Name = Darshan Lodha
Roll No. = 18

Probability And Statistics - 2

Assignment 2

Q. 1]

Answer: Given that: 'y' is the number of hours until Anand's next visit to an ATM and 'x' is the amount withdrawn.

(a) $\sum_{i=1}^{10} x_i = 850$ and $\sum_{i=1}^{10} x_i^2 = 92,500$

$\sum_{i=1}^{10} y_i = 1737$ and $\sum_{i=1}^{10} y_i^2 = 372717$

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

$$= 92500 - 10\left(\frac{850}{10}\right)^2$$

$$S_{xx} = 20250$$

$$S_{yy} = \Sigma y^2 - n(\bar{y})^2$$

$$= 372717 - 10\left(\frac{1737}{10}\right)^2$$

$$S_{yy} = 71000.1$$

$$S_{xy} = \Sigma xy - n(\bar{x})(\bar{y})$$

$$= 182560 - 10\left(\frac{850}{10}\right)\left(\frac{1737}{10}\right) = 34915$$

The equation of the regression line would be,

$$y_i = \alpha + \beta x_i$$

$$\therefore \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

and $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$

$$\therefore \hat{\beta} = \frac{34915}{20250}$$

$$\therefore \hat{\beta} = 1.724198$$

$$\therefore \hat{\alpha} = \left(\frac{1737}{10}\right) - 1.724198\left(\frac{850}{10}\right)$$

$$\therefore \hat{\alpha} = 27.14317$$

$$\therefore \text{Equation: } \boxed{y_i = 27.14317 + 1.724198 x_i}$$

(b) In the context of the question, the gradient of the regression line represents the amount of hours per rupee spent.

Q.2]
Answer:

x = Share price of Dell in US \$
y = Share price of IBM in US \$

(i) $\Sigma x = 385.2$

$\Sigma x^2 = 12666.58$

$\Sigma y = 1162.5$

$\Sigma y^2 = 119026.9$

$\Sigma xy = 38191.41$

and $n = 12$

$$S_{xx} = \Sigma x^2 - n(\bar{x})^2$$

$$= 12666.58 - (12)\left(\frac{385.2}{12}\right)^2$$

$$S_{xx} = 301.66$$

$$S_{yy} = \Sigma y^2 - n(\bar{y})^2$$

$$= 119026.9 - 12\left(\frac{1162.5}{12}\right)^2$$

$$S_{yy} = 6409.7125$$

$$S_{xy} = \Sigma xy - n(\bar{x})(\bar{y})$$

$$= 38191.41 - 12\left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right)$$

$$S_{xy} = 875.16$$

The least squares regression line of y over x is

$$y_i = \alpha + \beta x_i$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{875 \cdot 16}{301 \cdot 66}$$

$$\hat{\beta} = 2 \cdot 901147$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$= \frac{1162 \cdot 5}{12} - 2 \cdot 901147 \left(\frac{385 \cdot 2}{12}\right)$$

$$\hat{\alpha} = 3 \cdot 74818$$

$\therefore$ The least squares fit regression line is :

$$\boxed{y_i = 3 \cdot 74818 + 2 \cdot 901147 x_i}$$

(ii) Assuming that the random errors follow a normal distribution with mean 0 and constant variance σ^2.

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \left[S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right]$$

$$= \frac{1}{10} \left(6409 \cdot 7125 - \frac{(875 \cdot 16)^2}{301 \cdot 66} \right)$$

$\therefore \hat{\sigma}^2 = 387.0746$

Using the test statistic

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{(n-2)}$$

$t_{10, 2.5\%} = -2.228$
$t_{10, 97.5\%} = 2.228$

$$P(-2.228 < t_{10} < 2.228) = 0.95$$

$$P\left(-2.228 < \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} < 2.228\right) = 0.95$$

$$\therefore P\left(\hat{\beta} - 2.228\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta < \hat{\beta} + 2.228\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right) = 0.95$$

