CSC591: Foundations of Data Science (HW Theory-3)

Student Name: Rutvik Govind Kolhe

StudentID: 200258232

Unity ID: rkolhe

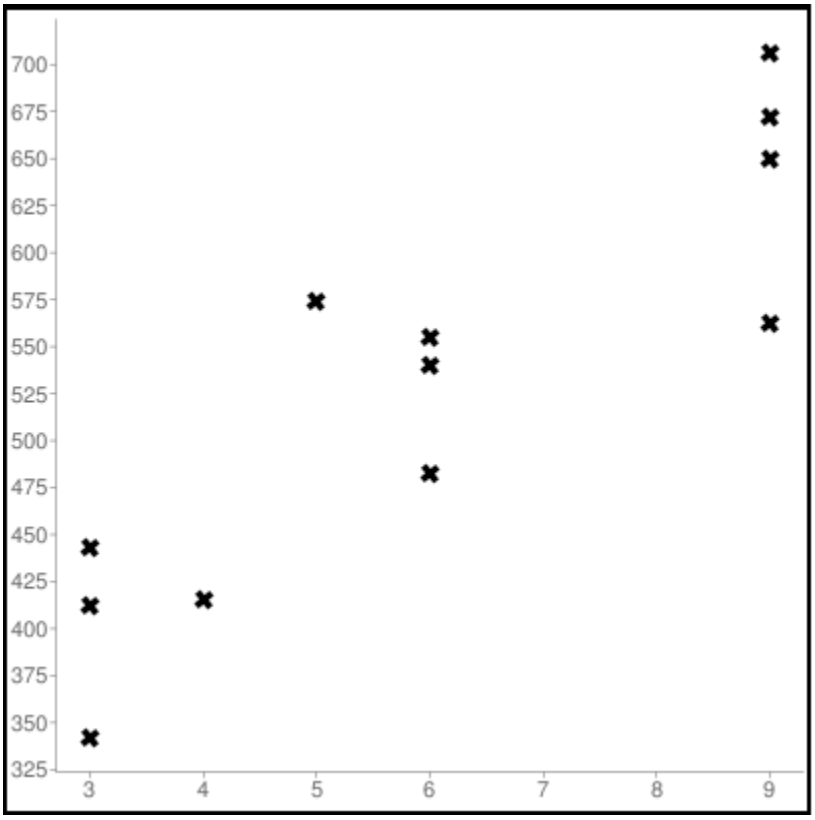
Solutions:

Question 1:

Consider the following table for Solution 1:

	x	y	x*y	x^2	y^2
	6	540	3240	36	291600
	4	415	1660	16	172225
	6	555	3330	36	308025
	9	650	5850	81	422500
	3	412	1236	9	169744
	9	562	5058	81	315844
	6	482	2892	36	232324
	3	443	1329	9	196249
	9	706	6354	81	498436
	5	574	2870	25	329476
	3	342	1026	9	116964
	9	672	6048	81	451584
Total ∑	**72**	**6353**	**40893**	**500**	**3504971**

a)



b) The relationship between x and y is **linear**. This is because as the values of x increases, the values of y increases. We can infer that by looking at the position of the points- their height increases with increase in the displacement.

c)

Q1.

(c)

To compute correlation between x and y:

Here $n = 12$

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

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

$$= \frac{33300}{\sqrt{816 \times 1699043}}$$

$$r = 0.8943$$

Since r (0.8943) is close to 1, the independent variable (x) and dependent variable y possess a strong correlation.

d)

Q1.

