Name:

Enrolment No:

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2018

Programme Name: MTech- Computer Science and Engineering **Semester : I**
Course Name : Predictive Modelling **Time : 03 hrs**
Course Code : CSDA7002 **Max. Marks: 100**
Nos. of page(s) : 05

SECTION A

S. No.		**Marks**	**CO**
Q 1	Is it appropriate to fit a simple linear regression model that predicts survival (live or die) of a mouse using the quantitative variable of drug dosage (in milligrams) as a predictor variable? Justify your answer.	05	CO1
Q 2	A sample of two variables of size 40 produces a correlation coefficient of r = 0.682. (t_c = 2.024) a. What is the point estimate for the population correlation coefficient, ρ? b. Construct a 95% confidence interval for ρ.	05	CO3
Q 3	Illustrate the goals of a Regression Analysis in Predictive Modelling.	05	CO5
Q 4	State and compare Multicollinearity and Multiple Regression.	05	CO4

SECTION B

Q 5	Suppose we are modeling house price as depending on house size, the number of bedrooms in the house and the number of bathrooms in the house. Price is measured in thousands of dollars and size is measured in thousands of square feet. Suppose our model is: $P = 20 + 50\,\mathrm{size} + 10\,\mathrm{nbed} + 15\,\mathrm{nbath} + \epsilon, \quad \epsilon \sim N(0, 10^2).$ (a) Suppose you know that a house has size =1.6, nbed = 3, and nbath =2. What is the distribution of its price given the values for size, nbed, and nbath. (b) Given the values for the explanatory variables from part (a), give the 95% predictive interval for the price of the house. (c) Suppose you know that a house has size =2.6, nbed = 4, and nbath =3. Give	10	CO4

the 95% predictive interval for the price of the house.

(d) In our model the slope for the variable nbath is 15. What are the units of this number?

Q 6 A restaurant decides to adopt a new strategy for attracting customers. Every week it advertises in the city newspaper. To measure how well the advertising is working, the restaurant owner records the weekly gross sales for the 45 weeks after the advertising began and the weekly gross sales for the 59 weeks before the advertising.

	Mean	Standard Deviation
After	5746	859
Before	5372	901

10 CO1

(a) Assuming equal variances, can we conclude that the advertising campaign is successful? Use the p-value method, state your hypothesis and interpret your finding.

(b) Find the power for the test explained in Part (a) if the significance level is 5% and if $\mu_{\text{After}} = 5900$ and $\mu_{\text{Before}} = 5850$. Explain your findings.

Q 7 A biologist assumes that there is a linear relationship between the amount of fertilizer supplied to tomato plants and the subsequent yield of tomatoes obtained. Eight tomato plants, of the same variety, were selected at random and treated, weekly, with a solution in which x grams of fertilizer was dissolved in a fixed quantity of water. The yield, y kilograms of tomatoes was recorded.

Plant	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

10 CO2

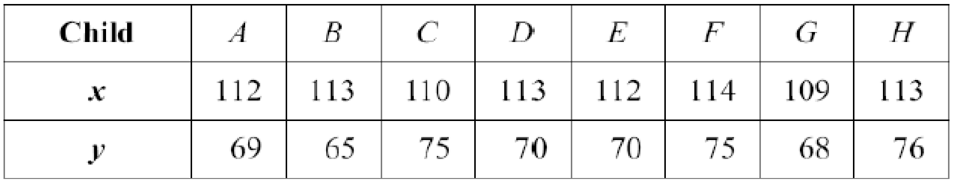
(a) Calculate the equation of the least squares regression line of y on x.

(b) Estimate the yield of a plant treated, weekly, with 3.2 grams of fertilizer.

(c) Plot a scatter diagram of yield y, against amount of fertilizer x.

OR

The table shows a Verbal Reasoning test score x and an English score y, for each of the random sample of 8 children who took both tests. 10 CO3

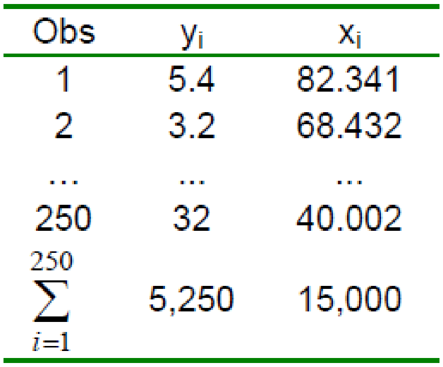


Child	A	B	C	D	E	F	G	H
x	112	113	110	113	112	114	109	113
y	69	65	75	70	70	75	68	76

(a) Calculate the value of the product moment correlation coefficient between the scores in Verbal Reasoning and English.

