

$$\alpha = \frac{\Sigma y - \beta \Sigma x}{n}$$

$$= \frac{4500 - 4.24(303)}{10}$$

$$\alpha = 321.5$$

(i) $\hat{y} = 321.5 + 4.24x$

(ii) For 65 units

$\hat{y} = 321.5 + 4.24(65)$

$= 321.5 + 275.6$

$= \text{Ksh } 597,100$

Multiple Linear Regression

This is a regression model of the form

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots \beta_n x_n.$$

This course we shall start by analysing a multiple linear regression model with two independent variables of the form

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Estimation of parameters

The parameters estimation should be such that the sum of the squared error term is minimized.

$$\sum e^2 = \sum (\hat{Y} - Y)^2$$

$$= \left(\sum Y - \beta_0 - \beta_1 x_1 - \beta_2 x_2\right)^2$$

We need to differentiate the sum of the squared error term partially with respect to each of the ~~independent~~ va. parameters We shall have three eqns with three unknown, the eqns are then solved simultaneously to obtain the estimate of each parameters.

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum e^2 = \sum (Y - \hat{Y})^2$$

$$= \sum (Y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2))^2$$

$$= \sum (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)^2$$

$$\frac{\partial \sum e^2}{\partial \beta_0} = 2 \sum (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)(-1)$$

$$\frac{0}{2} = \frac{-2}{2} \sum (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2$$

$$\sum (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2) = 0$$

$$\sum y - n\beta_0 - \beta_1 \sum x_1 - \beta_2 \sum x_2 = 0$$

$$\sum y = n\beta_0 + \beta_1 \sum x_1 + \beta_2 \sum x_2 \quad \cdots (i)$$

$$\frac{\partial \sum e^2}{\partial B_1} = 2 \sum (y - B_0 - B_1 X_1 - B_2 X_2)(-X_1)$$

$$\frac{0}{-2} = \frac{-2}{-2} \sum X_1 (y - B_0 - B_1 X_1 - B_2 X_2)$$

$$0 = \sum X_1 y - B_0 \sum X_1 - B_1 \sum X_1^2 - B_2 \sum X_1 X_2$$

$$\sum X_1 y = B_0 \sum X_1 + B_1 \sum X_1^2 + B_2 \sum X_1 X_2 \quad \text{- - - (ii)}$$

$$\frac{\partial \sum e^2}{\partial B_2} = 2 \sum (y - B_0 - B_1 X_1 - B_2 X_2)(-X_2)$$

$$\frac{0}{-2} = \frac{-2}{-2} \sum X_2 (y - B_0 - B_1 X_1 - B_2 X_2)$$

$$0 = \sum X_2 y - B_0 \sum X_2 - B_1 \sum X_1 X_2 - B_2 \sum X_2^2$$

$$\sum X_2 y = B_0 \sum X_2 + B_1 \sum X_1 X_2 + B_2 \sum X_2^2 \quad \text{- - (iii)}$$

Writing the above 3 eqns in matrix ~~form~~.

$$\begin{bmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum X_1 y \\ \sum X_2 y \end{bmatrix}$$

The matrix can be solved using either ~~traverse m~~

i) inverse method
ii) Grammers rule

Cramers rule

$$B_0 = \frac{\begin{vmatrix} \sum y & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{vmatrix}}{\begin{vmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{vmatrix}}$$

$$B_2 = \frac{\begin{vmatrix} n & \sum y & \sum X_2 \\ \sum X_1 & \sum X_1 y & \sum X_1 X_2 \\ \sum X_2 & \sum X_2 y & \sum X_2^2 \end{vmatrix}}{\begin{vmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{vmatrix}}$$

$$\beta_2 = \frac{\begin{vmatrix} n & \Sigma x_1 & \Sigma y \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 y \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2 y \end{vmatrix}}{\begin{vmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2^2 \end{vmatrix}}$$

Solve either using
* Artistic method
* co factor method.

Example 2.

x_1	x_2	y	x_1x_2	x_1^2	x_2^2	x_1y	x_2y
2	3	18	6	4	9	36	54
4	8	12	32	16	64	48	96
6	12	15	72	36	144	90	180
8	4	24	32	64	16	192	96
10	1	32	10	100	1	320	32
12	5	25	60	144	25	300	125
14	6	29	84	196	36	406	174
16	7	41	112	256	49	656	287
18	9	48	162	324	81	864	432
20	11	55	220	400	121	1100	605
110	66	299	790	1540	546	4012	2081

i) Estimated regression eqn of y on x_1 and x_2
ii) Estimated value of y if $x_1 = 30$ and $x_2 = 17$

$$\Sigma y = n\beta_0 + \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2$$
$$\Sigma x_1 y = \beta_0 \Sigma x_1 + \beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2$$
$$\Sigma x_2 y = \beta_0 \Sigma x_2 + \beta_1 \Sigma x_1 x_2 + \beta_2 \Sigma x_2^2$$

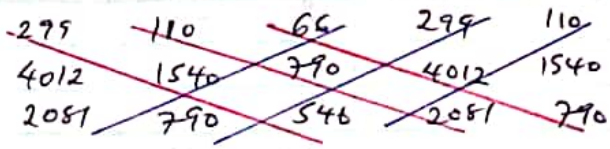
$$299 = 10\beta_0 + 110\beta_1 + 66\beta_2$$
$$4012 = 110\beta_0 + 1540\beta_1 + 790\beta_2$$
$$2081 = 66\beta_0 + 790\beta_1 + 546\beta_2$$

$$\begin{bmatrix} 10 & 110 & 66 \\ 110 & 1540 & 790 \\ 66 & 790 & 546 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 299 \\ 4012 \\ 2081 \end{bmatrix}$$

$$\beta_0 = \frac{\begin{vmatrix} 299 & 110 & 66 \\ 4012 & 1540 & 790 \\ 2081 & 790 & 546 \end{vmatrix}}{\begin{vmatrix} 10 & 110 & 66 \\ 110 & 1540 & 790 \\ 66 & 790 & 546 \end{vmatrix}}$$