95% confidence interval for β is given by

$$\boxed{(0.3773547, 5.42494)}$$

(iii) Given that: $x_0 = 40$

$$\therefore \hat{\mu}_0 = 3.74818 + 2.901147(40)$$

$$\therefore \hat{\mu}_0 = 119.79406$$

$$Var[\hat{\mu}_0] = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2$$

$$= \left[\frac{1}{12} + \frac{(40-32.1)^2}{301.66}\right] 387.0745$$

$$= 112.3375$$

Using the statistic,

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{Var[\hat{\mu}_0]}} \sim t_{(n-2)}$$

$t_{10,\ 2.5\%} = -2.228$

$t_{10,\ 97.5\%} = 2.228$

$$P\left(-2.228 < t_{10} < 2.228\right) = 0.95$$

$$P\left(-2.228 < \frac{\hat{\mu}_0 - \mu_0}{\sqrt{Var[\hat{\mu}_0]}} < 2.228\right) = 0.95$$

$$P\left(\hat{\mu}_0 - 2.228\sqrt{Var[\hat{\mu}_0]} < \mu_0 < \hat{\mu}_0 + 2.228\sqrt{Var[\hat{\mu}_0]}\right) = 0.95$$

$\therefore$ 95% confidence interval for μ_0 is given by

$$\boxed{(96.17963,\ 143.4085)}$$

Q.3]

Answer (i) From the plots of the 4 cities itself, we can infer that the mean time differences differ for all the cities.
The variation seems to the highest for city D.
But the results cannot be inferred solely on the basis of the plot, as it consists of a random sample of 7 commuters.

(ii) Assumptions required to carry out the ANOVA are:
→ The distribution of the population from which the random sample is determined must be normal.
→ The population should have a common variance
→ The observations should be independent.

(iii)

H_0: The mean time differences is the same for all four cities.

v/s

H_1: The mean time differences is not the same for all four cities

$$SS_{Between} = \left[\frac{64^2}{7} + \frac{44^2}{7} + \frac{8^2}{7} + \frac{(-13)^2}{7}\right] - \left[\frac{103^2}{28}\right]$$

$$SS_{Between} = 516.107$$

$$SS_{TOTAL} = 1495 - (28)\left(\frac{103}{28}\right)^2$$

$$SS_{TOTAL} = 1116.107$$

$SS_{Residual} = SS_{Total} - SS_{Between}$

$= 1116.107 - 516.107$

$= 600$

ANOVA table

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares
Between	3	516.107	172.036
Residual	24	600	25
Total	27	1116.107	

$$F = \frac{\text{Mean sum of squares (between)}}{\text{Mean sum of squares (residual)}}$$

$$= \frac{172.036}{25}$$

$F = 6.88144$

From tables,

$F_{3,24}(5\%) = 3.009$

$\because 6.88144 > 3.009$

We have sufficient evidence to reject H_0 at 5% level of significance. Thus, we conclude that the mean time differences is not the same for all four cities

(iv)

Σy for city A = 64

$\therefore$ $\bar{y}$ for city A = 9·14

Similarly,

$\bar{y}$ for city B = 6·29
$\bar{y}$ for city C = 1·14
$\bar{y}$ for city D = −1·86

Let $\bar{y}_1 = 9 \cdot 14$, $\bar{y}_2 = 6 \cdot 29$, $\bar{y}_3 = 1 \cdot 14$ and $\bar{y}_4 = -1 \cdot 86$
It implies that,

$$\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$$

$$\hat{\sigma}^2 = \frac{SS_{Residual}}{(n-k)} = 25 \quad \text{(From ANOVA table)}$$

