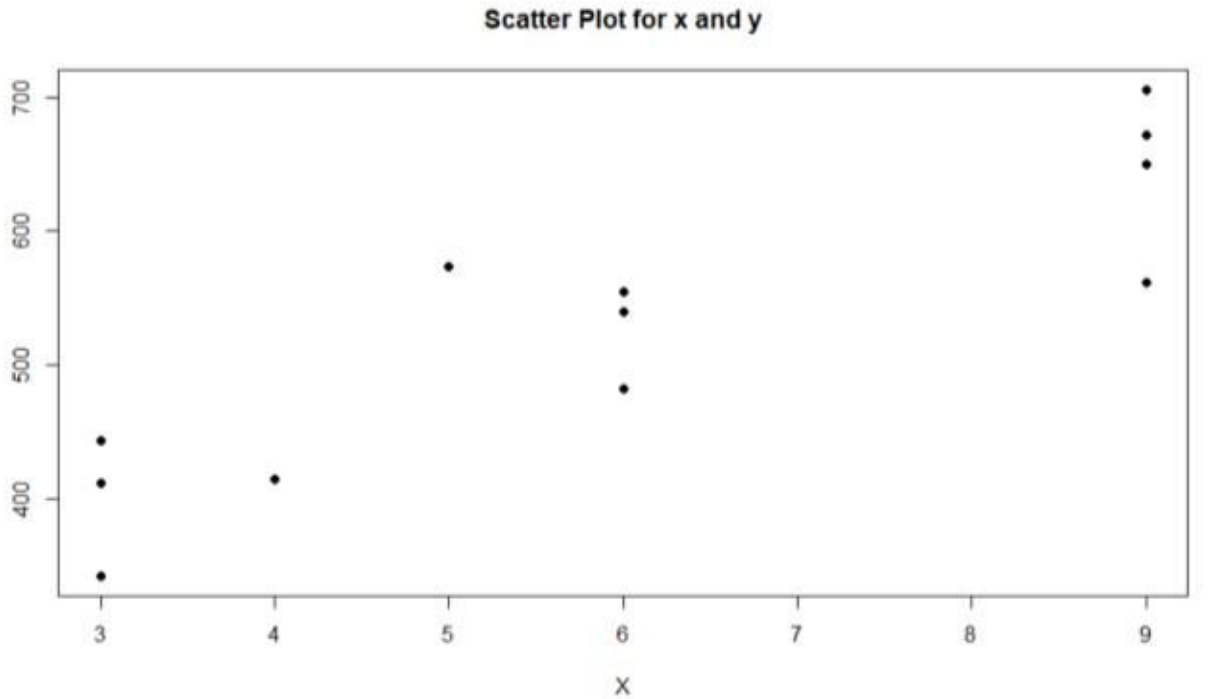
Ans 1)

a)



Theory Assignment 3

Ans 1-

b) X and Y have a positive linear relationship.

c)

X	Y	x^2	y^2	XY	$x-\bar{x}$	$(x-\bar{x})^2$
6	540	36	291600	3240	0	0
4	415	16	172225	1660	-2	4
6	555	36	308025	3330	0	0
9	650	81	422500	5850	3	9
3	412	9	169744	1236	-3	9
9	562	81	315844	5058	3	9
6	482	36	232324	2892	0	0
3	443	9	196249	1329	-3	9
9	706	81	498436	6354	3	9
5	574	25	329476	2870	-1	1
3	342	9	116964	1026	-3	9
9	672	81	451584	6048	3	9
ΣX = 72	ΣY = 6353	Σx^2 = 500	Σy^2 = 3504971	ΣXY = 40893	$\Sigma(x-\bar{x})$ = 0	$\Sigma(x-\bar{x})^2$ = 68

$$\text{Correlation, } r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(12 \times 40893) - (72 \times 6353)}{\sqrt{[(12 \times 500) - (72 \times 72)][(12 \times 3504971) - (6353)^2}}$$

$$= \frac{490716 - 457416}{\sqrt{816 \times 1699043}} = 0.89432$$

(suggesting strong positive linear relation)

