5

RAJARATA UNIVERSITY OF SRI LANKA
FACULTY OF APPLIED SCIENCES

B.Sc. (General) Degree in Applied Sciences
Third Year - Semester I Examination – November/December 2016

MAT 3203 – Regression Analysis

Time: Two (2) hours

Answer **All** Questions.

Calculators and **statistical tables** will be provided.

01

a) Match the statements below with the corresponding terms from the list. **(40 marks)**

a. Multicollinearity	g. Dummy variable
b. Extrapolation	h. Multiple regression model
c. R^2 adjusted	i. R^2
d. Quadratic regression	j. Residual
e. Residual plot	k. Influential points
f. Fitted equation	l. Outliers

i. Used when a numerical predictor has a curvilinear relationship with the response.
ii. Worst kind of outlier, can totally reverse the direction of association between x and y.
iii. Used to check the assumptions of the regression model.
iv. Used when trying to decide between two models with different numbers of predictors.
v. Proportion of the variability in y explained by the regression model.
vi. Is the observed value of y minus the predicted value of y for the observed x.
vii. A point that lies far away from the rest.
viii. Can give bad predictions if the conditions do not hold outside the observed range of x's.

ix. $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$

x. $\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$

xi. Problem that can occur when the information provided by several predictors overlaps.

xii. Used in a regression model to represent categorical variables.

b) The following table shows five observations of a response variable y and exploratory variable x, data follows a quadratic model $y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$; $\varepsilon_i \sim N(0, \sigma^2)$. **(60 marks)**

y_i	0.8	1.3	2.0	1.4	0.4
x_i	2.5	3.0	4.0	5.0	6.0

$$(X'X)^{-1} = \begin{pmatrix} 2.623 & -0.694 & 0.023 \\ -0.694 & 0.223 & -0.012 \\ 0.023 & -0.012 & 0.0014 \end{pmatrix}, X'Y = \begin{pmatrix} 5.900 \\ 23.299 \\ 98.100 \end{pmatrix}$$

i. Use the above information to estimate the $\hat{\beta} = (\beta_0, \beta_1, \beta_2)'$

ii. Find the dispersion matrix of the parameter vector $D(\hat{\beta})$.

iii. Find the 95% confidence interval for β_2.

02

a) Let $Y_1, Y_2, \ldots, Y_n$ are a set of uncorrelated random variables with common variance σ^2 and $E[Y_i] = \beta(X_i - \bar{X})$ for $i = 1,2,\ldots,n$ where X_i are known constants. **(50 marks)**

i. Determine the least square estimator, $\hat{\beta}$ of β.

ii. Show that the least square estimator of $\hat{\beta}$ is a linear function of Y_i.

iii. What is the distribution of $\hat{\beta}$? (Find the Mean and the Variance)

b) 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 were recorded. **(50 marks)**

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

i. Using a suitable graphical method comment on the relationship between the dissolved fertilizer and yield.

ii. Find regression coefficients and write down the fitted regression line of y on x.

iii. Estimate the yield of a plant weekly with 3.2 grams of fertilizer

03 In a study a random sample of 10 trees for a particular tree species were examined and the diameter and the age of each tree were recorded in order to find out whether there is a linear association exists between diameter and age. The following Excel output shows how the data were analyzed.

(100 marks)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.924895328
R Square	0.855431369
Adjusted R Square	0.83736029
Standard Error	**(A)**
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**(B)**	**(C)**	77.12911	**(D)**	0.000127045
Residual	**(E)**	13.03488608	**(F)**		
Total	9	**(G)**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	8.203062291	0.763263651	**(H)**	4.94364E-06
X Variable 1	0.20076296	**(I)**	**(J)**	0.000127045

$S_{xx} = 6620, S_{yy} = 1566.32\,, S_{xy} = 384.18$

i. What are the assumptions used in regression analysis.

ii. Find the value of A.

iii. Calculate H, I, J and Construct 95% confidence interval on β_0 and β_1.

iv. Test the hypothesis $H_0: \beta_i = 0$ Vs $H_1: \beta_i \neq 0$ at 5% significance level for i = 0 and 1.

v. Fill the values B, C, D, E, F and G.

vi. Write the relevant hypothesis for analysis of variance and state your conclusion.(Use 5% significance level)

vii. Interpret the model using coefficient of determination.

viii. What can you say about the appropriateness of the model.(Use $\alpha = 0.05$)

ix. Calculate the sample correlation coefficient (r) and interpret it.

x. Test whether the population correlation coefficient is significant or not (Use $\alpha = 0.05$).

04

a) Explain the usefulness of the following statistics in regression analysis. **(20 marks)**

i. Coefficient of determination

ii. Mallow's C_p statistics

iii. Adjusted Coefficient of multiple determination

iv. Mean Square Error (MSE)

b) Technicians measure heat flux as part of a solar thermal energy test. An energy engineer wants to determine how total heat flux is predicted by other variables: insolation(X_1), the position of the east (X_2), south (X_3), and north focal points (X_4), and the time of day (X_5). Therefore 13 models were built using different combinations of independent variables X_1, X_2, X_3, X_4 and X_5. Some statistics obtained for each model are as follows.

(50 marks)

Model	No of variables	R^2	$Adj\ R^2$	C_p	SSE	Variables				
						X_1	X_2	X_3	X_4	X_5
a	1	72.1	71.0	38.5	12.328				X	
b	1	39.4	37.1	112.7	18.154	X				
c	1	12.3	9.1	174.2	21.834					X
d	2	85.9	84.8	9.1	8.9321			X	X	
e	2	82.0	80.6	17.8	10.076				X	X
f	2	73.4	71.3	37.5	12.259	X			X	
g	3	87.4	85.9	7.6	8.5978		X	X	X	
h	3	86.5	84.9	9.7	8.9110	X		X	X	
i	3	86.4	84.7	10.0	8.9448			X	X	X
j	4	89.1	87.3	5.8	8.1698	X	X	X	X	
k	4	88.0	86.0	8.2	8.5550	X		X	X	X
l	4	87.5	85.4	9.4	8.7487		X	X	X	X
m	5	89.9	87.7	6.0	8.0390	X	X	X	X	X

i. Select the best 1 variable, 2 variable, 3 variable and 4 variable models and justify your answer.

ii. By giving reasons select the best model among all these 13 models.

c) **Briefly describe the procedure of forward selection method using an example.**

(30 marks)

END