The LSD between any pair of means is

$$t_{24, 2.57} \times \hat{\sigma} \sqrt{\frac{1}{7} + \frac{1}{7}}$$

$$= 2 \cdot 064 \times \sqrt{25} \cdot \sqrt{\frac{1}{7} + \frac{1}{7}}$$

$$= 5 \cdot 52$$

Now,

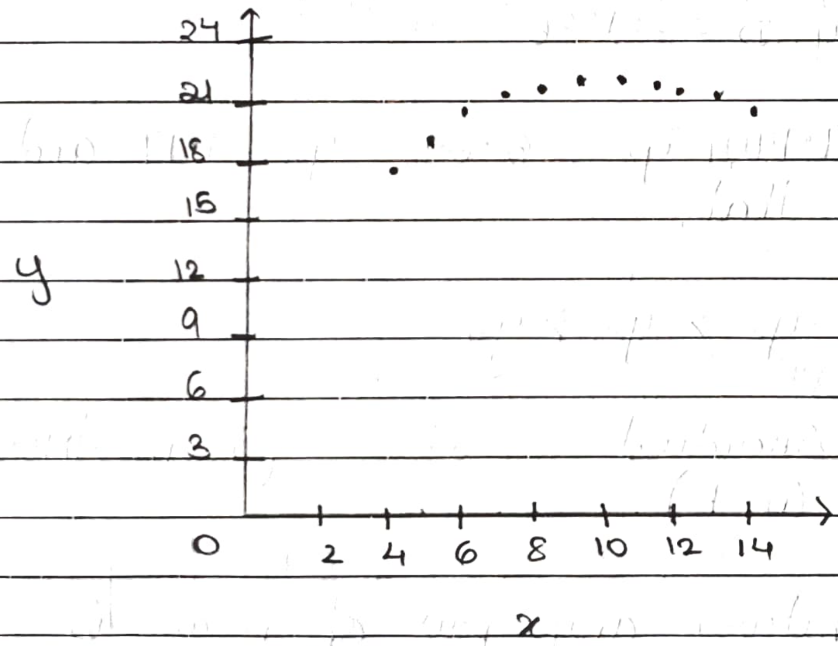
$\bar{y}_1 - \bar{y}_2 = 2.85$

$\bar{y}_2 - \bar{y}_3 = 5.15$

$\bar{y}_3 - \bar{y}_4 = 3$

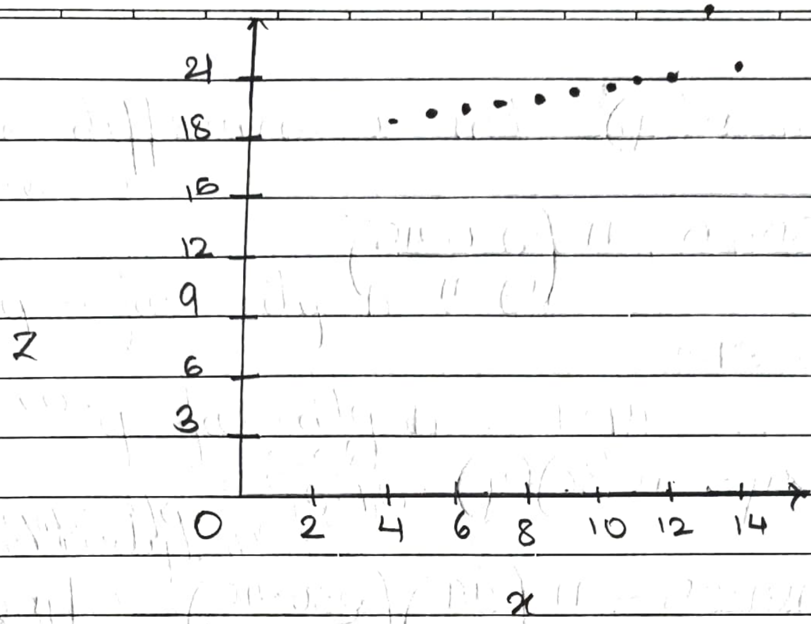
All these differences are less than the LSD 5.52.

Q.4]
Answer (a)



From the plot above, we can infer that a line of regression is not appropriate to model the relationship between minimum temperature and maximum temperature on weekday.

(b)



A line of regression is appropriate to model the linear relationship between minimum temperature and maximum temperature on weekend.
Only ~~one~~ one point seems to be an outlier.