(d) Formula for regression parameters b_0 and b_1 where:-

b_0 : intercept

b_1 : slope

$$b_0 = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

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

$$b_0 = 284.5637$$

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

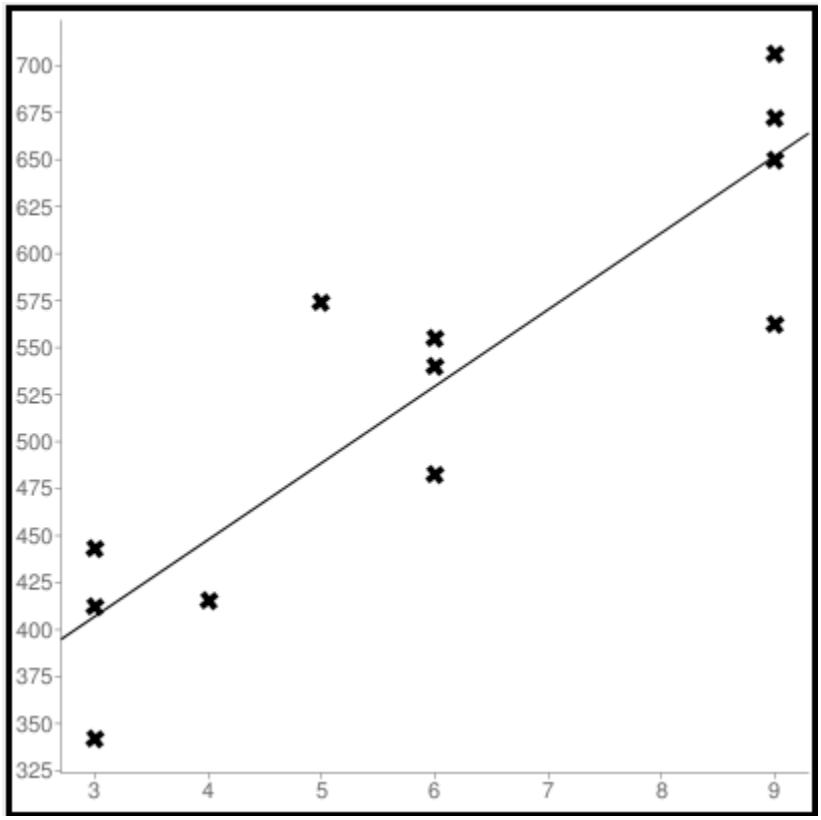
$$= \frac{12(40893) - 72 \times 6353}{12 \times (72)^2 - 12 \times 500 - (72)^2}$$

$$b_1 = 40.8088$$

Ans) The slope is 40.8088 and incept is 284.5637 for the eqn:

$$\hat{y} = 284.5637 + 40.8088x$$

e)



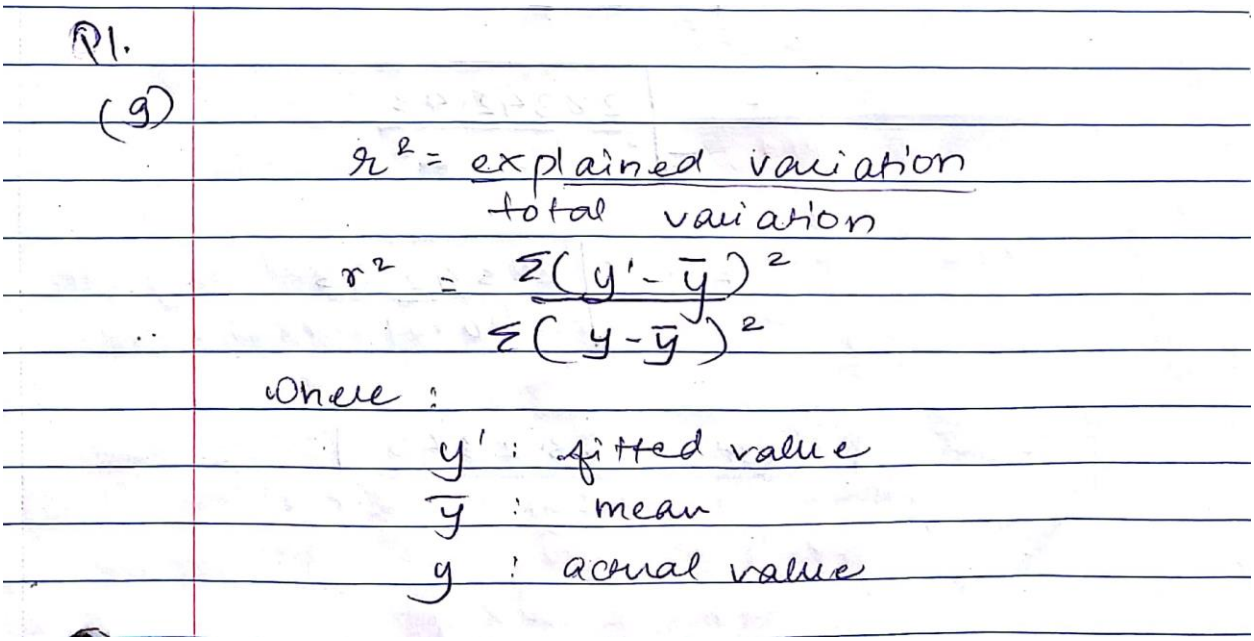
f) To compute the fitted values, we use the simple linear regression equation: y' = 284.5637 + 40.8088*x

