ASSIGNMENT DECISION SCIENCE

SECOND SEMESTER

PGDBM (BFM)

1 ANS: Define R and R Square:

R= Correlation coefficient

R is a statistical measure of the strength of the relationship between the relative movements of two variables.

R: It is the correlation between the observed values Y and the predicted values $\hat{Y}$.

Here, R =0.969

Therefore, 96.9% of strength in the relationship between two variables.

R square= coefficient of determination.

R square measures the proportion of variation in dependent variable is explained by independent variables.

The R square (Coefficient of Determination) is the square of r(Correlation Coefficient) between predict value and true value. The whole proof is here:

$$r(Y,\hat{Y}) = \frac{\sum_i (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}}$$

$$= \frac{\sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}}$$

$$= \frac{\sum_i [(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2]}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}}$$

$$= \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}}$$

$$= \sqrt{\frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}}.$$

so

$$r(Y,\hat{Y})^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

$$r(Y,\hat{Y})^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}$$

Here, R square= 0.939

Hence, 0.939 proportion of variation in y is explained by x3, x4 and x5.

2 ANS: What is the significance of F test:

The significance F gives the probability that the model is wrong.

Statistically, the significance F is the probability that the null hypothesis is our regression model cannot be rejected.

The overall significance of the F-test demonstrates whether your linear regression model is better suited to the data than a model that does not contain any independent variables. In this article, I look at how other regression statistics, such as R-squared, fit into the F-test of overall significance. R-squared informs you how well the knowledge matches the model, and the F-test applies to it.

An F-test is a type of very versatile statistical test. In a large range of environments, you can use them. Multiple model terms can be tested simultaneously by F-tests, which help us to compare the fits of various linear models. In addition, only one word at a time can be tested by t-tests.

3 ANS: Which variable has the highest impact on Y:

X4 variable has the highest impact on y.

Because t significance is less than alpha so we reject null hypothesis and support the claim.

Also β corresponding x4 is greater than other β's.

4 ANS: Write the regression equation:

The regression equation is:

$$y = 35.681 - 0.654 * x3 + 0.233 * x4 + 0.115 * x5$$

Regression analysis is a statistical technique which can evaluate the hypothesis that one or more other variables depend on a variable. In addition, regression analysis may provide an estimation of the magnitude of the effect of a shift on another variable. Of course, this last function is critical in predicting future values.

The study of regression is based on a functional relationship between variables and further assumes that the correlation is linear. This linearity assumption is needed because, for the most part, mathematicians and econometricians have not yet worked out the theoretical statistical properties of non-linear estimation well. In economic analysis, this poses us with some difficulties since many of our theoretical models are nonlinear. For example, the marginal cost curve is obviously nonlinear, as is the total cost function, if we believe in the effect of labor specialization and the Law of Decreasing Marginal Product.

5 ANS: Calculate Y if X1= 10, X2 =20 and X3 =5:

X1= 10, x2=20, x3=5

y= 35.681 - 0.654*5

= 32.41

Q: 2 ANS)

A: The frequency distribution table showing the class interval less than 10, 11 to 20 and so on frequency of each class interval and cumulative frequency are:

classes	Freq	C.F	Midvalue(x)	fyx	(f)(x)
0-10	0	0	5.5	0	0
11-20	0	0	15.5	0	0
21-30	0	0	25.5	0	0
31-40	28	28	35.5	994	35287
41-50	10	38	45.5	455	20702.5
51-60	21	59	55.5	1165.5	64685.25
61-70	20	79	65.5	1310	85805
71-80	09	88	75.5	679.5	51302.25
81-90	14	102	85.5	1197	102343.5
91-100	18	120	95.5	1719	164164.5
	120			7520	524290

B: ANS) The mean, median and mode quartiles of grouped data are:

Mean $\bar{x} = \frac{\sum fx}{n}$

$= \frac{7520}{120}$

= 62.6667

To find median class are:

= value of $\left(\frac{n}{2}\right)^{th}$ observation

= value of $\left(\frac{120}{2}\right)^{th}$ observation

= value of 60^{th} observation

From the column of cumulative frequency cf, we find that the 60^{th} observation lies in the class 61-70.

$\therefore$ The median class is 60.5 – 70.5.