Solving the numerator using artisti'c method.

$$\begin{vmatrix} 299 & 110 & 66 \\ 4012 & 1540 & 790 \\ 2081 & 790 & 546 \end{vmatrix} =$$

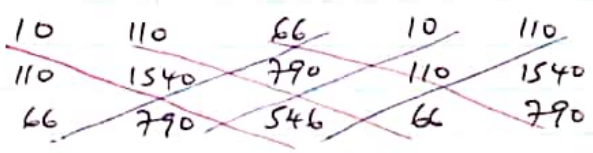


$= 251,411,160 + 180\,838\,900 + 209\,185,680$
$- (211,512\,840 + 186\,605\,900 + 240,960,720)$

$= 641\,435\,740 - 639\,079\,460$

$= 2,356\,280.$

$$\begin{vmatrix} 10 & 110 & 66 \\ 110 & 1540 & 790 \\ 66 & 790 & 546 \end{vmatrix} =$$



$= (8\,408\,400 + 5\,735\,400 + 5\,735\,400) - (6\,708\,240$
$+ 6\,241\,000 + 6,606,600)$
$= (19\,879\,200 - 19,555\,540)$

$= 323,360$

$$\beta_0 = \frac{2\,356\,280}{323\,360}$$

$$= 7.29.$$

$$\beta_1 = \frac{\begin{vmatrix} 10 & 299 & 66 \\ 110 & 4012 & 790 \\ 66 & 2081 & 546 \end{vmatrix}}{\begin{vmatrix} 10 & 110 & 66 \\ 110 & 1540 & 790 \\ 66 & 790 & 546 \end{vmatrix}} = \frac{10\,(546562) - 299\,(7920) + 66\,(-35\,882}{-110\,(7920) + 1540\,(1104) - 790\,(640)}$$

Using the co factor method. above.

$$= \frac{729\,328}{323\,360}$$

$$= 2.26$$

$$\beta_2 = \frac{\begin{vmatrix} 10 & 110 & 299 \\ 110 & 1540 & 4012 \\ 66 & 790 & 2051 \end{vmatrix}}{\begin{vmatrix} 10 & 110 & 66 \\ 110 & 540 & 790 \\ 66 & 790 & 546 \end{vmatrix}} = \frac{10(35\,260) - 110(-35\,882) + 299(-14\,740)}{66(-51\,260) - 790(640) + 546(-6700)}$$

$$= \frac{-107\,640}{323\,360}$$

$$= -0.33.$$

$$\therefore \hat{y} = 7.29 + 2.26x_1 - 0.33x_2.$$

ii) When $x_1 = 30$ $x_2 = 17$

$$y = 7.29 + 2.26(30) - 0.33(17)$$

$$= 2.29 + 67.8 - 5.61$$

$$= 69.48$$

Evaluation of Regression Model

This involves analysing how well the estimated regression model fits the data from ch it was obtained.

Evaluation of regression model can be classified into

i) Evaluation of the whole regression model
ii) Evaluation of b-co-efficient / slope.

Evaluation of regression model as a Whole

This involves determination of how well all the predictor variables in a regression model taken together predicts the depend, variable.

The methods used include :-

i) Standard error of estimate (S_e)
ii) Co efficient of determination (r^2)
iii) F-statistics

Standard error of estimate (S_e)

This is the average deviation of the estimated value from the actual value.

$$S_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-k}}$$

n - number of observations
k - number of parameters being estimated
n-k - d.f of error term

- The smaller the standard error the better the regression
- The std error of estimate is used in construction of confidence interval for the actual value of the predicted variable.

$$C.I = \hat{Y} \pm Z_{\alpha/2} S_e \quad \text{for} \quad n \geq 30.$$

$$\hat{Y} - Z_{\alpha/2} S_e \leq Y \leq \hat{Y} + Z_{\alpha/2} S_e.$$

or

$$CI = \hat{Y} \pm t_{\alpha/2} S_e$$

$$\hat{Y} - t_{\alpha/2} S_e \leq y \leq \hat{y} + t_{\alpha/2} S_e \quad \text{for} \quad n < 30.$$

Example 3.

Using the data in example 1

(i) Compute the std error of estimate

(ii) Construct a 95% Confidence Interval (C.I) for the production cost of 50 units

x	Y	$\hat{Y} = 321.5 + 4.24x$,	$(Y - \hat{Y})^2$
14	376	380.86	23.6196
18	378	397.82	0.0324
22	409	414.78	33.4054
27	458	435.98	484.8804
28	437	440.22	10.2400
32	459	457.18	3.3124
36	471	474.14	9.8596
39	482	486.86	23.6196
42	491	499.58	73.6164
45	519	512.3	44.89
			707.4788

i)

$$S_e = \sqrt{\frac{707.4788}{10-2}}$$

$$= \sqrt{88.43485}$$

$$= 9.40$$

$$V = df = n - k = 8$$

$$t_{0.975,8} = 2.306$$

ii) $\hat{Y} = 321.5 + 4.24(50)$

$= 533.5$

$$CI = \hat{y} \pm t_{\alpha/2} S_e$$

$$= 533.5 \pm 2.306 \times 9.4$$

$$= 533.5 \pm 21.6764$$

$$= 511.82 \leq Y \leq 555.19$$

Example 4.