d) Simple Linear Regression is given by

$$y = a_0 + a_1 x$$

Using table from the previous part for below calculations

Here, $a_0 = \dfrac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$

$$= \frac{(6353 \times 500) - (72 \times 40893)}{(12 \times 500) - (72 \times 72)}$$

$$= \frac{232204}{816} = 284.5637$$

$$a_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

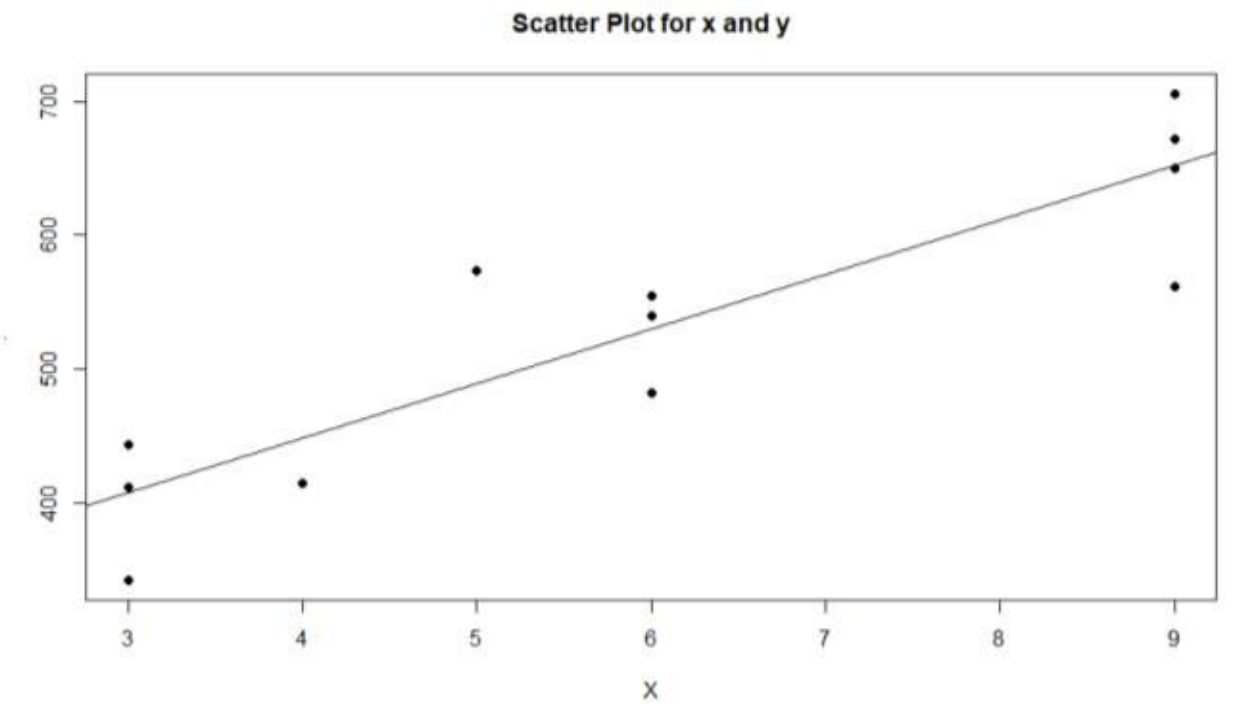
$$= \frac{(12 \times 40893) - (72 \times 6353)}{(12 \times 500) - (72 \times 72)}$$

$$= \frac{33300}{816} = 40.8088$$

Hence, the equation:

$$y = 284.5637 + 40.8088x$$

e)



f) Fitted values and residuals for each observation

x_i	y_i	$\hat{y}_i = a_0 + a_1 x_i$	$\varepsilon = y_i - \hat{y}$	$(y_i - \hat{y})^2$
6	540	529.4165	10.5835	112.01047
4	415	447.7990	-32.7990	1075.7744
6	555	529.4167	25.5833	654.5052
9	650	651.8431	-1.8431	3.3970
3	412	406.9902	5.0098	25.0980
9	562	651.8431	-89.8431	8071.7826
6	482	529.4167	-47.4167	2248.3434
3	443	406.9902	36.0098	1296.7057
9	706	651.8431	54.1568	2932.9589
5	574	488.6078	85.3921	7291.8107
3	342	406.9902	-64.9903	4223.7390
9	672	651.8431	20.1569	406.3006

