TECHNICAL UNIVERSITY OF MOMBASA

FACULTY OF APPLIED AND HEALTH SCIENCES

DEPARTMENT OF MATHEMATICS & PHYSICS

UNIVERSITY EXAMINATION FOR:

FOR THE FIRST SEMESTER IN THE FOURTH YEAR OF BACHELOR OF SCIENCE IN

MATHEMATICS & COMPUTER SCIENCE, BACHELOR OF SCIENCE IN STATISTICS & COMPUTER SCIENCE

AMA 4411: REGERSSION MODELLING

SPECIAL/ SUPPLIMENTARY EXAMINATIONS

SERIES: September 2018

TIME: 2HOURS

DATE: Pick DateSep2018

Instructions to Candidates

You should have the following for this examination

-Answer Booklet, examination pass and student ID

This paper consists of **FIVE** questions. Attemptquestion ONE (Compulsory) and any other TWO questions.

Do not write on the question paper.

Question ONE (30 Marks)

(a) Write notes on each of the following.

(i) Multicollinearity. (5 marks)

(ii) Heteroscedasticity. (5 marks)

(b) Each of the following problems could occur in fitting a multiple linear regression model. For each of the problems suggest a possible solution. Justify your answers.

(i) The **X'X** matrix contains some very large values. (2 marks)

(ii) Multicollinearity exists. (2 marks)

(iii) The residuals are uncorrelated but have a non-constant variance. (2 marks)

(iv) The overall model is statistically significant, but none of the individual parameter estimates is significant. (2 marks)

(c) Statistical packages contain various influence diagnostics.

(i) Explain what the hat matrix H is in the analysis of the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \left(\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}\right).$$

Obtain an expression for $\hat{\varepsilon}$ in terms of H.

[You may assume that the OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\ .]$$

State the properties of the residuals, and explain how the diagonal elements of the hat matrix are used to decide whether or not a data point (xi, yi) is influential. Explain how this helps in constructing a model. (6 marks)

(ii) Explain in detail how each of the following can be useful in constructing a model:

(1) Cook's distance

(2) DFFITS (6 marks)

Question 2 (20 Marks)

A group of astronomers carried out a study of the relationship between light intensity and surface temperature. Data gathered on 24 stars in the cluster CYG OB1 are given in the table below. Note that there are three outlying points indicated by an asterisk (*)

Log surface temperature (x)	*Log light intensity (y)*	*Log surface temperature (x)*	*Log light intensity (y)*	*Log surface temperature (x)*	*Log light intensity (y)*
4.37	5.23	4.56	5.74	4.23	3.94
4.26	4.93	4.56	5.74	4.23	4.18
4.30	5.19	4.46	5.46	4.29	4.38
3.48*	6.05	4.57	5.27	4.42	4.42
4.26	5.57	4.37	5.12	4.42	4.18
3.49*	5.73	4.43	5.45	3.49*	5.89
4.48	5.42	4.43	5.57	4.29	4.22
4.29	4.26	4.42	4.58	4.49	4.85

* indicates outlying point

(a) A regression analysis of the full data set was performed using a statistical package and produced the following output

Predictor	Coeff	St dev	*t*	*p*		
Constant	7.74	1.73	4.48	<0.001		
Slope	–0.628	0.403	–1.56	0.134	$s = 0.6207$	$R^2 = 9.9\%$

A quick look at the data suggests that there is a positive relationship between surface temperature and light intensity. However, the estimate of the slope is negative. Why is this? (4marks)

(b) The astronomers decided to asses the impact of the three outlying data points by deleting them and then calculating the following summaries.

$$\sum x = 92.13 \qquad \sum y = 103.70 \qquad \sum x^2 = 404.4303$$

$$\sum y^2 = 519.0608 \qquad \sum xy = 455.7101$$

Estimate the slope of the regression line after removing the outlying points and test the hypothesis that the slope is zero. (4marks)

(c) Construct a scatter plot of the full data set. Explain why the estimated slopes in (i) and (ii) have different signs (8 marks)

(d) What conclusions do you draw from these analyses? (4 marks)

Question 3 (20 Marks)

The table below shows the UK Gross Domestic Product (GDP, y) for the years 1989 − 1999, with years also coded as t = year − 1994. The figures are given in units of £bn, and are expressed in 1995 prices.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
t	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	655.2	659.5	649.8	650.3	665.4	694.6	714.0	732.2	757.9	777.9	794.4

Source: United Kingdom National Accounts, 2000 edn, table 1.1.