Using the data in example 2

i) Compute the std error

ii) Construct a 98% CI for y if $x_1 = 20$ $x_2 = 25$

x_1	x_2	y	$\hat{Y} = 7.287 + 2.255x_1 - 0.333x_2$	$(y-\hat{y})^2$
2	3	18	10.798	51.869
4	8	12	13.643	2.699
6	12	15	16.821	3.316
8	4	24	23.995	0.000
10	1	32	29.504	6.230
12	5	25	32.682	59.013
14	6	29	36.859	61.764
16	7	41	41.036	0.001
18	9	48	44.880	9.734
20	11	55	48.724	39.388
				234.014

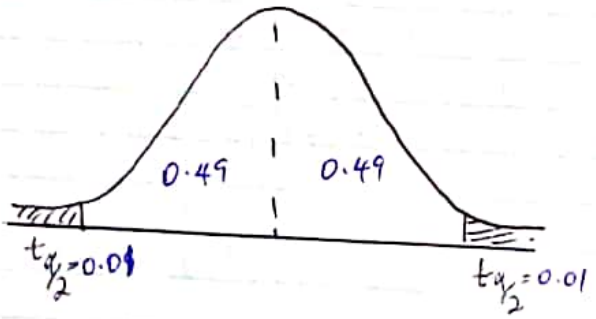
(i) $$S_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-k}} = \sqrt{\frac{234.014}{10-3}} = \sqrt{\frac{234.014}{7}}$$

$$= 5.782$$

(ii) $\hat{y} = 7.287 + 2.255x_1 - 0.333x_2$ but $x_1 = 20$, $x_2 = 25$

$= 7.287 + 2.255(20) - 0.333(25)$

$= 44.062$



$v = df = n-k$

$= 10-3$

$= 7$

$t_{0.99,7} = 2.998$

$$CI = \hat{y} \pm t_{\alpha/2} S_e$$

$= 44.062 \pm 2.998 \times 5.780$

$= 44.062 \pm 17.334$

$= 26.728 \leq y \leq 61.396$

06/02/2020

Coefficient of determination

Coefficient of determination is defined as the proportion of variations that is explained by the regression model
It can be computed in two different approaches

Sum of Squares

r^2 = coefficient of determination

Y – Actual
$\hat{Y}$ - estimated
$\bar{Y}$ - expected

Sum of square of Total Variation (SST)

$$SST = \sum (Y - \bar{Y})^2$$

Sum of squares of Error / Residual (SSE)

$$SSE = (Y - \hat{Y})^2$$

Sum of squares of regression model (SSR)

$$SSR = \sum (\hat{Y} - \bar{Y})^2$$

$$SST = SSR + SSE$$

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

~~Example 5~~

~~Using the data~~ The variations are summarized in a table known as analysis of variance (ANOVA) table.
The general format of the ANOVA table is as follows

Source of Variation	Sum of squares	degree of freedom	Mean sum of squares	F-Value
Regression	SSR	$k-1$	$MSR = \frac{SSR}{k-1}$	$\frac{SSR/k-1}{SSE/n-k} = \frac{MSR}{MSE}$
Error	SSE	$n-k$	$MSE = \frac{SSE}{n-k}$	
Total	SST	$n-1$		

Example 5

From the data in example 1,

(i) Construct ANOVA Table

(ii) Compute the coefficient of determination and comment on your results.

x	y	$\hat{y}$	SSE $(y-\hat{y})^2$	SSR $(\hat{y}-\bar{y})^2$	SST $(y-\bar{y})^2$
14	376	380.82	23.62	4780.34	5476
18	398	397.82	0.03	2722.75	2704
22	409	414.78	33.41	1240.45	1681
27	458	435.98	484.88	196.56	64
28	437	440.22	10.24	95.65	169
32	459	457.18	3.31	51.55	81
36	471	474.14	9.86	582.74	441
39	482	486.86	23.62	1358.66	1024
42	491	499.58	73.62	2458.18	1681
45	519	512.3	44.89	3881.29	4761
$\bar{y}=450$			707.48	17368.17	18,082

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

i)

Source of variation	Sum of squares	degree of freedom	Mean of squares	F - value
Regression	17 368.16	1	17 368.16	196.36
Error	707.60	8	88.45	
Total	18 082*	9		

ii) $r^2 = \frac{SSR}{SST} = \frac{17\,368.16}{18082} = 0.9605.$

$= 96.05\%.$

96.05% of the variations in production cost is explained by variations in productions units, the remaining proportion is explained by other factors ch include the error term.

Example 6

Using the data in example 2 Compute the coefficient of determination and comment on your results

x_1	x_2	y	$\hat{y}$	SSE $(y-\hat{y})^2$	SSR $(\hat{y}-\bar{y})^2$	SST $(y-\bar{y})^2$
2	3	18	10.80	51.87	364.81	141.61
4	8	12	13.64	2.70	264.39	320.41
6	12	15	16.82	3.32	171.09	222.01
8	4	24	24.	0.00	34.81	34.81
10	1	32	29.50	6.23	0.16	4.41
12	5	25	32.68	59.01	7.73	24.01
14	6	29	36.86	61.76	48.44	0.81
16	7	41	41.84	0.00	124.10	123.21
18	9	48	44.88	9.73	224.40	327.61
20	11	55	48.72	39.39	354.19	630.01
		$\bar{y}=29.9$		234.01	1594.12	1828.9

$$r^2 = \frac{SSR}{SST.} = \frac{1594.12}{1828.9} = 0.8720 = 87.20\%.$$

87.20% of the variations in y is explained by variations in x_1 and x_2, the remaining 12.8% is explained by other factors which include the error term.