We use this equation and substitute every value of x. Then, to calculate the residual, we subtract the individual y' values from the actual values of y. This calculation is shown in the following table:

x	y	y' = 284.5637 + 40.8088 * x	y- y'
6	540	529.4165	10.5835
4	415	447.7989	-32.7989
6	555	529.4165	25.5835
9	650	651.8429	-1.8429
3	412	406.9901	5.0099
9	562	651.8429	-89.8429
6	482	529.4165	-47.4165
3	443	406.9901	36.0099
9	706	651.8429	54.1571
5	574	488.6077	85.3923
3	342	406.9901	-64.9901
9	672	651.8429	20.1571

Taking the sum of the residual column (y-y'), we get the answer to be **0.002**. Hence we can conclude that the sum of residuals is approximately zero.

g) Coefficient of Determination is the measure of variation of dependent variable that is explained by the regression line and the independent variable. It is expressed in r^2. The formula for r^2 is:



	y	y'	(y'- mean(y)) ^ 2	(y- mean(y)) ^2
	540	529.4165	2.5E-07	111.999889
	415	447.7989	6661.514248	13091.24989
	555	529.4165	2.5E-07	654.489889
	650	651.8429	14988.10099	14540.25989
	412	406.9901	14988.34584	13786.75189
	562	651.8429	14988.10099	1061.651889
	482	529.4165	2.5E-07	2248.371889
	443	406.9901	14988.34584	7467.897889
	706	651.8429	14988.10099	31181.55589
	574	488.6077	1665.398966	1987.643889
	342	406.9901	14988.34584	35125.13189
	672	651.8429	14988.10099	20329.91189
Total			113244.3547	141586.9167

$$\therefore \quad r^2 = \frac{113244.3547}{141586.9167}$$

$$r^2 = 0.7998$$

$$r^2 \approx 0.8$$

h) Consider the following table

	y- y'	(y-y') ^ 2
	10.5835	112.0105
	-32.7989	1075.768
	25.5835	654.5155
	-1.8429	3.39628
	5.0099	25.0991
	-89.8429	8071.747
	-47.4165	2248.324
	36.0099	1296.713
	54.1571	2932.991
	85.3923	7291.845
	-64.9901	4223.713
	20.1571	406.3087
Total	0.002	28342.43

h) Standard error of estimate is given by:

$$S_{est} = \sqrt{\frac{\Sigma(y-y')^2}{n-2}}$$

$$= \sqrt{\frac{28342.43}{12-2}}$$

$$= \sqrt{\frac{28342.43}{10}}$$

$$\boxed{S_{est} = 53.2376}$$

i)

Q1.

(i) Step I: State Hypothesis

$H_0: \rho = 0 \qquad H_1: \rho \neq 0$

Step 2: for $\alpha = 0.01$

$d.f = n-2 = 10.$

Critical value $= \pm 3.169$

for 2 tailed

test from t-table

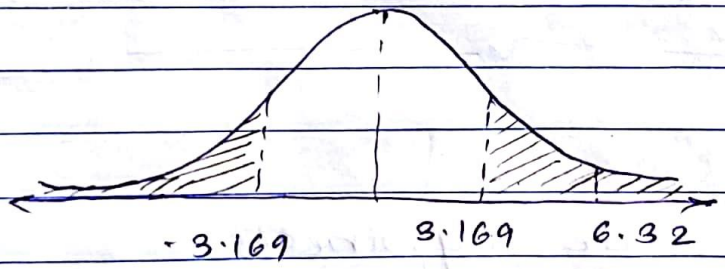
Step 3: Compute test value

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$= 0.8943\sqrt{\frac{10}{1-(.8943)^2}}$$

$$t = 6.32$$

Step 4: Make decision



since 6.32 lies in the critical region, we reject the null hypothesis H_0.

Step 5:- Hence, we can conclude that there is enough evidence to prove a significant relationship between x and y.

j)

Q1.

