Name: Enrolment No:	UPES UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Term Examination, December 2021

Course: Predictive Modelling — **Semester: I**
Program: M. Tech. (CSE) — **Time : 03 hours**
Course Code: CSDA7002 — **Max. Marks: 100**

Instructions: Attempt all the questions. Refer appendix for required distribution tables.

Section A
(Scan and Upload) **(5Qx 4M = 20 Marks)**

Q1	In a multiple regression analysis involving 12 independent variables and 166 observations, SSR = 878 and SSE = 122. Calculate the coefficient of determination.	**CO1**
Q2	How does regression differ from correlation? Lines are characterized by their slope and intercept. What does the slope tell you about the line? What does the intercept tell you?	**CO1**
Q3	A general practice based study sought to find out if people's ears increase in size as they get older. 206 patients were studied with ear size being assessed by the length of the left external ear from the top to the lowest part. Measurements were made simply, using a transparent plastic ruler. The relation between the patient's age and ear length was examined by calculating a regression equation. The mean age of the patients was 53.75 years (range 30 - 93) and the mean ear length was 67.5 mm (range 52.0 - 84.0 mm). The linear regression equation was ear length = 55.9 + 0.22 × age What is the use of the regression equation? What are the interpretations of the numbers 55.9 and 0.22 in the regression equation? Can we conclude that the mean ear size at birth is 55.9 mm?	**CO2**
Q4	A student scored 0, 90, 88, 96, 92, 88 and 95 in his math quizzes. Answer the following questions: a) Which score may be considered as an outlier? Use Inter Quartile Range (IQR) method? b) Calculate the mean, median and mode with and without the outlier and make a decision as to which central tendency better describes a typical central score.	**CO4**
Q5	For a multiple regression model, total sum of square (TSS) = 200 and Error sum of squares (ESS) = 50. Calculate the multiple coefficient of determination.	**CO2**

Section B
(Scan and Upload) **(4Qx10M = 40 Marks)**

Q6	The two regression lines were found to be 4X–5Y+33=0 and 20X–9Y–107=0. Find the mean values and coefficient of correlation between X and Y.	**CO2**
Q7	For a multiple regression model with 35 observations and 9 independent variables (10 parameters), SSE = 134 and SSR = 289, state and test the null hypothesis that all of the regression parameters are zero at the 0.05 level. Use the following F table for the required significance level.	**CO3**

<table>
<tr><td>Q8</td><td>Regression diagnostics consists of autocorrelation, normality and homoscedasticity conditions on the residuals of the regression. Discuss these terms in brief. Assume the following (X,Y) data points:
(1,4), (2,5), (3,7), (3,7), (4,7), (5,9)
The equation for the best fit line of this data is: Y= X+3.
Compute the Durbin Watson statistic to find autocorrelation of residuals.</td><td>CO4</td></tr>
<tr><td>Q9</td><td>Discuss the method to compute the coefficient of multiple linear regression with the help of an analytical method termed as Normal Equation method.
OR
Consider the estimated regression equation: $\hat{y} = 3536 + 1183x_1 - 1208x_2$. Suppose the model is changed to reflect the deletion of x_2 and the resulting estimated simple linear equation becomes $\hat{y} = -10663 + 1386x_1$.
a) How should we interpret the meaning of the coefficient on x_1 in the estimated simple linear regression equation $\hat{y} = -10663 + 1386x_1$?
b) How should we interpret the meaning of the coefficient on x_1 in the estimated multiple regression equation $\hat{y} = 3536 + 1183x_1 - 1208x_2$?
c) Is there any evidence of multicollinearity? What might that evidence be?</td><td>CO1</td></tr>
<tr><td colspan="3">SECTION-C
(Scan and Upload) (2Qx 20M= 40 Marks)</td></tr>
<tr><td>Q10</td><td>The number of officers on duty in a Delhi and the number of robberies for that day are:
<table><tr><td>Officers</td><td>10</td><td>15</td><td>16</td><td>1</td><td>4</td><td>6</td><td>18</td><td>12</td><td>14</td><td>7</td></tr><tr><td>Robberies</td><td>5</td><td>2</td><td>1</td><td>9</td><td>7</td><td>8</td><td>1</td><td>5</td><td>3</td><td>6</td></tr></table>Calculate the regression line for this data, and the residual for the first observation, (10; 5). What percentage of variation is explained by the regression line?
OR
A study involved comparing the per capita income (in thousands) to the number of medical doctors per 10,000 residents. Six small cities in Uttarakhand had the observations:
<table><tr><td>Per capita income</td><td>8.6</td><td>9.3</td><td>10.1</td><td>8.0</td><td>8.3</td><td>8.7</td></tr><tr><td>Doctors</td><td>9.6</td><td>18.5</td><td>20.9</td><td>10.2</td><td>11.4</td><td>13.1</td></tr></table>Calculate the regression line for this data. What percentage of variation is explained by the regression line? Predict the number of doctors per 10,000 residents in a town with a per capita income of 8500.</td><td>CO3</td></tr>
<tr><td>Q11</td><td>Consider the table below. It shows three performance measures for five students. Using multiple regression, develop a regression equation to predict test score, based on IQ and the number of hours that the student studied.
<table><tr><th>Test scores</th><th>IQ</th><th>Study hours</th></tr><tr><td>100</td><td>110</td><td>40</td></tr><tr><td>90</td><td>90</td><td>32</td></tr><tr><td>80</td><td>75</td><td>28</td></tr><tr><td>70</td><td>60</td><td>24</td></tr><tr><td>60</td><td>50</td><td>20</td></tr></table></td><td>CO2</td></tr>
</table>

Appendix

Critical Values of the F-Distribution: $\alpha = 0.05$

Denom.	Numerator Degrees of Freedom									
d.f.	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236