Now,

L = Lower Boundary Point Of median class = 60.5

n = Total frequency = 120

cf = cumulative frequency of the class preceding the median class = 59

f = frequency of the median class = 20

c = class length of the median class = 10

Median M = L+ $\frac{n/2 - cf}{f}$. c

$= 60.5 + \frac{60-59}{20}$. 10

$= 60.5 + \frac{1}{20}$. 10

$= 60.5$ + 0.5

= 61

To find mode class are:

Here, maximum frequency is 28.

The mode class is 30.5 – 40.5.

L = Lower Boundary point of mode class = 30.5

f_1= frequency of the mode class = 28

f_0= frequency of the preceding class = 0

f_2= frequency of the succedding class = 10

$\therefore c$ = class length of mode class = 10

$$Z = L + \left(\frac{f_1 - f_0}{2 \, . \, f_1 - f_0 - f_2}\right).c$$

$$= 30.5 + \left(\frac{28 - 0}{2\,.28 - 0 - 10}\right).10$$

$$= 30.5 + \left(\frac{28}{46}\right).10$$

= 30.5+6.087

= 36.587

C: ANS) The standard deviation, variance and range are:

Population variance σ^2 = $\frac{\sum f \, . \, x^2 - \frac{(\sum f \, . x)^2}{n}}{n}$

= $\frac{524290 \; - \; \frac{(7520)^2}{120}}{120}$

= $\frac{524290 - 471253.3333}{120}$

= $\frac{53036.6667}{120}$

= 441.9722

Range=L-S= 98-31=67

Thus range is 67.

Population standard deviation σ = $\sqrt{\dfrac{\sum f.x^2 - \dfrac{(\sum f.x)^2}{n}}{n}}$

= $\sqrt{\dfrac{524290 - \dfrac{(7520)^2}{120}}{120}}$

= $\sqrt{\dfrac{524290 - 471253.3333}{120}}$

= $\sqrt{\dfrac{53036.6667}{120}}$

= $\sqrt{441.9722}$

= 21.0231

Here, n = 120

Q1 Class:

Class with $\left(\frac{n}{4}\right)^{th}$ value of the observation in *cf* column

= $\left(\frac{120}{4}\right)^{th}$ value of the observation in *cf* column

= $(30)^{th}$ value of the observation in *cf* column

And it lies in the class 41 – 50.

Q1 Class: 40.5 – 50.5

The lower boundary point of 40.5 – 50.5 is 40.5.

L = 40.5

$Q\,1 = L + \frac{\frac{n}{4} - cf}{f} \cdot c$

$= 40.5 + \frac{30 - 28}{10} \cdot 10$

$= 40.5 + \frac{2}{10} \cdot 10$

= 40.5 + 2

= 40.5 + 2

= 42.5

Q1 class:

Class with $\left(\frac{2n}{4}\right)^{th}$ value of the observation in *cf* column

$= \left(\frac{2 \cdot 120}{4}\right)^{th}$ value of the observation in *cf* column

$= (60)^{th}$ value of the observation in *cf* column

And it lies in the class 61- 70.

Q 2 Class: 60.5 – 70.5

The lower boundary point of 60.5 – 70.5 is 60.5.

$\therefore$ L = 60.5

$$Q\,2 = L + \frac{\frac{2n}{4} - cf}{f} \cdot c$$

$$= 60.5 + \frac{60 - 59}{20} \cdot 10$$

$$= 60.5 + \frac{1}{20} \cdot 10$$

$$= 60.5 + 0.5$$

$$= 61$$

Q 3 class:

Class with $\left(\frac{3n}{4}\right)^{th}$ value of the observation in *cf* columm

= $\left(\frac{3 \cdot 120}{4}\right)^{th}$ value of the observation in *cf* column

= $(90)^{th}$ value of the observation in *cf* column

And it lies in the class 81 – 90.

Q 3 Class: 80.5 – 90.5

The lower boundary point of 60.5 – 70.5 is 60.5.

L = 60.5

Q 2 = L + $\frac{\frac{2n}{4} - cf}{f} \cdot c$

= 60.5 + $\frac{60 - 59}{20} \cdot 10$

= 60.5 + $\frac{1}{20} \cdot 10$

= 60.5 + 0.5

= 61

Q 3 class:

Class with $\left(\frac{3n}{4}\right)^{th}$ value of the observation in *cf* column

= $\left(\frac{3 \cdot 120}{4}\right)^{th}$ value of the observation in *cf* column

=$(90)^{th}$ value of the observation in *cf* column

And it lies in the class 81 – 90.