F - statistics

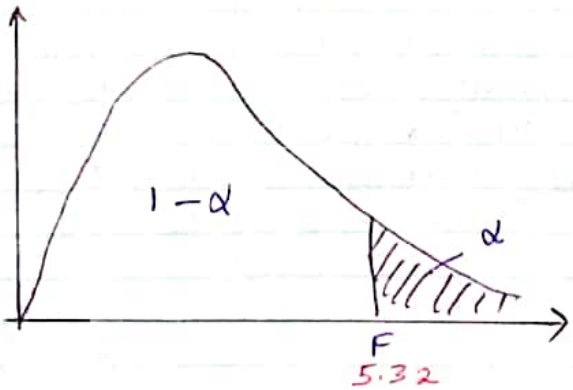
This statistics is used to carry out hypothesis testing on the significance of the regression model
The hypothesis is tested as follows: —

$H_0 : r^2 = 0$ Regression model is not statistically significant.
$H_a : r^2 \neq 0$. Regression model is statistically significant

F - statistic

$$F = \frac{SSR/k-1}{SSE/n-k}$$

F - Distribution
$F_{k-1, n-k}$



Example 7.

Test on the significance of the regression model obtained in example 1 at 5% significance level.

$H_0 : r^2 = 0$
$H_1 : r^2 \neq 0$.

$F_{k-1, n-k}$

$F_{1,8} = 5.32$

$$F = \frac{SSR/k-1}{SSE/n-k} = 196.36$$

* At 5% significance level we reject the null hypothesis and conclude that the regression model is statistically significant.

Example 8

Using the data in example 2 test whether the estimated regression model is statistically significant at 1% significance level.

$H_0 : r^2 = 0$

$H_1 : r^2 \neq 0$

F-distribution

$F_{k-1, n-k} = F_{2,7}$

$= 9.55$

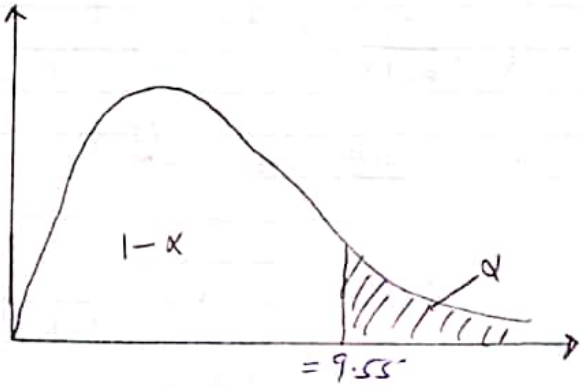
F-Statistic

$$F = \frac{SSR/k-1}{SSE/n-k}$$

$$= \frac{1594.12/3-1}{234.01/10-3}$$

$$= \frac{797.06}{33.43}$$

$$= 23.84$$



At 1% significance level we reject the null hypothesis and conclude that the regression model is statistically significant.

r^2 = power of the model.

Evaluation of the regression model using the slope / B-coefficient

In this case we evaluate how well a given independent variable predicts the dependent variable holding other factors constant

The following measures are used :-

i) Correlation coefficient $(r, 1)$
ii) The std error of the slope (S_b)
iii) t or Z statistic

Correlation coefficient $(r, 1)$

Correlation is the measure of the degree and nature of rlship btwn two variables.

Correlation is measured using correlation coefficient ch ranges btwn -1 and 1.

A negative correlation implys that if one of the variables is increasing the other variable is decreasing and vice versa.

A positive correlation implies that the two variables vary in the same direction ie, if one of the variables is increasing the other variable is also increasing and vice versa.

The nature of correlation can be obtained by checking the nature of the coefficient of the independent variable in the regression model.

eg

$$y = 4.2 - 3.6x_1 + 8.2x_2$$

x_1	x_2	y	
2	3	21.6	> 3.6
3	3	18.	

Interpretation of coefficients

-3.6 Holding other factors constant if x_1 was increased by 1 unit the estimated value of y would decrease by 3.6 units

8.2 This implies if x_2 was increased by 1 unit holding other factors constant the estimated value of y will increase by 8.2 units.

Std. error of the slope (S_b).

This is the average error term for the b coefficient.

$$S_b = \frac{S_e}{\sqrt{\Sigma (x - \bar{x})^2}}$$

The smaller the standard error the better the predictor variable in predicting the dependent variable.
The std error of the slope is used in construction Confidence Interval for the b-coefficient.

$$CI = \hat{b} \pm Z_{\alpha/2} S_b \quad \text{- for large sample size } n \geq 30$$

$$CI = \hat{b} \pm t_{\alpha/2} S_b \quad \text{- for small sample size}$$

$$\hat{b} - Z_{\alpha/2} S_b \leq b \leq \hat{b} + Z_{\alpha/2} S_b.$$

Example 9.