$\Sigma (y_i - \hat{y}) = 28342.4259$

$$\sum_{i=1}^{12} \varepsilon = 0 \left(10.5835 - 32.7990 + 25.5833 - 1.8431 + 5.0098 - 89.8431 - 47.4167 + 36.0098 + 54.1568 + 85.3921 - 64.9903 + 20.1569\right)$$

$$= 0$$

g) Variance of y as explained by x is given by Coefficient of Determination, r^2.

$r^2 = (\text{Correlation coeff.})^2$

$= 0.7998$

h) $$\sigma_{est} = \sqrt{\frac{\Sigma (y_i - \hat{y})^2}{N-2}}$$

$$= \sqrt{\frac{28342.4259}{10}}$$

{ Using the $(y_i - \hat{y})^2$ values from the previous table

$= 48.5990$ $= 53.2376$

i) Hypothesis Testing to test for significance of 'r'

$H_0: \rho = 0$ (no correlation)

$H_a: \rho \neq 0$ (correlation exists)

Test statistic, $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$ (with $n-2$ degrees of freedom)

$$= \frac{0.89432}{\sqrt{\frac{1-(0.89432)^2}{10}}}$$

$$= \frac{0.89432}{\sqrt{0.0200}} = 6.3238$$

From t-table, at $\alpha = 0.01$

$t_{critical} = t_{0.0}$ $= 3.169$

Since, $t > t_{cri}$, we reject the null hypothesis and say that correlation exists b/w the variables x and y.

j) Prediction interval

$$\hat{y}^* \pm t_{\alpha/2} s_{pred}$$

where $\hat{y}^*$ is the predicted value of y & s_{pred} is the standard deviation of the predicted value of y^*

Here, $\hat{y} = a_0 + a_1 x \Rightarrow 284.5637 + 40.8088x$

$t_{\alpha/2}$ with 10 degrees of freedom is 3.169

$$s_{pred} = s^2 + s_{y^*}^2$$

where s^2 is the standard error of estimate and $s_{y^*}^2$ is variance because of using y^*

Here, s^2 = standard error of estimate

and $s_{\hat{y}^*}^2 = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

Here, $s = 53.2376$

$$s_{y^*}^2 = 53.2376 \sqrt{\frac{1}{12} + \frac{(x-6)^2}{68}}$$

$$s_{pred} = (53.2376)^2 + 53.2376 \sqrt{\frac{1}{12} + \frac{(x-6)^2}{68}}$$

Prediction interval,

$$\left[(284.5637 + 40.8088x) \pm 3.169 \left((53.2376)^2 + 53.2376 \sqrt{\frac{1}{12} + \frac{(x-6)^2}{68}} \right) \right]$$

Here, substituting different values of the independent variable(x) will give the corresponding prediction intervals ~~different values of corresponding y values~~

Ans 2-a) Assumptions of Simple Linear Regression:

i) Linear Relationship → The outcome Y has a roughly linear relationship with the independent variable X.

ii) Homoscedasticity → For each value of X, the distribution of residuals has the same variance.

iii) Independent errors → Residuals should be uncorrelated (independent of each other).

iv) Normally distributed residuals → Residuals should be normally distributed.

Assumptions of Multiple Regression (Multivariable Regression) are same as that of Simple Linear Regression with some additions.

i) No multicollinearity → This occurs when independent variables are too highly correlated to each other. Hence, the assumption states that the independent variables shouldn't be correlated to each other.