(c) Given that: $\sum x = 99$

$\sum x^2 = 1001$

$\sum y = 230.45$

$\sum y^2 = 4850.30$

$\sum xy = 2104.85$

The least squares regression equation of y on x would be given by,

$$S_{xx} = \sum x^2 - n(\bar{x})^2$$

$$= 1001 - 11\left(\frac{99}{11}\right)^2$$

$$S_{xx} = 110$$

$$S_{yy} = \Sigma y^2 - n(\bar{y})^2$$

$$= 4850.30 - 11\left(\frac{230.45}{11}\right)^2$$

$$S_{yy} = 22.3725$$

$$S_{xy} = \Sigma xy - n(\bar{x})(\bar{y})$$

$$= 2104.85 - 11\left(\frac{99}{11}\right)\left(\frac{230.45}{11}\right)$$

$$S_{xy} = 30.8$$

Now,

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta} = \frac{30.8}{110} = 0.28$$

And

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\therefore \hat{\alpha} = \frac{230.45}{11} - 0.28\left(\frac{99}{11}\right)$$

$$\therefore \hat{\alpha} = 18.43$$

$\therefore$ The least squares fitted regression line of y on x is

$$\boxed{y_i = 18.43 + 0.28x_i}$$

(d) $\sum z = 230.45$

$\sum z^2 = 4850.30$

$\sum xz = 2104.85$

$$S_{zz} = \sum z^2 - n(\bar{z})^2$$

$$= 4850.30 - 11\left(\frac{230.45}{11}\right)^2$$

$$S_{zz} = 22.3725$$

$$S_{xx} = 110$$

$$S_{xz} = \sum xz - n(\bar{x})(\bar{z})$$

$$= 2104.85 - 11\left(\frac{99}{11}\right)\left(\frac{230.45}{11}\right)$$

$$S_{xz} = 30.8$$

$$\hat{\beta} = \frac{S_{xz}}{S_{xx}} = \frac{30.8}{110} = 0.28$$

$$\hat{\alpha} = \bar{z} - \hat{\beta}\bar{x}$$

$$= \left(\frac{230.45}{11}\right) - (0.28)(9)$$

$$\hat{\alpha} = 18.43$$

Thus, the least squares fitted regression line of z on x is as same as obtained in (c) $= 18.43 + 0.28x_i$

(e) (i) Correlation coefficient of x and y $= \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}}$

$$= \frac{30.8}{\sqrt{110} \cdot \sqrt{22.3725}}$$

$$= \underline{0.620865}$$

(ii) Correlation coefficient of x and z $= \frac{S_{xz}}{\sqrt{S_{xx}} \cdot \sqrt{S_{zz}}}$

$$= \frac{30.8}{\sqrt{110} \cdot \sqrt{22.3725}}$$

$$= \underline{0.620865}$$

Though the relationship between x and y; and x and z ~~a~~ is different, their correlation coefficients are the same indicating a same relationship

(f)

After omitting the values for city 3

$\Sigma x = 86$ $\Sigma y = 209.1$ $\Sigma z = 205.14$

$\Sigma x^2 = 832$ $\Sigma y^2 = 4394.48$ $\Sigma z^2 = 4209.704$

(i)

$$S_{xx} = \Sigma x^2 - n(\bar{x})^2$$
$$= 832 - 10\left(\frac{86}{10}\right)^2$$

$$S_{xx} = 92.4$$

$$S_{zz} = \Sigma z^2 - n(\bar{z})^2$$

$$= 4209.704 - 10\left(\frac{205.14}{10}\right)^2$$

$$S_{zz} = 1.46204$$

$$\Sigma xz = 1775.82$$

$$S_{xz} = \Sigma xz - n(\bar{x})(\bar{z})$$

$$= 1775.82 - 10\left(\frac{86}{10}\right)\left(\frac{205.14}{10}\right)$$

$$S_{xz} = 11.616$$

$$\hat{\beta} = \frac{S_{xz}}{S_{xx}} = \frac{11.616}{92.4} = 0.125714$$

$$\hat{\alpha} = y\,\bar{z} - \hat{\beta}\bar{x}$$

$$= \left(\frac{205.14}{10}\right) - 0.125714\left(\frac{86}{10}\right)$$

$$\therefore \hat{\alpha} = 19.4329$$

$\therefore$ The regression line of z on x is $\underline{19.4329 + 0.125714x}$

(ii) Correlation coefficient between x and z $= \dfrac{S_{xz}}{\sqrt{S_{xx}} \cdot \sqrt{S_{zz}}}$

$= \dfrac{11.616}{\sqrt{92.4} \times \sqrt{1.46204}}$

$= 0.999404$

(g) After removing the value of city 3 which was an outlier the relationship (linear) between variables x and z. The correlation coefficient in f(ii) implies that there is almost a perfectly positive correlation between x and z.