(j) Formula for prediction interval:

$$y' \pm t_{\frac{\alpha}{2}} \text{Sed.} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}$$

$\Rightarrow$ We substitute the values:

$$\Rightarrow y' \pm (6.32)(53.2376) \times \sqrt{1 + \frac{1}{12} + \frac{(x-6)^2}{500 - \frac{(72)^2}{12}}}$$

$$\Rightarrow y' \pm 336.4616 \sqrt{1.083 + \frac{(x-6)^2}{68}}$$

Hence prediction interval for y:

$$y' - 336.4616 \sqrt{1.083 + \frac{(x-6)^2}{68}} < y < y' + 336.4616 \sqrt{1.083 + \frac{(x-6)^2}{68}}$$

Assume $x = 10$.

$\Rightarrow \quad y' = 284.5637 + 40.8088(10)$

$y' = 692.6517.$

∴ prediction interval =

$$692.6517 \pm 6.32 \times 53.2376 \times \sqrt{1 + \frac{1}{12} + \frac{12(10-6)^2}{12.500 - (72)^2}}$$

$$\Rightarrow 692.6517 \pm 6.336.46 \sqrt{\frac{269}{204}}$$

$$\Rightarrow 692.6517 \pm 386.3620$$

Hence prediction interval for $x=10$ at $\alpha = 0.01$ is $(306.2896, 998.9413)$

ie for $x=10$ y may lie between the interval.

Question 2:

a)

Some assumptions for Simple Linear Regression are [1]:

1. **Linearity assumption**: We assume that there is always a linear relationship between the dependent variable (y) and independent variable (x). This can be checked by observing the scatterplot. The position of the points on the plot indicate whether the regressor and regressand are linear or not.
2. **Normality assumption**: For any given value of independent variable, the value of dependent variable is always normally distributed. It can be checked with a histogram. If the data violates this condition, a variable transformation (ex: log transformation) would solve the problem.
3. **Independence assumption**- We assume that the dependent variable (y) are independent of each other. Any value which is previously predicted does not affect the predication in future

4. **Homoscedasticity:** We assume that the variance of error at any point of an independent variable should be constant. Here the variance of error tells us how far the fitted point lies from it's expected position.
5. The distribution of the error or residuals should be normal.

Assumptions for Multiple Regression

They are like that of simple regression. The assumptions are as follows:

1. **Equal-variance assumption or Homoscedasticity**: Here, we assume that the variances of the dependent variable (y) are same for each value of the independent variable (x1, x2, x3,…xk).
2. **Nonmulticollinearity assumption**: Like that of the independence assumption in linear regression with a difference that, here, the values of independent variables are not correlated to each other.
3. There should not be any influential cases or outliers that are biasing the model. These points may hamper the representation of our model as they are located at the extremes. Hence, we assume that there are no outliers or influential points present. All these points must be identified and removed during the initial stages of data preparation.

Assumptions for Logistic regression [2]

1. **Assumption of appropriate outcome structure**: The logistic regression model confines the output of the dependent variable usually, to either one of the two outcomes or classes. (For ex: male/female, win/loss, yes/no etc.). Hence the first assumption is that it requires the dependent variable to be discrete but mostly dichotomous.
2. This model of regression provides us with the probability of every outcome (P(Y=1)). In most of the cases the outcomes are binary, hence it is assumed that the dependent variable is calculated such that the desired outcome should be one
3. The model should be fitted correctly. It must be ensured that only the necessary variables are considered. There should not be any over-fitting or underfitting of variables to sustain space and prevent loss of data.
4. Even though the logistic regression does not require the independent and dependent variables to be linearly related, it requires the independent variables to be **linearly** related to the **log odds** of any event.
5. It requires **large sample sizes** since the maximum likelihood estimates which are used to find the unknown parameters of the logistic regression equation are less powerful than that of least squares used in linear regression.

b)

Consider the three assumptions from the groups- Simple, Multiple and Logistic regression respectively:

1. **Simple Linear Regression**- In this model, we assume that the independent variable and dependent variable are linear in relation, based on which we perform all the calculations. Since they are linear, we derive the regression line equation having the form of y= mx + c; which is true for every linear line. Based on this knowledge, we predict the values of dependent variables, given an independent variable. If this assumption is violated, then the equation of the

regression line based on which we predict values becomes insignificant, thereby producing a wrong output.