Q 3 Class: 80.5 – 90.5

The lower boundary point of 80.5 – 90.5 is 80.5.

L = 80.5

Q 3 = L + $\frac{\frac{3n}{4} - cf}{f} \cdot c$

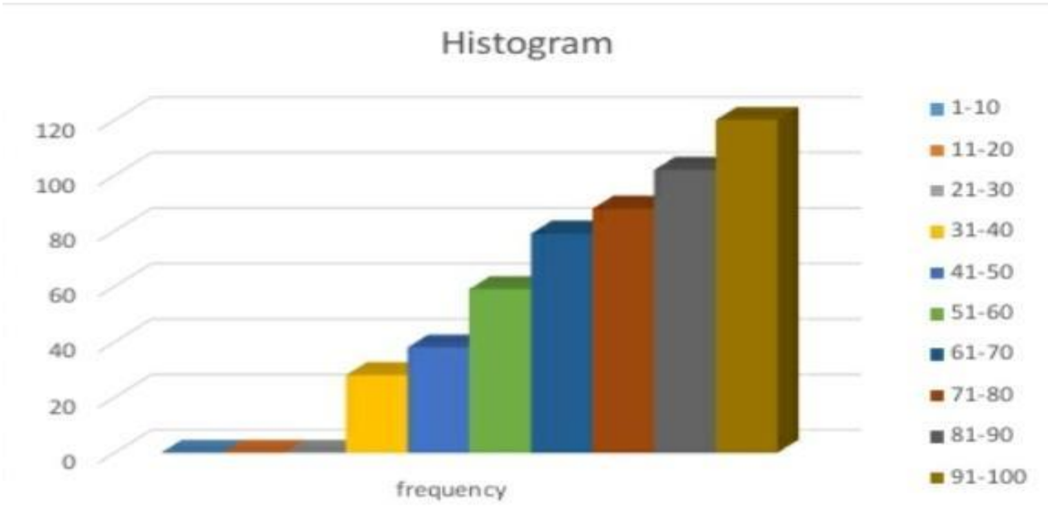
$=80.5 + \frac{90-88}{14} \cdot 10$

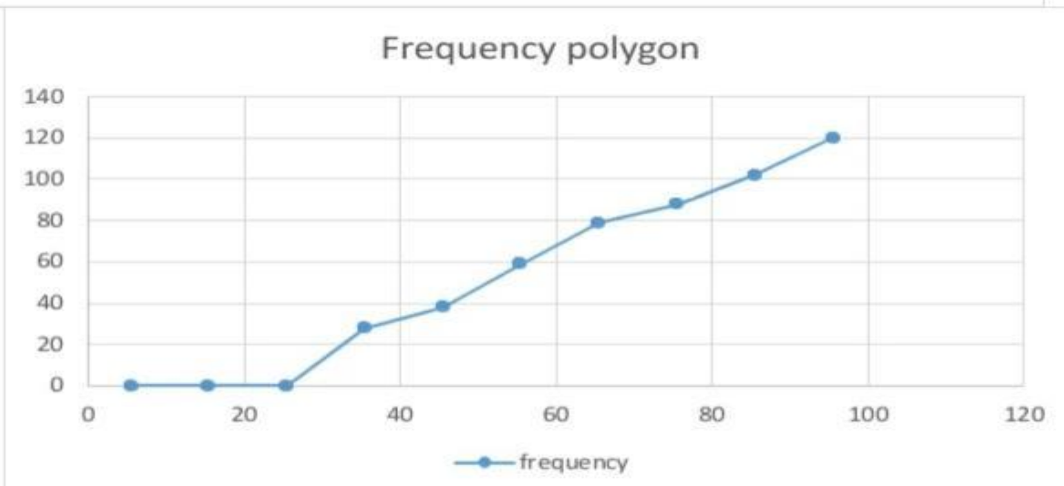
$= 80.5 + \frac{2}{14} \cdot 10$

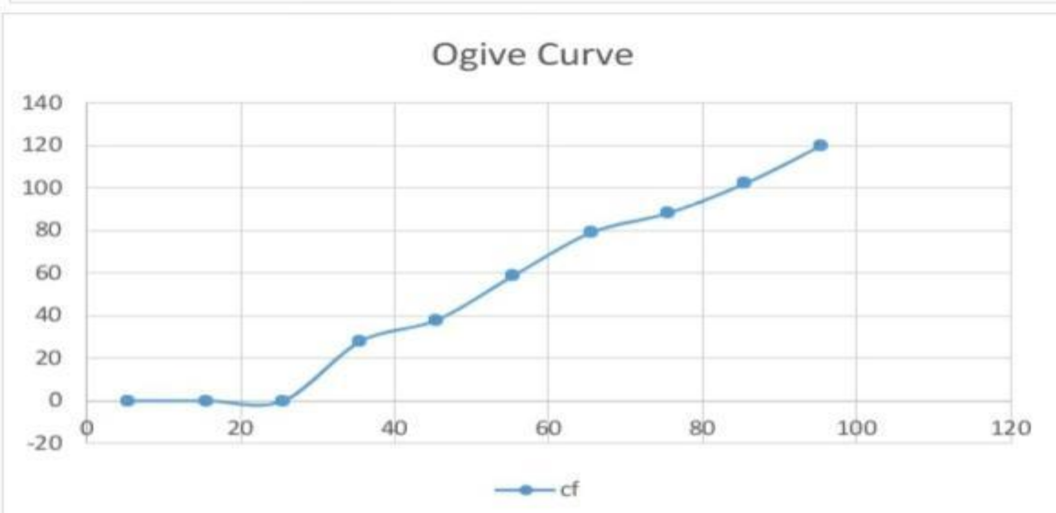
$= 80.5 + 1.4286$

$= 81.9286$

ANS D) The following diagrams are:







ANS 3:) A:)

This question is based on Bayes' probability equation.

Bayes' Theorem is used to find out a probability of certain event.

Equation of Bayes Theorem is:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Where P (A|B) is conditional probability. Probability of occurring of an event A when B is true.

P (A) and P (B) are respective probability of event A and B

Now we are defining an event

A= Email detected as spam

B=Email is a spam

B'=Email is not spam (Complement of B)

Over here we have given 50% mails are spam hence we can say that

P(B)=0.5 (Email that is a spam)

P(B')=0.5 (Email, that is not a spam)

Now certain software claims it detect 99% spam

Hence P(A|B)=0.99 (99% of the spam are detected (from all spam mail))