Assumptions of Logistic Regression:

i) Logistic Regression requires that the dependent variable should be binary.

ii) It requires that the observations should be independent of each other.

iii) It requires that there should not be any multicollinearity among the independent variables.

iv) It assumes a linear relationship b/w independent

variables and log odds.

v) It requires a large sample size.

b) Violation of assumptions :-

i) Linearity violation for simple linear regression → If we fit a linear model to data which is are nonlinearly related, the predictions are likely to be seriously in error, especially when we extrapolate beyond the range of sample data.

ii) Multicollinearity violation for multiple linear regression → If we have independent variables that are highly correlated to each other, then it is hard to distinguish b/w the effects of the two or more related independent variables to the dependent variable. 1 unit increase in one of the independent variable causes an increase in the other independent variable as well as the dependent variable, thus making it hard to understand which independent variable actually caused the change in value of the dependent variable.

c) Independence of outcome violation for logistic regression → If the outcomes are duplicated,

the error values will be similarly correlated and errors will follow a particular pattern OR trend.

Ans 3-

a) In logistic regression, since the outcome variable is binary and only the predicted mean value (sigmoid) is modeled, then there only two possible values for error (1-m) OR (m-0).

b) Variance in logistic regression is given by $\mu_M P(1-P)$ where P is the probability of getting a binary 1 OR True value. Hence, it varies with the probability of binary outcome 1 OR true., being maximum when P = 0.5 ie. 0.25.

Ans 4-

a) When x & y ~~are~~ ~~no~~ have a non-linear relationship, we can use transformations (eg:- log transforms) to convert it into linear relation. Other transforms can also be applied depending on the distribution of data.

b) To prove that if X1 and X2 are independent, then $Cov(X1, X2) = 0$

~~$Cov(X1, X2) = E[X1 - E(X1)] * E[X2 - E(X2)]$~~

$$Cov(X1, X2) = E[(X1 - E(X1)) * (X2 - E(X2))]$$
$$= E[X1 \cdot X2 - X1 * E(X2) - X2 E(X1) + E(X1) * E(X2)]$$
$$= E(X1 * X2) - E(X1 * E(X2)) - E(X2 \cdot E(X1)) + E(E(X1) * E(X2))$$
$$= E(X1 * X2) - E(X2) * E(X1) - E(X2) * E(X1) + E(X1) * E(X2)$$
$$= E(X1 * X2) - E(X1) * E(X2)$$

Now, if X1 & X2 are statistically independent

Then, $E(X1 * X2) = E(X1) * E(X2)$

Therefore,

$$Cov(X1, X2) = E(X1) \cdot E(X2) - E(X1) \cdot E(X2)$$
$$= 0$$

Hence, proved.

Parts a & b combined

Ans 5- a) & b) We have, $y_i = a_0 + a_1 x_i + \varepsilon$

& $\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i$

Now, $F(a_0, a_1) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

$$F(a_0, a_1) = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x))^2$$

This is the SSE and should be minimum

Hence,

$$\frac{\partial F}{\partial a_0} = \sum_{i=1}^{n} 2(y_i - (a_0 + a_1 x)) = 0 \quad \text{——— (1)}$$

$$\frac{\partial F}{\partial a_1} = \sum_{i=1}^{n} 2(y_i - (a_0 + a_1 x))\, x = 0 \quad \text{——— (2)}$$

From eq^n (1),

$$\Sigma y_i - a_0 n - a_1 \Sigma x_i = 0$$

$$\Rightarrow \quad \Sigma y_i = a_0 n + a_1 \Sigma x_i$$

Dividing by n

$$\frac{\Sigma y_i}{n} = a_0 + a_1 \frac{\Sigma x_i}{n} \quad \text{——— (3)}$$

$$\bar{y} = a_0 + a_1 \bar{x}$$

↳ This shows that $(\bar{x}, \bar{y})$ lie on the regression (Part (b)) line.

From eq^n ②,

~~$\Sigma y_i - (a_0 n + a_1 \Sigma x_i)$~~

$$\Sigma y_i x_i - (a_0 \Sigma x + a_1 \Sigma x^2) = 0$$

$$\Rightarrow \Sigma y_i x_i = a_0 \Sigma x + a_1 \Sigma x^2$$

{ Substituting a_0 from eq^n ③

$$\Sigma y_i x_i = \left(\frac{\Sigma y_i}{n} - \frac{a_1 \Sigma x_i}{n}\right) \Sigma x_i + a_1 \Sigma x_i^2$$

