MURANG'A UNIVERSITY OF TECHNOLOGY

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND ACTUARIAL SCIENCE

UNIVERSITY ORDINARY EXAMINATION

2018/2019 ACADEMIC YEAR

4TH YEAR **1ST** SEMESTER EXAMINATION FOR, BACHELOR OF SCIENCE APPLIED STATISTICS WITH PROGRAMMING

AMS 427 STATISTICAL MODEL BUILDING

DURATION: 2 HOURS

DATE: 25/04/2019

TIME: 9:00-11:00 AM

Instructions to candidates:

1. Answer question One and Any Other Two questions
2. Mobile phones are not allowed in the examination room.
3. You are not allowed to write on this examination question paper.

SECTION A: ANSWER ALL QUESTIONS IN THIS SECTION

QUESTION ONE (30 MARKS)

a) i) Define Ridge Regression analysis as used in statistical model building. 2marks

ii) Discuss any three model validation techniques used in statistical model building 3marks

b) The following output was obtained after filling a multiple regression model in R.

Call:

Lm (formula =$y \sim x_1 + x_2 + x_3 + x_4$, data = data)

Residuals

Min	IQ	Median	3Q	Max
-2.303	-1.2524	-0.2449	1.0718	4.3589

Coefficients

	Estimate	**Std. Error**	**t value**	**Pr (>t)**	
(Intercept)	39.76	6.32	6.28	0.000237	xxx
X_1	1.890	0.223	8.459	0 .000029	xxx
X_2	0.659	0.04709	14.764	0.0000004	xxx
X_3	0.4552	0.212	2.147	0.064	
X_4	0.088	0.054	1.620	0.143	

Residual standard error 2.128 on degrees of freedom

Multiple R-squared 0.987, Adjusted R-squared 0.98

F- statistic 147.9 on 4 and 8 df p-value 1.57 x 10^{-7}

i) Write the model. 2marks

ii) Explain all the parameters involved. 2marks

iii) Discuss the goodness of fit of the model. 1mark

c) Outline the steps for variable selection and model building. 7marks

d) For the models shown below, determine whether it is linear, an intrinsically linear model or a non-linear model. If the model is intrinsically linear, show how it can be linearised by a suitable transformation. 5marks

i $y = \theta_1 e^{\theta_2 + \theta_3 x} + \varepsilon$

ii) $y = \theta_1 + \theta_2 x_2 + \varepsilon$

iii) $y = \theta_1 + \frac{\theta_2}{\theta_1} + \epsilon$

iv) $y = \theta_1 x_1^{\theta_2} + \varepsilon$

v) $y = \theta_1 + \theta_2 e^{\theta_3 x} + \varepsilon$

e) Consider the following observations

X	Y	
0.5	0.68	1.58
1	0.45	2.66
2	2.50	2.04
4	6.19	7.85
8	56.1	54.2
9	89.8	90.2
10	147.7	146.3

Write a well documented programme in R that does the following

i) Reads the data in R. 1mark

ii) Fits the non-linear regression model.

$y = \theta_1 e^{\theta_2 x}$ 2marks

iii) Test the significance of regression. 1mark

iv) Estimates the error variance δ^2 2marks

v) Analyses the residuals from this model 2marks

SECTION B – ANSWER ANY TWO QUESTIONS IN THIS SECTION

QUESTION TWO (20 MARKS)

a) Discuss the following methods as used in evaluating subset regression models clearly showing the underlying equations.

i. Coefficient of multiple determinations. 4marks

ii. Residual mean square. 4marks

iii. Alkaike information criterion. 4marks

iv. Bayesian information criterion. 4marks

b) Discuss the data splitting technique as applied in model validation clearly showing underlying equations. 4marks

QUESTION THREE (20 MARKS)

a) The data below concerns the heat evolved in calories per gram of cement (y) as a function of the amount of each of four ingredients in the mix; aluminate (x_1), silicate (x_2), ferrite (x_3) and dicalcium (x_4).

Y = 78.5, 74.3, 104.3, 87.6, 95.9, 109.2, 102.7, 72.5, 93.1, 115.9, 83.8, 113.3, 109.4.

X_1 = 7, 1, 11, 11, 7, 11, 3, 1, 2, 21, 1, 11, 10

X_2 = 26, 29, 56, 31, 52, 55, 71, 31, 54, 47, 40, 66, 68

X_3 = 6, 15, 8, 8, 6, 9, 17, 22, 18, 4, 23, 9, 8

X_4 = 60, 52, 20, 47, 33, 22, 22, 6, 44, 22, 26, 34, 12

Write a well documented programme in R that does the following

i) Reads the data in R. 4marks

ii) Regresses Y on X_1, X_2, X_3, X_4 through multiple linear regression with the intercept and gives the summary. 2marks

iii) Performs the steps backward elimination to determine the best subset regression model

2marks

iv) Perform the step forward elimination to determine the best subset regression model 2mks

v) Splits the data into two i.e. the estimation and prediction. 3mks

vi) Performs ridge regression of Y on X_1, X_2, X_3, and X_4. 2marks

b) Discuss the steps in response surface methodology stating the underlying equations. 5marks

QUESTION FOUR (20 MARKS)

a) The general form of the logistic regression model is given by

$$y_i = E(y_i) + \varepsilon_i$$

Where the observations y_i are independent Bernoulli random variables with expected values.

$$E(y_i) = \pi_i = \frac{\exp(x_i'b)}{1 + \exp(x_i'b)}$$

Use the method of maximum likelihood to estimate the parameters in their linear predictor

$$x_i'b$$

10marks

b) The Poison probability density function is

$$f(y,\lambda) = \frac{\lambda^y e^{\lambda}}{y!}$$

Show that the Poisson distribution is a member of the exponential family. 10marks