Introduction to Linear Models and Regression

Full Marks: 50 Time : 3 hrs

Group A is mandatory. From Group B, answer as many as you want but maximum you can score in this group is 30

[Bold-faced letters are used to denote vectors]

Group A

1. An experiment was conducted to determine the effects of different dates of planting and different methods of planting of sugarcane. The following data shows the yields of sugarcane (in quintal). Carry out an analysis of variance for the data.

Method of planting	Date of Planting			
	October	November	February	March
I	7.10	3.69	4.70	1.90
II	10.29	4.79	4.58	2.64
III	8.30	3.58	4.90	1.80

Consider $F_{0.05\ 2,\ 6} = 5.14$, $F_{0.05\ 3,\ 6} = 4.76$, $F_{0.05\ 2,\ 11} = 3.98$, $F_{0.05\ 3,\ 11} = 3.59$,
$t_{0.05,\ 6} = 1.943$, $t_{0.025,\ 6} = 2.447$, $t_{0.05,\ 11} = 1.796$, $t_{0.025,\ 11} = 2.201$ (12)

2. Suppose a group of 500 workers in a factory who have been diagnosed as having breathing problem, are identified as case group and a group of 1000 workers who do not have breathing problem are identified as control group. Some of the members of these groups were exposed to a particular chemical suspected of causing breathing problem. The data on chemical exposure and breathing problem are given in the following table:

	Breathing Problem (Case)	No Breathing Problem (control)
Chemical Exposure	112	216
No Chemical Exposure	388	784

Fit a logistic regression model for examining the effect of chemical exposure on the breathing problem. (8)

Group B

3. Under the usual notations of general linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, if $\widehat{\boldsymbol{\beta}}$ is any particular solution to the corresponding set of normal equations, then show that it minimizes the residual sum of squares. (6)

4. Let **X** be distributed as $N_p(\boldsymbol{\mu}, \Sigma)$. Define any two fixed p X 1 vectors **l** and **λ** orthogonal to **μ** such that $\mathbf{l}'\Sigma\boldsymbol{\lambda} = 0$. If $Y_1 = \mathbf{l}'\mathbf{X}$ and $Y_2 = \boldsymbol{\lambda}'\mathbf{X}$, then find $P(Y_1 > 0, Y_2 > 0)$. (6)

5. Define $r_{1(2\,.\,34\ldots p)} = \text{Corr}(x_1, e_{2\,.\,34\ldots p})$ as the correlation coefficient between x_1 and the residual of x_2 given $x_3, x_4, \ldots, x_p$. Then show that, $r^2_{1(2\,.\,34\ldots p)} \leq r^2_{12.34\ldots p}$, where $r^2_{12.34\ldots p}$ denotes the partial correlation coefficient between x_1 and x_2 excluding the impacts of the rest of the variables. (8)

6. Prove that – if every BLUE is expressed in terms of the observations **Y** as **a**'**Y,** the coefficient vector **a** is a linear combination of the columns of X and vice-versa. (8)

7. Consider the following data.

Sequence no.	X	Y
1	3.3	1.8
2	4	2.2
3	5.3	3.5
4	5.7	3.4
5	4	2.8
6	5.3	2.8
7	2	2.8
8	2	1.5
9	6	3.2
10	5.3	2.1
11	3.7	3.7
12	1.3	2.3
13	6	3
14	6.3	3
15	4.7	1.9
16	6.7	5.9
17	2.7	2.2
18	5	1.8
19	3.7	1.7
20	4	2.8
21	4.7	3.2
22	3.3	3.8
23	1.3	1.8

The regression line of y on x is found to be Y = 1.426 + 0.316x. The sum of squares due to regression is 5.499 with degrees of freedom 1 and the sum of squares due to residual is 15.278 with degrees of

freedom 21. Using the repeated observations in the data, check whether the model adequately fits the data or not.

Consider $F_{0.05\ 1,\ 21} = 4.32$, $F_{0.05\ 11,\ 10} = 2.85$ (8)

8. Define any unbiased estimator $\mathbf{a}'\mathbf{Y}$ of the parametric function $\boldsymbol{\lambda}'\boldsymbol{\beta}$. Show that P can be used as a projection operator on $\mathbf{a}'\mathbf{Y}$ to find the corresponding BLUE. What will happen if the same projection operator is applied again to BLUE, thus found? Justify your answer. (8)