And probability of false positive is 5% (non spam email detected as spam)

Hence P(A|B')=0.05 (5% of times non spam email detected as spam)

Now we have to find a probability of detected email is actually not spam is

$$P(B'|A) = \frac{P(A|B').P(B')}{P(A)}$$

$$P(B'|A) = \frac{P(A|B').P(B')}{\mathrm{P\ (A|B).P(B) + P(A|B').P(B')}}$$

$$P(B'|A) = \frac{0.05 * 0.5}{(\ 0.99 * 0.5) + (0.05 * 0.5)}$$

$$P(B'|A) = \frac{0.025}{0.495 + 0.025}$$

$$P(B'|A) = \frac{0.025}{0.52} = \frac{25}{520} = \frac{5 * 5}{5 * 104} = \frac{5}{104}$$

Q: 3 B):

A) Less than 19.5 hours:

$$Given\ \mu = 20, \sigma = 2$$

$$we know that\ z = \frac{x-\mu}{\sigma}$$

$$(a)\ when\ x = 19.5,$$

$$z = \frac{19.5-20}{2} = \frac{-0.5}{2} = -0.25$$

$$p(x < 19.5)$$

$$= p(z < -0.25)$$

$$= 0.4013$$

The probability that a car can be assembled less than 19.5 *hrs* = 0.4013

ANS : b) Between 20 and 22 hours:

$$(b)\ when\ x = 20,$$

$$z = \frac{20-20}{2} = 0$$

$$when\ x = 22,$$

$$z = \frac{22-20}{2} = 1$$

$$p(20 < x < 22)$$

$$= p(0 < z < 1)$$

$$= 0.3413$$

The probability that a car can be assembled between 20 and 22 hrs = 0.3413

THE END