Q.5]

Answer (a) The mathematical model of one way ANOVA is given as:

$$Y_{ij} = \mu + T_i + e_{ij} \quad ; \quad i = 1, 2, \ldots, k \text{ and } j = 1, 2, \ldots, n_i$$

where, k = number of treatments
n_i = number of responses from the i^{th} treatment
Y_{ij} is the j^{th} response from the i^{th} treatment
T_i is the i^{th} treatment
μ is the overall mean response
e_{ij} is the error term.

The assumption is: e_{ij} is i.i.d. $N(0, \sigma^2)$

(b) H_0: The mean claim amounts of five companies are equal.

v/s

H_1: The mean claim amounts of five companies are not equal.

$$SS_{\underset{Between}{\cancel{Total}}} = \sum y^2 - n(\bar{y})^2$$

$$= \left[\frac{354^2}{7} + \frac{386^2}{8} + \frac{89^2}{6} + \frac{645^2}{7} + \frac{384^2}{5}\right] - \frac{1856^2}{33}$$

$$SS_{\underset{Between}{\cancel{Total}}} = 22325.6892$$

$$SS_{Total} = \sum y^2 - n\bar{y}^2$$

$$= 174316 - \frac{1856^2}{33}$$

$$SS_{Total} = 69930.061$$

$$SS_{Residual} = SS_{Total} - SS_{Between}$$

$$= 69930.061 - 22325.6892$$

$$SS_{Residual} = 47604.372$$

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares
Between	4	22325.6892	5581.4223
Residual	28	47604.372	1700.1561
Total	33	69930.061	

$$F = \frac{\text{Mean Sum of squares (Between)}}{\text{Mean Sum of squares (Residual)}}$$

$$= \frac{5581.4223}{1700.1561}$$

$$F = 3.282889$$

$$F_{4,28,(5\%)} = 2.714$$

$3.282889 > 2.714$

We have sufficient evidence to reject H_0 at 5% level of significance.

Hence, we conclude that the mean claim amounts of five companies are not equal.

(c) H_0: The salaries of actuarial students are independent of the actuarial papers they have cleared

v/s

H_1: The salaries of actuarial students are not independent of the actuarial papers they have cleared.

Chi-square contingency table test

Observed values (f_i)

Papers cleared	Salary per annum (in Rs. lacs)				Total
	3-5	5-8	8-10	10-12	
0-3	45	20	6	5	76
4-6	7	20	9	6	42
7-9	5	8	15	12	40
Total	57	48	30	23	158

Expected values (e_i)

Papers cleared	Salary per annum (in Rs. lacs)				Total
	3-5	5-8	8-10	10-12	
0-3	27.418	23.089	14.43	11.063	76
4-6	15.152	12.7595	7.975	6.114	42
7-9	14.4304	12.152	7.595	5.823	40
Total	57	48	30	23	158

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

$$= \frac{(45 - 27.418)^2}{27.418} + \frac{(7 - 15.152)^2}{15.152} + \cdots + \frac{(12 - 5.823)^2}{5.823}$$

$$= 49.919$$

$$\chi^2_{6, 5\%} = 12.59$$

$\therefore$ $49.919 > 12.59$

We have sufficiently strong evidence to reject H_0 at 5% level of significance.

Thus, we conclude that the salaries of actuarial students depend on the number of paper they have cleared.

Q.6]

Answer (i) The assumptions required for one way analysis of variance are:

→ The populations must be normal

→ The populations must have a common variance

→ The observations are independent.

The sample variance observed for the four rates appear to be very different from each other. Thus, we can clearly see that the assumption that the underlying populations must have a common variance will not hold for the data as they are.

(ii)(a) Considering $x \rightarrow \sqrt{x}$, the value of the sample mean of rate 1 will be

$$\frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3} = \underline{4.52}$$

(b) Considering $x \rightarrow \log_e x$, the value of sample variance for rate 2 will be

$$\frac{1}{2}\left[(\ln 180 - 4.827)^2 + (\ln 90 - 4.827)^2 + (\ln 120 - 4.827)^2\right]$$

$= \underline{0.1213}$

(iii) The scientist is correct in asserting that the log transformation must be done before carrying out a one-way analysis of variance as for this transform-ation, the assumption of common variance will hold true.

