Name:

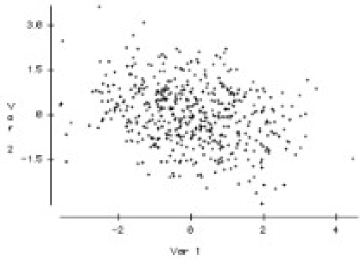
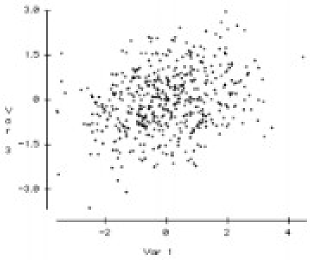
Enrolment No:

UPES
UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2019

Course: Predictive Modeling **Semester: I**
Program: MTECH Computer Science **Time : 03 hrs.**
Course Code: CSDA 7002 **Max. Marks: 100**

SECTION A

S. No.		**Marks**	**CO**
Q 1	What are the applications of predictive modeling?	**4**	**CO1**
Q 2	How to handle missing values?	**4**	**CO2**
Q 3	How to treat outliers?	**4**	**CO2**
Q 4	What is p-value and how it is used for variable selection?	**4**	**CO4**
Q 5	Difference between Linear and Multiple Regression with suitable example.	**4**	**CO1**
	SECTION B		
Q 6	How would you suggest to a franchise where to open a new store?	**8**	**CO5**
Q 7	Justify: Why might it be preferable to include fewer predictors over many?	**8**	**CO[2-3]**
Q 8	Height and weight are well known to be positively correlated. Ignoring the plot scales (the variables have been standardized), which of the two scatter plots (plot1, plot2) is more likely to be a plot showing the values of height (Var1 – X axis) and weight (Var2 – Y axis).  **Plot 1**  **Plot 2**	**8**	**CO3**
Q 9	Define: F, significance F, t Stat, P-value	**8**	**CO5**
Q 10	**Attempt any one**	**8**	

City planners believe that larger cities are populated by older residents. To investigate the relationship, data on population and median age in 6 large cities were collected.

City	Population	Median age
Chicago	2.833	31.5
Dallas	1.233	30.5
Houston	2.144	30.9
Los Angeles	3.849	31.6
New York	8.214	34.2
Philadelphia	1.448	34.2

a) Plot this data on a scatter diagram with median age as the dependent variable. (2)

b) Find the correlation coefficient (2)

c) A regression analysis was performed and the resulting regression equation is Median age = 31.4 + 0.272 population. Interpret the meaning of the slope. (4)

OR

List and discuss all the steps in developing a multivariate regression model and how to interpret all of the relevant statistics along with the necessary null and alternative hypotheses.

CO[1-4]

SECTION-C

Q 11

The following results were obtained from a multiple regression analysis.

Sum of variation	Degrees of freedom	Sum of Squares	Mean Square	Freedom
Regression	.....	288	48	.....
Error	.....	.....	20	.....
Total	.....	588	.....	.....

a) How many independent variables were involved in this model? (5)

b) How many observations were involved? (5)

c) Determine the value of the F statistic. (10)

CO[3-5]

Q 12

In the following output, some of the numbers have been accidentally erased. Recompute them, using the numbers still available. There are n = 20 in the data set.

The regression equation is: Y= (a)+0.19X				
Predictor	Coef	SE Coeff	T	P
Constant	(a)	0.43309	0.688	(b)
X	0.18917	0.065729	(c)	(d)
S= 0.67580	R-sq=31.0%			

OR

The number of murders and robberies per 100,000 population for a random selection of states are:

Murders (X)	2.4	2.7	3.6	2.6	2.1	3.3	7.6	3.7
Robberies(Y)	25.3	34.3	71.6	51.1	30	49	173	55.8

a) Create a scatter plot of the data. (4)

b) Compute the value of the correlation coefficient. (4)

c) Explain the strength, direction and form for this relationship in context. (4)

d) What is the regression equation? (4)

e) Compute the number of expected robberies when you have 3.5 murders. (4)

20

CO[1-5]