(b) Plot a scatter graph and comment briefly on the results obtained in (a).

Q 8 Consider y = β 0 + β1x +ε , where y is financial aid and x is parental income. Both are measured in $1,000's. A random sample 250 observations is drawn.



Obs	y_i	x_i
1	5.4	82.341
2	3.2	68.432
…	…	…
250	32	40.002
$\sum_{i=1}^{250}$	5,250	15,000

10 CO5

Σxi yi = 270,900 Σxi2 = 1,076,400 Σyi2 = 405,370

(a) Estimate the coefficients and interpret them.

(b) The parents of a particular student have an income of $92,000. Make an inference about the student's financial aid. (Answer in dollars and use α = 0.05.)

(c) Construct the 95% confidence interval for the slope. Interpret it.

SECTION-C

Q 9 A newspaper claims that last year mutual fund portfolios declined 20 percentage points on average. To test the validity of the claim, you randomly sample 121 mutual fund portfolios. The sample shows an average decline of 18.8 with a sample standard deviation of 7. 20 CO4

(a) Use 5% significance level and test the research (alternative) hypothesis that the mean decline in mutual fund portfolios is less than 20 percentage points.

(b) Find the approximate p-value of the test by using the fact that for a large sample size like 121 the Student t distribution and the Standard Normal distribution are very close. In a maximum of three sentences interpret the p-value and make a conclusion (inference).

(c) Suppose that, in truth, last year mutual fund portfolios declined 18 percentage points on average. Find the power of the hypothesis test considered in part (a) if the significance level remains at 5%. (Hint: Again, use the normal approximation to the Student t distribution.)

(d) Suppose that in testing the hypotheses we wish the power of the test to be at least 95% when the true decline in the population mean is 18 and Type I error is 5%. What sample size is needed? (Hint: Again, use the normal approximation to the Student t distribution.)

OR

A used car dealer offers a number of models for sale during a clearance event.
The following data on price, Y (in $1000's), and age of the car, X (in years), are obtained from past experience:

Y	39.9	32	25	20	16	20	13	13.7	11	12	20	9	9	12.5	7
X	1	2	4	5	6	6	10	11	11	12	12	12	12	13	15

$$\overline{X} = 8.8, \overline{Y} = 17.34, \quad \sum_{i=1}^{n} X_i Y_i = 1789.1, \quad \sum_{i=1}^{n} X_i^2 = 1430, \quad \sum_{i=1}^{n} Y_i^2 = 5685.95$$

20 CO5

(a) Estimate the regression equation for price as a function of age and interpret the parameter estimates.

(b) Is there sufficient evidence to indicate a linear relationship between selling price and age? Test the appropriate hypotheses using 5% as the significance level.

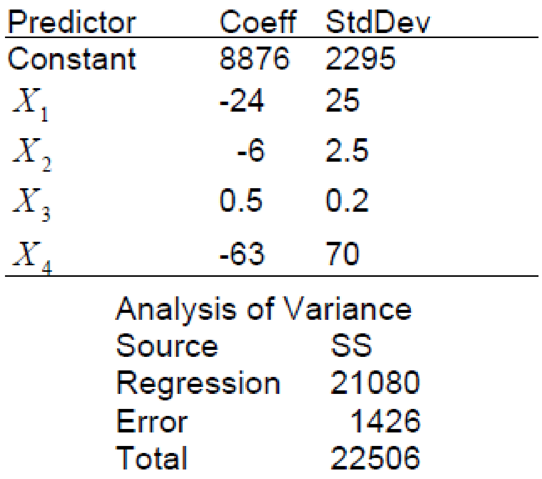
(c) The car dealer claims that a one-year increase in the age of the car reduces the price by more than $1500. Test this claim using $\alpha = 0.05$.

(d) Suppose a car owner asks, "What is the predicted selling price of my 5 year old car? Give me a range." Provide a 95% interval that can be used for such a prediction.