(iv) (a) The ANOVA will be performed on the transformed data.
Assuming the following model:

$$Y_{ij} = \mu + T_i + \varepsilon_{ij} \quad , \quad i = 1,2,3,4 \text{ and } j = 1,2,3$$

(b) H_0: The mean number of germinations is the same for all four rates.
v/s
H_1: The mean number of germinations is not the same for all four rates.

Modified data:

Rate	$\log_e(x)$		
	Mean	Variance	Y_L^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.716	0.186	294.053
4	6.575	0.148	389.060
Total	20.110	0.618	973.373

Calculating sum of squares:

$$SS_{Rate} = \frac{973.373}{3} - \frac{(3 \times 20.110)^2}{12}$$

$$= 21.153$$

$$SS_{Residuals} = \sum_{i=1}^{4} (2 \times \text{variance})$$

$$= 2 \times 0.618$$

$$= 1.236$$

ANOVA table:

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
Total	11	22.389	

$F = \dfrac{\text{Mean sum of squares (Rate)}}{\text{Mean sum of squares (Residual)}}$

$= \dfrac{7.051}{0.155}$

$= 45.4903$

$F_{3,8}(1\%) = 7.951$

$\because$ $45.4903 > 7.951$

(c) $\therefore$ We have sufficiently strong evidence to reject H_0 even at the 1% level of significance.

Thus, we conclude that the mean number of germination is not the same for all four rates.

Q.7]

Answer: Given: $\Sigma x = 39$, $\Sigma x^2 = 207$

$\Sigma y = 562$, $\Sigma y^2 = 40508$ $\Sigma xy = 2853$

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

$= 207 - 10\left(\dfrac{39}{10}\right)^2$

$S_{xx} = 54.9$

$S_{yy} = \Sigma y^2 - n(\bar{y})^2$

$$= 40508 - (10)\left(\frac{562}{10}\right)^2$$

$$S_{yy} = 8923 \cdot 6$$

$$S_{xy} = \Sigma xy - n(\bar{x})(\bar{y})$$

$$= 2853 - 10\left(\frac{39}{10}\right)\left(\frac{562}{10}\right)$$

$$S_{xy} = 661 \cdot 2$$

Correlation coefficient (r) $= \dfrac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}}$

$$= \frac{661 \cdot 2}{\sqrt{54.9} \cdot \sqrt{8923 \cdot 6}}$$

$$= \underline{0.9446624}$$

(ii) $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{661 \cdot 2}{54 \cdot 9} = 12 \cdot 04372$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\therefore \hat{\alpha} = \frac{562}{10} - (\cancel{54 \cdot 9})\left(\frac{39}{10}\right)(12 \cdot 04372)$$

$$\hat{\alpha} = 9 \cdot 2295$$

$\therefore$ The fitted regression line is $y_i = \hat{\alpha} + \hat{\beta} x_i$

$$\therefore \underline{y_i = 9 \cdot 2295 + 12 \cdot 04372\, x_i}$$

(iii) $\underline{SS_{TOT} = SS_{RES} + SS_{REG}}$

$\underline{SS_{TOT} = S_{yy} = 8923.6}$

$SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

$= 8923.6 - \frac{(661.2)^2}{54.9}$

$\underline{SS_{REG} = 960.2951}$

$SS_{REG} = SS_{TOT} - SS_{RES}$

$= 8923.6 - 960.2951$

$\underline{SS_{REG} = 7963.3049}$

(iv) Coefficient of determination (R^2) $= \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}}$

$= \frac{SS_{REG}}{SS_{TOT}}$

$= \frac{7963.3049}{8923.60}$

$\underline{R^2 = 0.892387}$

We know,

$$r^2 = R^2$$

$$\therefore r = \sqrt{R^2}$$

$$\therefore r = \sqrt{0.892387}$$

$$\therefore r = 0.9446624$$

which is the same as r calculated in (i)