2. **Multiple Regression**: Homoscedasticity is assumed for this model which states that there is a constant variance in the error term. The least squares is used to predict the unknown parameters while minimizing the standard errors. Hence, if the variances across all the independent variables is not the same, there is a larger disturbance towards certain observations. This violation is called as heteroscedasticity. It also causes the standard errors to be biased. This is a serious issue as the standard errors are used in significance testing and to compute confidence intervals. In turn, it may result in improper hypothesis conclusions and may also lead up to incorrect conclusions about significance of regression coefficients.
3. **Logistic Regression**- [3][4]

   For this model, the most important assumption is that the structure of the outcome is based on the groups on which the dependent variable needs to be classified. Hence, the value of Y is the probability of the dependent variable belonging to a class of output. If this assumption is violated, then the value of Y would not be the expected value (value between 0 and 1) . It would mean that the probability for that independent variable not between 0 and 1 which is false.

Question 3: [5] [4] [6]

a) In the logistic regression model, we analyze the relationship between multiple independent variables and a categorical dependent variable. In this, we compute the probability of a given value of dependent variable belonging to a certain category by fitting it into a logistic curve.

For ex: We have a data of heights and weights of people based on which we determine whether they are male or female. Here, the independent variables are height and weight whereas the categories of output are either male/ female. Given the values of the independent variables, we plot the given variables in a logistic graph function (sigmoid). Given values of height and weight of a person, we compute the probability of that person belonging to either male or female class. We define the general logistic regression equation as follows:

$$\text{Log } (P(X) / P(1-X)) = \beta o + \beta 1 \, X$$

Since we need the categorical response variable Y, we use the logit function so that the output is the probability of a point belonging to a certain class and the value of output lies in between 1 and 0. Hence, for any given predictor values, there are only two possible error values- $1-\pi$ when the value is π or $-\pi$ when that value is $1-\pi$. This would be similar for every independent variable. Hence, the output variables and thus the errors also follow a binomial distribution.

Thus, we can summarize the above explanation that since the outputs in logistic regression gives a value of probability, the error can only lie between 0 and 1, thus concluding errors in logistic regression can’t be normally distributed.

b) In logistic regression, the dependent variable is dichotomous in nature. If the dependent response variable is binary, it has a Bernoulli distribution. When there are more than 2 response categories, it follows binomial distribution. The formula for variance for these distributions is:

$$Var(x) = N * \pi (1-\pi)$$

#N = 1 for Bernoulli Distribution

The variance represents the error of the outcome. Here, π represents the probability of a positive outcome and $1-\pi$ represents the probability of a negative outcome. Therefore, unless the value of π is the same for all independent variables, the error variance would not remain constant. As the size of the dataset used for logistic regression is high, there cannot be the same value of probability for all the independent variables. Hence, the error variance is not constant.

Question 4: [7] [8]

a) When the relationship between x and y is not linear, we transform the variables to achieve linearity so that we can apply SLR to the data. Based on the type of variable to be transformed (dependent or independent), we have different methods as follows:

Method	Transform	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	DV = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	DV = sqrt(y)	$\sqrt{y} = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	DV = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	IV = log(x)	$y = b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	DV = log(y) IV = log(x)	$\log(y) = b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

Table: Methods of transforming variables to achieve linearity [8]

Each method transforms a specific type of variable using a regression equation given in column three. Note that this equation for each non-linear method is in the form of SLR equation. The predicted values of y are transformed back again to its original scale. This process is known as back transformation.

In any case of non-linearity, the type of method suitable for the data is identified using a trial and error method. The coefficient of determination is computed for the original variables, after which a transformation method is implemented on the data. Then, the coefficient of determination is computed and compared with the original value. If the value for the transformed data is more than the original value, it states that there is a stronger linear relationship between the dependent and independent variables. Thus, the transformation was a success. If the transformed value is lower than the original value of R^2, it shows that the

transformation is a failure and thus we must try a different transformation method. This method also depends on the type of data that is used as the input.

b)

Q4.

b) To prove, $\text{Cov}(X_1, X_2) = 0$

Given: X_1 and X_2 are independent

The formula for covariance is:

$$\underset{(X,Y)}{\text{Cov}} = E[(X - \mu_x)(Y - \mu_y)]$$

$$\therefore \text{cov}(X_1, X_2) = E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]$$

$$= E[X_1 X_2 - X_1 \mu_{X_2} - X_2 \mu_{X_1} + \mu_{X_1} \cdot \mu_{X_2}]$$

$$= E[X_1 X_2) - E(X_1 \mu_{X_2}) - E(X_2 \mu_{X_1}) + E(\mu_{X_1} \cdot \mu_{X_2})$$

$$= E(X_1 X_2) - \mu_{X_2} E(X_1) - \mu_{X_1} E(X_2) + \mu_{X_1} \cdot \mu_{X_2}$$