Q 10 The objective of a study is to estimate a multiple regression model to predict sales of cotton fabric. The explanatory variables are: 20 CO2

X1 : Whole sale price index
X2 : Quantity of Imported Fabric
X3 : Quantity of Exported Fabric
X4 : Time

With 28 observations, the following estimates are given in part of some computer output:



Predictor	Coeff	StdDev
Constant	8876	2295
X_1	-24	25
X_2	-6	2.5
X_3	0.5	0.2
X_4	-63	70

Analysis of Variance

Source	SS
Regression	21080
Error	1426
Total	22506

(a) Write down the estimated regression equation. Interpret the coefficient estimates for X2 and X3 .

(b) Test if the overall regression model is significant using a 0.05 significance level.

(c) Test the statistical significance of each regression coefficient. Use significance level 0.05.

(d) Suppose you believe that outside air temperatures (weather) may affect the sales of cotton fabric. Furthermore, suppose you have classified the weather in two categories: cool and warm. Write down a regression model that allows sales to depend on weather category while still controlling for the effects X1, X2, X3, X4 have on sales. Without doing any calculations, describe how to test the hypothesis that weather, measured as cool or warm, has a significant impact on sales of cotton fabric.

Name:

Enrolment No:

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2018

Programme Name:	**MTech- Computer Science and Engineering**	**Semester**	**: I**
Course Name	**: Predictive Modelling**	**Time**	**: 03 hrs**
Course Code	**: CSDA7002**	**Max. Marks:**	**100**
Nos. of page(s)	**: 04**		

SECTION A

S. No. **Marks** **CO**

Q 1 Compute the simple linear regression equation if: 05 CO1

	mean	stdev	correlation
x	163.5	16.2	-0.774
y	874.1	54.2	

Q 2 The capacity of an office-building elevator is 8 persons or 640 kg. Assume that the weights of the office workers are normally distributed with mean equal to 75 and standard deviation equal to 9 kg. What is the probability that the total weight of 8 randomly selected office workers exceeds the capacity of an elevator? 05 CO3

Q 3 A researcher has a large number of data pairs (age, height) of humans from birth to 70 years. He computes a correlation coefficient. 05 CO5

a. Would you expect it to be positive or negative? Why?

b. What would you suggest to be a major problem with this approach?

Q 4 State and compare Multicollinearity and Multiple Regression. 05 CO4

SECTION B

Q 5 Suppose we are modeling house price as depending on house size, the number of bedrooms in the house and the number of bathrooms in the house. Price is measured in thousands of dollars and size is measured in thousands of square feet. Suppose our model is: 10 CO4

$$P = 20 + 50\,\text{size} + 10\,\text{nbed} + 15\,\text{nbath} + \epsilon, \quad \epsilon \sim N(0, 10^2).$$

(e) Suppose you know that a house has size =1.6, nbed = 3, and nbath =2. What is the distribution of its price given the values for size, nbed, and nbath.

(f) Given the values for the explanatory variables from part (a), give the 95% predictive interval for the price of the house.

(g) Suppose you know that a house has size =2.6, nbed = 4, and nbath =3. Give the 95% predictive interval for the price of the house.

(h) In our model the slope for the variable nbath is 15. What are the units of this number?

Q 6 Using the following data, perform a one way analysis of variance using α=.05.Write up the results in APA format. 10 CO1

$$\begin{bmatrix} \text{Group1} \\ 51 \\ 45 \\ 33 \\ 45 \\ 67 \end{bmatrix} \begin{bmatrix} \text{Group2} \\ 23 \\ 43 \\ 23 \\ 43 \\ 45 \end{bmatrix} \begin{bmatrix} \text{Group3} \\ 56 \\ 76 \\ 74 \\ 87 \\ 56 \end{bmatrix}$$

Q 7 A mathematics teacher recorded the length of time, y minutes, taken to travel to school when leaving home x minutes after 7 am on seven selected mornings. The results are as follows.

x	0	10	20	30	40	50	60
y	16	27	28	39	39	48	51

10 CO2

(a) Plot the data on Scatter Diagram.

(b) Calculate the equation of least square regression line of y on x, writing your answer in form y=a+bx

(c) The mathematics teacher needs to arrive at school no later than 8:40 am. The number of minutes by which the mathematics teacher arrives early at school, when leaving home x minutes after 7 am, is denoted by z.
Deduce that Z= (100 - a) – (1 + b) x

OR

Consider the research question of whether woman exercise more regularly than men do. A random sample of 200 women and 150 men yields these results:

	Men	Women
Exercise regularly	88	130
Do not exercise regularly	62	70
Total	150	200

10 CO3

(a) Construct a 95% confidence interval estimate of the difference in the proportion of women and men who exercise regularly. Interpret the interval.

