INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Department of Chemical Engineering

CL 651 – Foundations of Data Science for Engineers

Assignment-5 **Function Approximation Methods** **Date:10/04/2022**

Q1. For the following data find the pearson correlation for the pairs $(x;\ y_1)$ and $(x;\ y_2)$ and comment on the relationships between variables in each pair.

x	1	3	10	2	9	6	10	1	5	2
y_1	3	9	30	6	27	18	30	3	15	6
y_2	4	12	103	7	84	39	103	4	28	7

Q2. For the data given below, find the Spearman's correlation and comment on the relation between x and y.

x	2	3	-1	2	2	-1	-2	1	4	0
y	4	9	1	4	4	1	4	1	16	0

Q3. For the data given below, Find Kendall's correlation and test the hypothesis that the two quantiles are uncorrelated

x	23	14	17	16	24	13	15	11	32	20
y	10	6	12	11	12	5	15	9	11	12

Q4. Consider following data set where x is the number of units sold and y is the profit. If the true relation between $x\ and\ y$ is linear, $y = \beta_0 + \beta_1 x$ find the parameters β_0 and β_1, that best fits the data. Comments on the validity of assumption made on the noise/error using q-q plot

x	1	2	3	4	5	6	7	8	9	10
y	4.04	6.56	10.67	13.73	16.14	19.14	21.94	25.67	27.98	30.64

Q5. If the test set is as shown below, calculate R^2 and comment on the performance of the linear model estimated in Q4.

x	25	26	27	28	29	30
y	101.75	105.75	109.75	113.75	117.75	121.75

Q6. For the problem discussed in Q4, before production, the manager derived from first principles that the relation between $y\ and\ x$ is $y = 1.25 + \beta_1 x$ with $\beta_1 = 3$. Comment whether the data given can be thought as coming from the true model given that the error variance is 1.2.

Q7. A food delivery chain was analyzing the time required to deliver the food and recorded the following information

Distance between hotel and destination (d)	8.2	10.7	9	8.2	6.4	9.7	6.6	13.4	14.5	5.8
Number of items in the order (n)	7	5	5	7	1	2	1	7	6	7
Time for delivery (t)	32.35	38.08	33.03	32.11	21.74	32.77	22.69	47.73	50.53	24.85

i. Find out the linear model relating time for delivery (t) to independent variables distance d and the number of items (n).

ii. Find r-squared and adjusted r-squared for the model built in part (i).

iii. Fit a normal distribution to the residual/error between true time and predicted time. Draw the q-q plot and comment on the validity of assumption.

Q8. A company produces 3 different items A,B and C. The data below shows the sale of these items in one day and the profit made by the company on that day

x_1 (Sales of item A)	8	11	9	8	6	10	7	13	14	6
x_2 (Sales of item B)	6	4	5	7	1	1	0	7	6	7
x_3 (Sales of item C)	49	40	23	39	6	32	7	47	26	21
Profit (y)	93.26	89.76	60.78	79.34	28.23	75.83	32.74	105.59	79.68	48.86

i. find the best multi-linear model using backward elimination

ii. find the best multi-linear model using forward elimination

Q9. Consider the data given in Q8, Build the best multi-linear model using Lasso regression.

Q.10. Manufacturing cost of cylindrical tank of various surface areas are given below: If the model to predict y is given by, $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Perform ridge regression to find the optimal parameters of the model.

Curved Surface Area (x_1)	15	17	16	15	14	16	14	19	20	14
Base Area (x_2)	7	5	5	7	1	2	1	7	6	7
Total Area (x_3)	22	22	21	22	15	18	15	26	26	21
Cost (y)	797.2	793.8	758.7	792.5	536.4	651.1	543.2	937.1	943.1	755.2