Construct a 95% Confidence interval for the b-coefficient using the data in example 1.

x	y	$(x - \bar{x})^2$
14	376	265.69
18	398	151.29
22	409	68.89
27	458	10.89
28	437	5.29
32	459	2.89
36	471	32.49
39	482	75.69
42	491	136.89
45	519	216.09
303		966.1

$\bar{x} = 30.3$

$$S_b = \frac{9.40}{\sqrt{966.1}} = \frac{9.40}{31.08} = 0.3024$$

$$CI = 4.24 \pm 1.96 \times 0.3024 \quad (2.306)$$

$$= 4.24 \pm 0.59$$

$$= 3.65 \leq b \leq 4.83$$

$$3.54 \leq b \leq 4.94$$

Example 10

Construct 99% c.I for the b coefficient of the estimated regression model in example 2.

X_1	X_2	y	$(X_1-\bar{X}_1)^2$		$(X_2-\bar{X}_2)^2$
2	3	18	102.01	81	12.96
4	8	12	65.61	49	1.96
6	12	15	37.21	25	20.16
8	4	24	16.81	9	6.76
10	1	32	4.41	1	31.36
12	5	25	0.01	1	2.56
14	6	29	3.61	9	0.36
16	7	41	15.21	25	0.16
18	9	48	34.81	49	5.76
20	11	55	62.41	81	19.36
$\bar{X}_1 = 11.0$	$\bar{X}_2 = 6.6$		342.10	330	110.40

$S_e = 5.782$

$$S_{b_1} = \frac{S_e}{\sqrt{\Sigma(X_1-\bar{X}_1)^2}}$$

$$= \frac{5.782}{\sqrt{342.1}} \quad \frac{5.782}{\sqrt{330}}$$

$$= 0.313 \quad = 0.318$$

$$CI = \hat{b}_1 \pm t_{\alpha/2} S_{b_1}$$

$$= 2.26 \pm 2.576 \times 0.313$$

$$= 2.26 \pm 0.81$$

$$1.45 \leq b_1 \leq 3.07$$

$$= 2.26 \pm 3.499 \times 0.38$$

$$= 2.26 \pm 1.11$$

$$= 1.15 \leq b_1 \leq 3.37$$

$$S_{b_2} = \frac{S_e}{\sqrt{\Sigma(X_2-\bar{X}_2)^2}}$$

$$= \frac{5.782}{\sqrt{110.40}}$$

$$= 0.550$$

$$CI = \hat{b}_2 \pm t_{\alpha/2} S_{b_2}$$

$$= -0.33 \pm 2.576 \times 0.55$$

$$= -0.33 \pm 1.42$$

$$= -1.75 \leq b_2 \leq 1.09$$

$$= -0.33 \pm 3.499 \times 0.55$$

$$= -0.33 \pm 1.92$$

$$= -2.25 \leq b_2 \leq 1.59$$

13/02/2020

3. t or Z Statistics

This statistic is used to test whether a given predictor variable is statistically significant in predicting the dependent variable

$$t \text{ or } Z = \frac{b}{S_b}$$

The hypothesis is stated as follows :-

$H_0 : b = 0$ (predictor variable is not statistically significant)

$H_1 : b \neq 0$ (predictor variable is statistically significant).

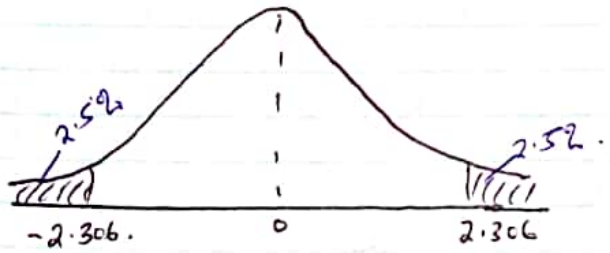
Example 11

At 5% significance level test whether production units is a statistically significant predictor variable for production cost using the data in example 1.

$H_0 : \beta = 0$ (prdn units is not a statistically significant predictor variable for prdn cost)

$H_1 : \beta \neq 0$ (prdn units is a statistically significant predictor variable for prdn cost).

$$t = \frac{b}{S_b} = \frac{4.24}{0.3024} = 14.02.$$



$t_{\alpha/2} = t_{0.975,8} = 2.306$

Conclusion: At 5% significance level we reject the null hypothesis and conclude that prdn units is a statistically significant predictor variable for prdn cost.

Example 12

At 1% significance level test whether x_1 and x_2 are significant predictor variables for y using the data in example 2.

$H_0 : \beta_1 = 0$ (x_1 is not statistically significant predictor of y)

$H_1 : \beta_1 \neq 0$ (x_1 is statistically significant predictor of y)

$$t = \frac{b_1}{S_{b_1}} = \frac{2.26}{0.318} = 7.11$$

$t_{\alpha/2} = t_{0.995,7} = 3.499$ ~~2.998~~

At 1% significance level we reject the null hypothesis and conclude that x_1 is statistically significant predictor of y.

$H_0: \beta_2 = 0$ (x_2 is not a statistically significant predictor of y

$H_1: \beta_2 \neq 0$ (x_2 is statistically significant predictor of y)

$$t_2 = \frac{\beta_2}{S_{\beta_2}} = \frac{-0.33}{0.55} = -0.6$$

$t_{\alpha/2} = 3.499$.

At 1% significance level we fail to reject the null hypothesis and conclude that x_2 is not statistically significant predictor of y.