Since $E(X) = \mu_X$ — (1)

$$\therefore = E(X_1 X_2) - \mu_{X_2} \mu_{X_1} - \mu_{X_1} \mu_{X_2} + \mu_{X_1} \mu_{X_2}$$

$$= E(X_1 X_2) - \mu_{X_2} \mu_{X_1}$$

Since X_1 and X_2 are independent -

$$E(X_1 X_2) = E(X_1) * E(X_2)$$

$$\Rightarrow E(X_1) E(X_2) = \mu_{X_2} \mu_{X_1}$$

From (1)

$$\Rightarrow \mu_{X_1} \mu_{X_2} - \mu_{X_2} \mu_{X_1}$$

$$= 0.$$

i.e. $\text{cov}(X_1, X_2) = 0$

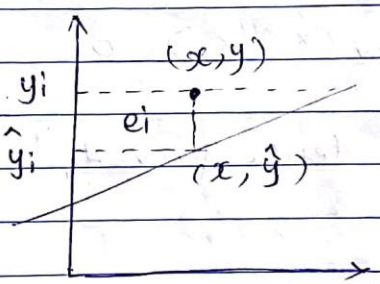
Hence proved.

Q5.

(a) The equation for least square line is:

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

Consider the graph that represents residuals:



Here y: actual value of dependent variable

$\hat{y}$: fitted value

e: residual

e = Observed response − predicted response

$$e_i = y_i - \hat{y}_i$$

$$e_i = y_i - b_0 - b_1 x$$

In the least squares technique, we find the minimum of the residual to make error as small as possible but not negative

Hence, we find.

$$S = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

We minimize this error by find the partial deriveate of S wrt the regression coefficients b_0, b_1, and equate it to zero.

$$\therefore \frac{\partial S}{\partial b_0} = 0 \quad \text{and} \quad \frac{\partial S}{\partial b_1} = 0$$

$$\therefore \frac{\partial S}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \quad -(I)$$

$$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0 \quad -(II)$$

Consider: $\frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$

$$\therefore \sum_{i=1}^{n} x_i = n\bar{x} \quad -(i)$$

Similarly

$$\sum_{i=1}^{n} y_i = n\bar{y} \quad -(ii)$$

From (I).

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} b_0 + \sum_{i=1}^{n} b_1 x_i$$

$$\Rightarrow \sum_{i=1}^{n} y_i = b_0 \sum_{i=1}^{n}(1) + b_1 \sum_{i=1}^{n} x_i$$

From (i) and (ii)

$$n\bar{y} = b_0 \cdot n + b_1 \cdot n\bar{x}$$

$$\Rightarrow \boxed{\bar{y} = b_0 + b_1\bar{x}} \quad -(III).$$

$$\therefore \quad b_0 = \bar{y} - b_1\bar{x} \quad -(IV)$$

use equation (II).

$$\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} b_0 x_i - \sum_{i=1}^{n} b_1 x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

- from (i)

$$\sum_{i=1}^{n} x_i y_i = b_0 n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2$$

From (IV)

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1 \bar{x}) n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = n\bar{x}\bar{y} - b_1 n\bar{x}^2 + b_1 \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = b_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$$

$$\therefore b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i y_i - \bar{x} y_i)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$ from (ii)

From (i)

$$b_1 = \frac{\sum_{i=1}^{n} (x_i y_i - \bar{x} y_i)}{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)} \quad \text{— (V)}$$

Now consider

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

$$= \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{x} n \bar{y} - \bar{y} n \bar{x} + \bar{x}\bar{y} n$$

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

$$= \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i) \quad \text{— (iii)}$$

Substitute (iii) in V

$$\boxed{b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \quad \text{— VI}$$

Now, substituting the value of b_1 in (IV) we get b_0:-

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

Q5)

(b) Consider the equation of the regression line:-

$$\hat{y} = b_0 + b_1 x_1$$

In the above solution Q5 (a), we derived equation III as:

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

Since both the points $(\bar{x}, \bar{y})$ satisfy the equation of regression line, we can say that $(\bar{x}, \bar{y})$ lie on the regression line.

