III-MTech(CSE)-RM (3.3)

M.Tech (CSE) 3rd Sem-2019
Sub: Research Methodology

Time: 3 Hours **Full Mark: 70**

Answer any <u>FIVE</u> questions. Marks are indicated alongside the questions.

Question 1 **(6 + 4 + 4)**

a. Fill in the blanks for the following ANOVA output of the solution to a simple regression model:

	SS	*df*	*MS*	F_0
Regression	900	?	?	?
Residual	?	23	?	
Total	1360	?		

b. Find the values of (a) number of observations, (b) standard error of the residuals, (c) *R*-square, and (d) adjusted *R*-Square.

c. Match the items in the List A to those in List B (More than one item of List A can match with one or more than one item of List B):

List A: Bernoulli distribution, *F*-statistic, ROC curve, Centroid, Odds ratio, Adjusted *R*-square, Log-likelihood ratio, and Euclidean distance

List B: *K*-Means clustering, Classification, Multiple regression, and Logistic regression

Question 2 **(2 + 2 + 2 + 2 + 6)**

Based on 10 observations, a simple regression equation has been obtained as under:

$\hat{y} = 3 + 50x$
SE: 2 4

a. Are the slope and the intercept significant at α = 0.05, given $t_{.05,\,8}$ = 1.860.
b. Write the regression equation for this model (dropping the insignificant factor, if any).
c. What is the most likely predicted value of the dependent variable *y* when *x* takes a value of 10?
d. If the observed value of *y* is 495, what is the value of the residual when *x* = 100?
e. The residuals for 10 such observations in this regression model are obtained as 5, –8, 6, –6, –8, 2, –8, –5, 3, and 4. Write the values of the coordinates of the points to be plotted on a normal probability graph paper, for testing the normality assumption of the errors.

Question 3 **(4 + 3 + 3 + 4)**

a. Give a real-life example of a classification problem which can be solved by a binary logistic regression model (do not give the example given in Question 4).
b. Develop the expression for such a model with logit as the explained variable.
c. Why can one not minimize the sum-of-the-squared error to estimate the parameter values?
d. The estimated optimal values of β_0 and β_1 of such a model have been obtained as –6.2 and 0.36. What information does it give on the probability of success (*Y* = 1)?

Question 4 **(1 4)**

A binary logistic regression model has been applied to a classification problem on testing patients for cancer. The actual occurrence of 10 cancer cases (Y: Positive, N: Negative) and their probability values that have been computed using the logistic model are given below.

	Y	Y	N	Y	Y	N	N	N	Y	N
Prob.	0.8	0.75	0.7	0.7	0.6	0.55	0.5	0.5	0.45	0.45

Draw the ROC curve for the classifier.

Question 5 **(12 + 2)**

A software company wishes to divide its 10 newly recruited programmers into 3 classes, based on their error-proneness, by using the K-means approach. Number of errors per kilo-line of code (KLOC), detected in the codes written by them in the month of October, are tabulated below.

Programmer Id	1	2	3	4	5	6	7	8	9	10
Detected Errors/KLOC	10	25	30	5	20	12	50	35	42	32

Three initial clusters were defined as under (numbers within the curls indicate the programmer id.):

C_1: {1}; C_2: {2}; C_3: {3, 4, 5, 6, 7, 8, 9, 10}

a. Find the clusters to which the programmers will be assigned.

b. The company wishes to send the most error-prone programmers for a three-month training. Which programmers would be sent for such a training?

Question 6 **(2 + 4 + 8)**

A researcher is trying to divide 20 text documents into two clusters using the K-means approach, depending on the number of pages and the number of words in each document.

Collected data on pages of manuscripts and number of words in the manuscripts vary from 10 to 1,000 and from 3,000 to 350,000, respectively. The researcher wanted to normalize the data so as to bring both of them in the range of 0 to 1.

a. Which scheme can effect such a normalization?

b. Two documents have pages 25 and 300 and words 10,500 and 150,000, respectively. What are their normalized values according to this normalization scheme?

c. If the initial clusters, C_1 and C_2, respectively, have means of 0.4 and 0.7 normalized pages and 0.25 and 0.3 normalized words, in which clusters should these two documents be included in the next iteration?

Question 7 **(3 + 2 + 2 + 2 + 2 + 1 ½ + 1 ½)**

a. For the M. Tech. thesis that you are working on (i) suggest a title, write a copyright statement, and give five keywords, (ii) state the research objectives/questions, (iii) the methods used in the literature and the methods chosen, and (iv) the contribution that your thesis is likely to make to the world of knowledge.

b. Write a sentence using each of the following Latin abbreviations: e.g. and i.e..

c. Write the SI forms of 15Kw, 20 *mg.,* and 20,00,000.

d. Write the following in scientifically accepted forms:

(i) 15 candidates were interviewed.

(ii) Out of 15 candidates five were selected.

(iii) Sita & Mohan are good friends.