Example Question

A study was conducted to examine the predictive ability of economists and finance specialists on their future performance of their country's economy. Equal random samples of economist and finance specialists were asked to make predictions of the percentage change in the gross domestic product GDP of their countries. A year later the actual percentage change in GDP (y) was regressed against the economists prediction (x_1) and the finance specialists prediction x_2 using a statistical software. Below is part of the output:

Parameters estimates

Variable	Co-efficient	Standard Error	t-ratio
Constant	0.983	A	0.34
x_1	B	0.257	2.42
x_2	0.587	0.225	2.61

Analysis of Variance (ANOVA)

Source	Degrees of Freedom	Sum of Squares	Mean of Sum of squares	F-value
Regression	2	2904.8	D	E
Error	11	100.1	9.1	
Total	13	C		

Required

i) Compute the values A, B, C, D, E — 5mks

ii) The estimated regression model — 2mks

iii) The explanatory power of the model 3mks

iv) Test the significance of each of the predictor variables at 5% significance level (6mks)

v Comment on the statistical significance of the model at 5% significance level.

(i)

A

$t = \frac{b}{S_b}$

$S_b = \frac{b}{t}$

$= \frac{0.983}{0.34}$

$= 2.891$

B

$b = t \times S_b$

$= 2.42 \times 0.257$

$= 0.622$

C

$C = SST = SSR + SSE$

$= 2904.8 + 100.1$

$= 3004.9$

D

$MSR = \frac{SSR}{k-1}$

$= \frac{2904.8}{2}$

$= 1452.4$

$F = \frac{MSR}{MSE} = \frac{1452.4}{9.1} = 159.60.$

ii $\hat{y} = 0.983 + 0.622x_1 + 0.587x_2.$

iii) $r^2 = \frac{SSR}{SST} \times 100 = \frac{2904.8}{3004.9} \times 100 = 96.67\%$

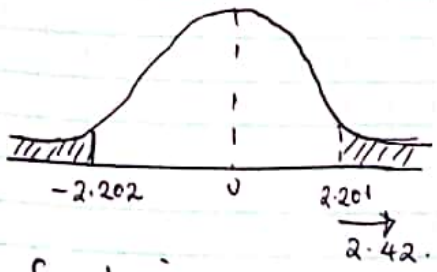
96.67% of the change in GDP is explained by the economists and finance specialists prediction while the remaining proportion is explained by other factors including the error term.

iv

For x_1

$t_{tables} = 2.201$ $H_0: x_1 = 0$

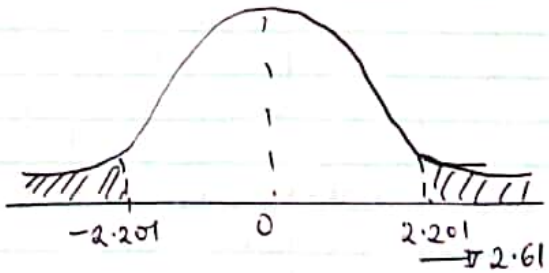
$t_c = 2.42.$ $H_1: x_1 \neq 0$



For x_2

$t_{tables} = 2.201$ $H_0: x_2 = 0$

$t_c = 2.61$ $H_1: x_2 \neq 0$



Conclusion

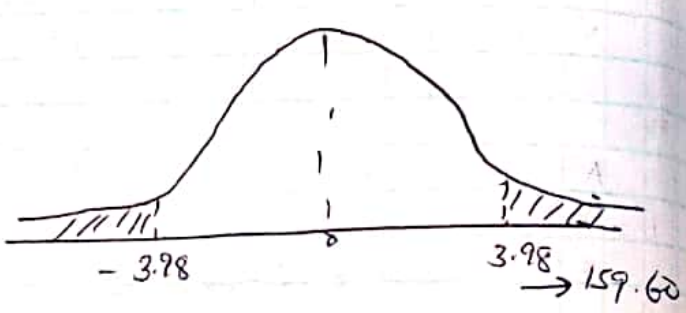
At 5% significance level, we reject the null hypothesis and conclude that the economists prediction is a statistically significant predictor variable of GDP

At 5% significance level, we reject the null hypothesis and conclude that the finance specialists prediction is a statistically significant predictor variable of GDP.

$H_0: r = 0$ (regression model is not statistically significant)
$H_1: r \neq 0$ (regression model is statistically significant)

$f_{k-1, n-k} = f_{2,9} = 3.98$

$F_c = 159.60$



At 5% significance level we reject the null hypothesis and conclude that the regression model is statistically significant.

Question 2.

KCA University research department plans to do an investigation on the impact of students' final score and their linguistic aptitude at entry point. For the purposes of this survey, the following sample data has been assembled for 12 subject areas under investigation. The University applies a 5% significance level in this kind of statistical analysis.

Subject	Linguistic aptitude	Final Score
1	5	5
2	10	20
3	6	4
4	8	15
5	4	11
6	4	9
7	3	12
8	10	18
9	2	7
10	6	2
11	7	14
12	9	17

Required:

i) Obtain the estimated regression model
ii) Comment on the meaning of the coefficient values
iii) Comment on the significance of linguistic aptitude in predicting the final score
iv) Is the regression model statistically significant.