Q6.

(a)

① $H_0 : \sigma = 15$ and $H_1 : \sigma < 15$

② Find critical values:

Since our hypothesis predicts the value of standard deviation we use the chi square X^2 test.

To find CV:

∵ It is a left tailed test

CV → intersection of $(1-\alpha)$ and df.

Here, degrees of freedom = 12 − 1 = 11.

∴ for $\alpha = 0.05$ $1-\alpha = 0.95$

critical value = 4.575.

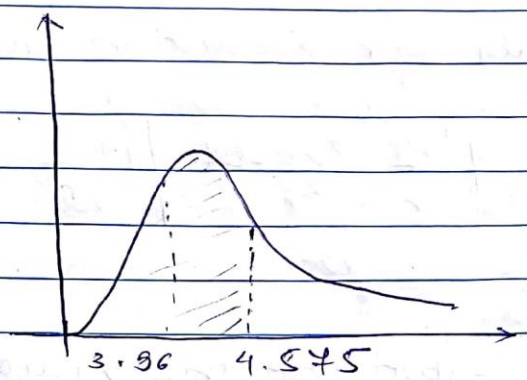
③ We compute test value using X^2 test:-

$$X^2 = \frac{(n-1)\,s^2}{\sigma^2}$$

$$X^2 = \frac{(11)(9)^2}{(15)^2}$$

$$X^2 = 3.96.$$

④ Make decision based on graph:



Since $3.96 < 4.575$, we reject the null hypothesis.

⑤ Summarize / Conclude:
We have enough evidence to support the claim that standard deviation of number of aircrafts stolen each year in United States is less than 15.

Q6.

(b)

(1) $H_0: \mu_1 - \mu_2 = 0$ and $H_1: \mu_1 - \mu_2 \neq 0$

(2) ~~Compute test value at α.~~

Find the critical values:

at $\alpha = 0.05$

$\therefore$ critical values $= \pm 1.960$.

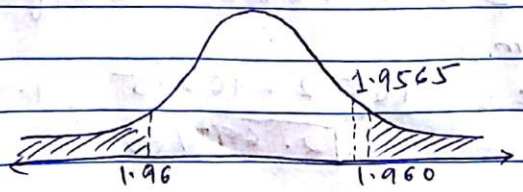
(3) Compute test value:-

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}}$$

$$= \frac{(165.2 - 162.5) - 0}{\sqrt{\frac{6.9^2}{50} + \frac{6.92}{50}}}$$

$$z = 1.9565.$$

(4) Make decision using graph



$\therefore$ accept H_0.

Since test value does not lie in the critical region, we accept H_0.

(5) Summarize:
We do not have enough evidence to support the claim. Hence, we conclude that there is no change in average weight of current batch of students.

Q6.

b)

b) using pvalue method.

(1) $H_0: \mu_1 - \mu_2 = 0$ and $H_1: \mu_1 - \mu_2 \neq 0$.

(2) Compute test value:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}}$$

$z = 1.9565$ (Refer Q6.ba(a))

From,

$$z = \frac{165.2 - 162.5}{\sqrt{2.69^2/50}} = 1.9565$$

(3) Find the P value.

for $z = 1.9565 \approx 1.96$.

$\therefore$ p value $= 0.9750$,

Since it is two tailed test, the area is doubled.

$\therefore$ p value $= 2 \times 0.9750 = 1.95$

(4) Make decision using Pvalue decision rule.

since, p value $> \alpha$

i.e. $1.95 > 0.05$

We do not reject null hypothesis.

(5) Summarize

We do not have enough evidence to support the claim that there is a change in average weight of current batch of students.

References:

[1] http://www.statisticssolutions.com/assumptions-of-linear-regression/

[2] An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain [Park, Hyeoun-Ae]

[3] Logistic and Linear Regression Assumptions: Violation Recognition and Control - Deanna Schreiber-Gregory

[4] https://en.wikipedia.org/wiki/Logistic_regression

[5] Logistic Regression, Part I: Problems with the Linear Probability Model (LPM) - Richard Williams, University of Notre Dame

[6] https://www.theanalysisfactor.com/what-is-logit-function/

[7] https://academic.macewan.ca/burok/Stat378/notes/remedies.pdf

[8] https://stattrek.com/regression/linear-transformation.aspx