(b) Conduct a hypothesis test to determine if women exercise more regularly than men do. For a 5% significance level, find the standardized rejection region AND the p value. Make a conclusion for both the rejection region approach and p-value approach.

(c) If 65% of the women and 55% of the men exercise regularly, what is the power of your test in Part (b) for α = 0.05?

Q 8 A chain of stores sells radio and video equipment. It gathers data on 15 stores to study the relationship between sales volume, the local population size and parking. Below is a partial computer printout from a regression analysis.

y	sales volume ($1,000's)
x_1	number of households within a 30 km radius of the store's location (1,000's)
x_2	=1 if the store has a free parking lot and 0 otherwise

Predictor	Coef	Standard Error
Constant	15.0	6.20
x_1	0.87	0.04
x_2	28.4	4.46

$R^2 = 70\%$

10 CO5

Analysis of Variance:

	SS	*df*	*MS*	*F*
Regression				
Error				
Total	1000			

(a) Write out the estimated regression equation. Interpret the estimated coefficient on x1 taking into account the units of measurement.

(b) Test the overall statistical significance of the model.

SECTION-C

Q 9 A newspaper claims that last year mutual fund portfolios declined 20 percentage points on average. To test the validity of the claim, you randomly sample 121 mutual fund portfolios. The sample shows an average decline of 18.8 with a sample standard deviation of 7. 20 CO4

(a) Use 5% significance level and test the research (alternative) hypothesis that the mean decline in mutual fund portfolios is less than 20 percentage points.

(b) Find the approximate p-value of the test by using the fact that for a large sample size like 121 the Student t distribution and the Standard Normal distribution are very close. In a maximum of three sentences interpret the p-value and make a conclusion (inference).

(c) Suppose that, in truth, last year mutual fund portfolios declined 18 percentage points on average. Find the power of the hypothesis test considered in part (a) if the significance level remains at 5%. (Hint: Again, use the normal approximation to the Student t distribution.)

(d) Suppose that in testing the hypotheses we wish the power of the test to be at least 95% when the true decline in the population mean is 18 and Type

I error is 5%. What sample size is needed? (Hint: Again, use the normal approximation to the Student t distribution.)

OR

A researcher is investigating the prevalence of working laptop computers among students and would like to get a better sense of how common it is for students to have one.

(a) Outline how to do a statistical analysis to address the research question such that a researcher with appropriate data could simply follow your outline. Include necessary formulas.

(b) Consider collecting a sample of students by randomly selecting classes, asking students with a working laptop to raise their hands, and counting the total number of students and the number that raised their hands. Make an argument that this data collection method could yield a biased estimate of the parameter of interest. Indicate the direction of bias and whether it would affect both the point and interval estimates. 20 CO5

(c) Provide a detailed recommendation about how many students should be sampled to address the goal. Your answer should recommend a sample size (a number or range of numbers) with supporting work and justification. You will need to make some specific choices to answer this part. Provide a brief justification for each choice.

Q 10 A newspaper claims that last year mutual fund portfolios declined 20 percentage points on average. To test the validity of the claim, you randomly sample 121 mutual fund portfolios. The sample shows an average decline of 18.8 with a sample standard deviation of 7.

(a) Use 5% significance level and test the research (alternative) hypothesis that the mean decline in mutual fund portfolios is less than 20 percentage points.

(b) Find the approximate p-value of the test by using the fact that for a large sample size like 121 the Student t distribution and the Standard Normal distribution are very close. In a maximum of three sentences interpret the p-value and make a conclusion (inference). 20 CO2

(c) Suppose that, in truth, last year mutual fund portfolios declined 18 percentage points on average. Find the power of the hypothesis test considered in part (a) if the significance level remains at 5%. (Hint: Again, use the normal approximation to the Student t distribution.)

(d) Suppose that in testing the hypotheses we wish the power of the test to be at least 95% when the true decline in the population mean is 18 and Type I error is 5%. What sample size is needed? (Hint: Again, use the normal approximation to the Student t distribution.)