- $y = a + bx$

$$a = \frac{\Sigma y - b \Sigma x}{n} \qquad b = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$

x	y	$(x-\bar{x})^2$	xy	x^2	$\hat{y}$	$(y-\hat{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\bar{y})^2$
5	5	5.76	25	25	9.46	19.89	2.72	38.07
10	20	6.76	200	100	16.71	~~48.82~~	30.69	77.97
6	4	1.96	24	36	10.91	~~47.75~~	0.07	51.41
8	15	0.36	120	64	13.81	~~33.72~~	6.97	14.67
4	11	11.56	44	16	8.01	~~18.84~~	9.99	0.03
4	9	11.56	36	16	8.01	~~16.82~~	9.99	4.71
3	12	19.36	36	9	6.56	29.59	20.34	0.69
10	18	6.76	180	100	16.71	1.66	30.69	46.65
2	7	29.16	14	4	5.11	3.57	36.72	17.39
6	2	1.96	12	36	10.91	79.39	0.07	84.09
7	14	0.16	98	49	12.36	2.69	1.42	8.01
9	17	2.56	153	81	15.26	3.03	16.73	33.99
74	134	~~5.07~~ 6.925	942	536		209.77	166.6	377.68

$$b = \frac{12 \times 942 - 74 \times 134}{12 \times 536 - 74^2}$$

$$= 1.45$$

$$a = \frac{134 - 1.45(74)}{12}$$

$$= 2.21$$

i) $\hat{y} = 2.21 + 1.45x$

The minimum final score is 2.21

ii) Coefficient of x – 1.45 → When the value of lingnistic aptitude is increased by one the final score is increased by 1.45.

iii) $H_0: b = 0$ (linguistic aptitude is not a statistically significant predictor variable of the final score

$H_1: b \neq 0$ (linguistic aptitude is statistically significant predictor variable of the final score)

$$S_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-k}}$$

$$= \sqrt{\frac{209.77}{12-2}}$$

$$= 4.58$$

$$S_b = \frac{S_e}{\sqrt{\Sigma(x-\bar{x})^2}}$$

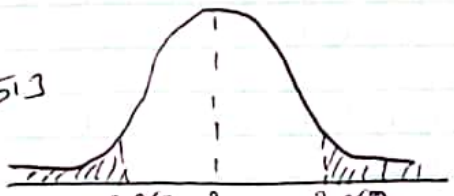
$$= \frac{4.58}{8.93}$$

$$= 0.513$$

$$t = \frac{b}{S_b}$$

$$= \frac{1.45}{0.513}$$

$$= \frac{2.734}{2.82}$$



$t_{0.975,9} = 2.262$

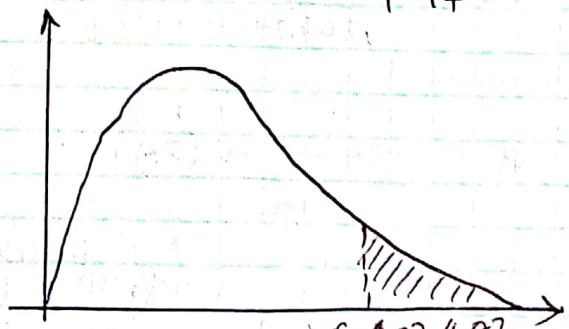
At 5% significance level, we reject the null hypothesis and conclude that linguistic aptitude is statistically significant predictor variable of the final score.

iv) $H_0: r^2 = 0$ (regression model is not statistically significant)
$H_1: r^2 \neq 0$ (regression model is statistically significant)

$$r^2 = \frac{SSR}{SST} = \frac{166.6}{377.68} \times 100 = 44.11\%$$

$$F \text{ value} = \frac{\frac{SSR}{k-1}}{\frac{SSE}{n-k}} = \frac{\frac{166.6}{1}}{\frac{209.77}{12-2}} = 7.94$$

$F_{k-1, n-k} = F_{1,10} = 4.97$



At 5% significance level, we reject the null hypothesis and conclude that the regression model is statistically significant.

Example 13.

The job ~~performance~~ ~~satisfaction~~ of an employee is said to be influenced by
(i) remuneration / benefits (X_1)
(ii) experience (X_2)
(iii) Gender (X_3)
(iv) academic qualifications (X_4)

The following data has been obtained from the HR records relating to 20 employees.
The performance is believed to assume a linear regression model.

	y	X_1	X_2	X_3	X_4
Y: 1-10	8.4	7	5	M	4
X_1: 3-10	6.2	6	3	M	4
	3.9	5	1	F	3
X_2: 1-6	4.5	3	2	M	2
X_3: F, M	~~6.2~~				
	5.2	4	4	F	1
X_4: 1-5	3.5	7	6	F	2
	8.0	9	2	M	5
	4.7	6	4	F	4
	3.3	4	2	F	5
	7.7	3	6	M	4
	8.1	7	6	F	3
	7.3	6	5	F	3

y	x_1	x_2	x_3	x_4
6.0	8	4	M O	2
7.5	7	5	F 1	3
88	6	6	M O	4
3.4	5	1	F 1	5
9.2	5	6	M O	3
7.0	6	5	M O	4
8.6	4	6	M O	3
9.9	8	6	F 1	5

$\hat{y} = 1 + 0.25x_1 + 0.76x_2 - 1.384x_3 + 0.459x_4$

Required.

(i) Interpret the coefficients for each of the variables including the constant.

Intercept - 1. The minimum job performance level is 1.

$x_1 = 0.248$. If the remuneration scale increases by 1 the job performance of an employee will increase by 0.248

$x_2 = 0.76$ - If the experience scale increases by 1 the employee performance would increase by 0.76

x_3 - Holding other factors constant the performance of a male employee is higher than the performance of a female employee by 1.384

x_4 - If the academic qualification scale of the employee is increased by 1 the job performance would increase by 0.459.

Test the significance of each of the predictor variables at 5% significance level

Construct a 90% CI for x_1 and x_2 and comment on your results

Test the power of model

05/03/2020

Dummy variable

A dummy variable is used in regression analysis to indicate absence or presence of some categorical outcome.