~~$\Sigma y_i = \frac{\Sigma y_i}{n} + a_1 \left(\Sigma x_i - \frac{\Sigma x_i}{n}\right)$~~.

$$\Sigma y_i x_i = \frac{\Sigma y_i \Sigma x_i}{n} - \frac{a_1 (\Sigma x_i)^2}{n} + a_1 \Sigma x_i^2$$

$$a_1 \left(\frac{(\Sigma x_i)^2}{n} - \Sigma x_i^2\right) = \frac{\Sigma y_i \Sigma x_i}{n} - \Sigma y_i x_i$$

$$a_1 = \frac{\frac{\Sigma y_i \Sigma x_i}{n} - \Sigma y_i x_i}{\left(\frac{(\Sigma x_i)^2}{n} - \Sigma x_i^2\right)}$$

Taking n common from above,

$$a_1 = \frac{\Sigma y_i \Sigma x_i - n \Sigma y_i x_i}{(\Sigma x_i)^2 - n \Sigma x_i^2} \quad \text{——— ④}$$

Substituting a_1 to find a_0

$$a_0 = \frac{\Sigma y_i}{n} - \frac{a_1 \Sigma x_i}{n}$$

$$= \frac{\sum y_i}{n} - \frac{\sum x_i}{n}\left(\frac{\sum y_i \sum x_i - n\sum y_i x_i}{(\sum x_i)^2 - n\sum x_i^2}\right)$$

$$= \frac{\sum y_i}{n} - \frac{\sum y_i (\sum x_i)^2 - n\sum x_i \sum y_i x_i}{n\left((\sum x_i)^2 - n\sum x_i^2\right)}$$

$$= \frac{\sum y_i \left((\sum x_i)^2 - n\sum x_i^2\right) - \sum y_i (\sum x_i)^2 + n\sum x_i \sum y_i x_i}{n\left((\sum x_i)^2 - n\sum x_i^2\right)}$$

$$= \frac{-n\sum y_i \sum x_i^2 + n\sum x_i \sum y_i x_i}{n\left((\sum x_i)^2 - n\sum x_i^2\right)}$$

$$a_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Ans 6- a) ① $H_0: \sigma = 15$

$H_a: \sigma < 15$

Calculating using the t value since sample size = 12

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{(9 - 15)}{}$$

② Since the test is left tailed test, critical value for $\alpha = 0.05$ is obtained at the intersection of $(1-\alpha)$ and df. -ie. 0.95 & 11 intersection

$$\chi^2_{critical} = 4.575$$

(3) Obtaining the test value :

$$t = (N-1)\left(\frac{s}{6}\right)^2$$

$$= 11 \times \frac{9^2}{15}$$

$$= \frac{891}{225} = 3.96$$

(4) Since the test value is less than $\chi_{critical}$, the decision is to reject the null hypothesis. ie. there is enough evidence to support the claim that the standard deviation of number of aircrafts stolen every year in United States is less than 15.

b)

a-subpart) (1) $H_0 : \mu = 162.5$

$H_a : \mu \neq 162.5$

(2) Two-tailed test with $\alpha = 0.05$ meaning critical values are at $z_{0.025}$ and $z_{0.975}$ -ie. $z_{critical} = -1.96, 1.96$

(3) Finding the test statistic,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{(165.2 - 162.5)}{6.9/\sqrt{50}} = 2.77$$

(4) Since the test value is within the critical region, we fail to reject the null hypothesis and hence, there is no reason to believe that there is a change in average

④ Since the test value falls outside the critical region, we fail to accept the null hypothesis and believe that the average weight of the students has changed.

b-subpart〉

① $H_0 : \mu = 162.5$

$H_a : \mu \neq 162.5$

② Computing the value of test statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = 2.77$$

Obtaining the area for $z = 2.77$ from the z-table gives 0.0028.

Hence, $p\text{-value} = 2 \times 0.0028$

$p\text{-value} = 0.0056$

~~Given, $\alpha = 0.05$~~

③ Given $\alpha = 0.05$, since $p\text{-value} < \alpha$, we fail to accept the null hypothesis and belive that there is a change in average weight of current batch of students.