A dummy variable will take value 0 or 1. A dummy independent variable ch for some observations has a value of zero will make the variables coefficient to have no role or influence in the dependent variable whereas when the dummy variable takes a value of 1 its coefficients will have an influence on the dependent variable

Examples of dummy variables

Gender can either be male or female

Political affiliation e.g democratics or republican

Academic qualification graduate or nongraduate

Example

The following data relates to the final score of the students as influenced by gender and class attendance

Final score out of 100	Gender		Class attendance 0-13
48	F	1	10
64	M	0	11
72	F	1	11
55	F	1	8
69	M	0	9
57	F	1	7
46	M	0	12
66	M	0	8
50	F	1	9
61	F	1	10

Required

i) Regression model for the above data
ii) Interprete the meaning of the coefficients for each of the independent variables
iii) Test whether each of the predictor variable is statistically significant at 1% significance level.
iv) Comment on the significance of the regression model at 5% significance level.

Y	X_1	X_2	X_1X_2	X_1^2	X_2^2	X_1Y	X_2Y
48	1	10	10	1	100	48	480
64	0	11	0	0	121	0	704
72	1	11	11	1	121	72	792
55	1	8	8	1	64	55	440
69	0	9	0	0	81	0	621
57	1	7	7	1	49	57	399
46	0	12	0	0	144	0	552
66	0	8	0	0	64	0	528
50	1	9	9	1	81	50	450
61	1	10	10	1	100	61	610
588	6	95	55	6	925	343	5576

$H_0: r^2 = 0$

$H_0: r^2 \neq 0$

$$F = \frac{\sum(\hat{y}-\bar{y})^2/k-1}{\sum(y-\hat{y})^2/n-k}$$

$F_{2,7}$

$$\Sigma y = nB_0 + B_1\Sigma x_1 + B_2\Sigma x_2$$
$$\Sigma x_1 y = B_0\Sigma x_1 + B_1\Sigma x_1^2 + B_2\Sigma x_1x_2$$
$$\Sigma x_2 y = B_0\Sigma x_2 + B_1\Sigma x_1x_2 + B_2\Sigma x_2^2$$

$$\begin{bmatrix} \Sigma y \\ \Sigma x_1 y \\ \Sigma x_2 y \end{bmatrix} = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1x_2 \\ \Sigma x_2 & \Sigma x_1x_2 & \Sigma x_2^2 \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$B_0 = \frac{\begin{vmatrix} \Sigma y & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 y & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 y & \Sigma x_1 x_2 & \Sigma x_2^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2^2 \end{vmatrix}}$$

$$+ \quad - \quad +$$
$$- \quad + \quad -$$
$$+ \quad - \quad +$$

$$= \frac{\begin{vmatrix} 588 & 6 & 95 \\ 343 & 6 & 55 \\ 5576 & 55 & 925 \end{vmatrix}}{\begin{vmatrix} 10 & 6 & 95 \\ 6 & 6 & 55 \\ 95 & 55 & 925 \end{vmatrix}}$$

$$= \frac{588(5550 - 3025) - 6(317275 - 30680) + 95(18865 - 33456)}{10(5550 - 3025) - 6(5550 - 5225) + 95(330 - 570)}$$

$$= \frac{34985}{500}$$

$$= 69.97$$

$$B_1 = \frac{\begin{vmatrix} 10 & 588 & 95 \\ 6 & 343 & 55 \\ 95 & 5576 & 925 \end{vmatrix}}{\begin{vmatrix} 10 & 6 & 95 \\ 6 & 6 & 55 \\ 95 & 55 & 925 \end{vmatrix}}$$

$$= \frac{10(317275 - 30680) - 588(5550 - 5225) + 95(33456 - 32585)}{500}$$

$$= \frac{-2405}{500}$$

$$= -4.81$$

$$B_2 = \frac{\begin{vmatrix} 10 & 6 & 588 \\ 6 & 6 & 343 \\ 95 & 55 & 5576 \end{vmatrix}}{\begin{vmatrix} 10 & 6 & 95 \\ 6 & 6 & 55 \\ 95 & 55 & 925 \end{vmatrix}}$$

$$B_2 = \frac{10(33456 - 18865) - 6(33456 - 32585) + 588(330 - 570)}{500}$$

$$= \frac{-436}{}$$

$$= -0.872.$$

$$y = 69.97 - 4.81X_1 - 0.872X_2.$$

$F=1$, $M=0$

(iii) Score male student 5 classes

$$y = 69.97 - 4.81(0) + 0.872(5)$$
$$= 65.61$$

$F=0$ $M=1$

$$y = 65.16 + 4.81(1) - 0.872(5)$$
$$= 65.61.$$

ii) $B_0 = 69.97$

$B_1 = -4.81$ → ~~Holding other factors constant~~, the male student final score will be higher by 4.81 than a female student who attended the same number of classes

$B_2 - 0.872$ - Holding other factors constant when the class attended is increased by one the final score would decrease by 0.872.

iii) $H_0 : B_1 = 0$

$H_0 : B_1 \neq 0$

$$t = b / S_b$$

$$S_b = \sqrt{\frac{\Sigma(y-\bar{y})^2/n-k}{\Sigma(x_i-\bar{x})^2}}$$

$$S_{b_1} = \sqrt{\frac{681.77}{7}} = \frac{6.37}{2.4}$$

$$S_{b_2} = \sqrt{\frac{681.77}{7}} = 2.08$$