Note: $\Sigma y = 7751.2$, $\Sigma y^2 = 5491108.76$, $\Sigma t = 0$, $\Sigma t^2 = 110$, $\Sigma ty = 1706.3$.

(a) A model $y = \alpha + \beta t + \varepsilon$, where ε is a random error term with the usual properties, is proposed for the data. Obtain least squares estimates of α and β, and calculate r^2 (the coefficient of determination). Also estimate σ^2, the variance of ε, and obtain estimates of the standard errors of the coefficients α and β. (6 marks)

(b) What are "the usual properties" of the errors? How realistic is the assumption that the errors have these properties? (You are not expected to describe or conduct any tests.) (2marks)

(c) Interpret the value of r^2, and the values of your estimates of α and β. (3marks)

(d) Draw a time chart of the data and superimpose your estimated function on it. (5marks)

(e) Use your estimated model to predict GDP in 2000 and 2010, and comment on your predictions in the light of your graph. (4marks)

Question 4 (20 Marks)

A scientist has collected a set of data (*xi*, *yi*) in a situation in which he believes that the underlying model is of the form $y = ae^{bx}$.

He has read about two ways of fitting such a model and does not know how to proceed.

(a) One method is to use log(*y*) rather than *y* as the response variable.

(i) Write down the appropriate statistical form of the model, stating the assumptions with regard to the distribution of errors for the *untransformed* data. (2 marks)

(ii) Derive the normal equations for this model and hence derive expressions for the parameter estimates. (5marks)

(b) The second method uses the statistical model

$$y_i = ae^{bx_i} + \varepsilon_i .$$

For a non-linear model $y_i = f(x_i) + \varepsilon_i$, normal equations for any parameters can be obtained by differentiating $S = \Sigma\{y_i - f(x_i)\}^2$ suitably and equating the derivatives to zero. Derive the normal equations for the model $y_i = ae^{bx_i} + \varepsilon_i$ and hence derive a nonlinear equation for the estimate of the parameter *b*, and an expression for the estimate of *a* in terms of that of *b*. (6 marks)

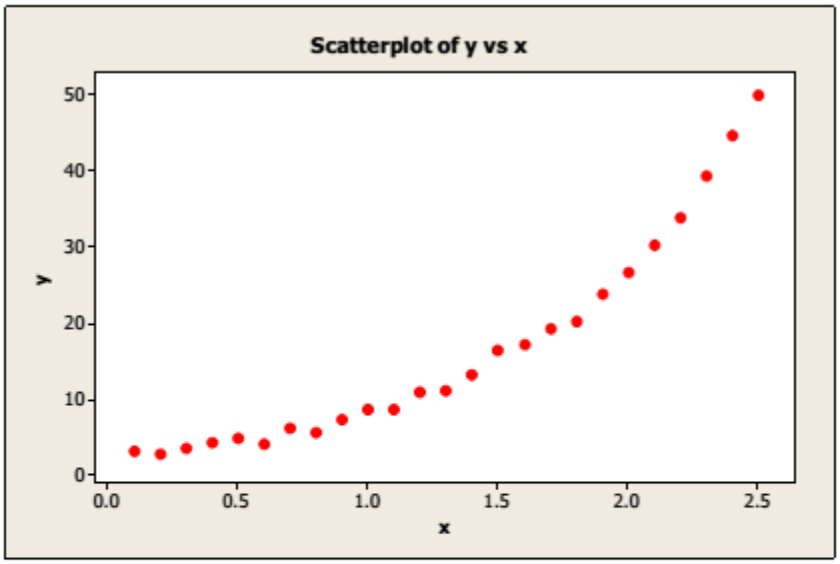
(c) The scientist proceeds to fit each model to his data. He has 25 data points, shown in the scatter lot below (Fig.1)

(i) Compare the estimates of *a* and *b* given by the two models. (1mark)

(ii) Computer output has provided some residual plots, which are given in the figures 2(a) & (b) below . Based on all of the available information, which model would you prefer, and why? (3 marks)

(iii) What further information would you wish to have in order to decide which is the better model? (3 marks)

Fig. 1



The resulting equations are

$\log(y) = 0.918 + 1.19x$ from the first model

and

$y = 2.45e^{1.20x}$ from the second model.

Fig.2(a)

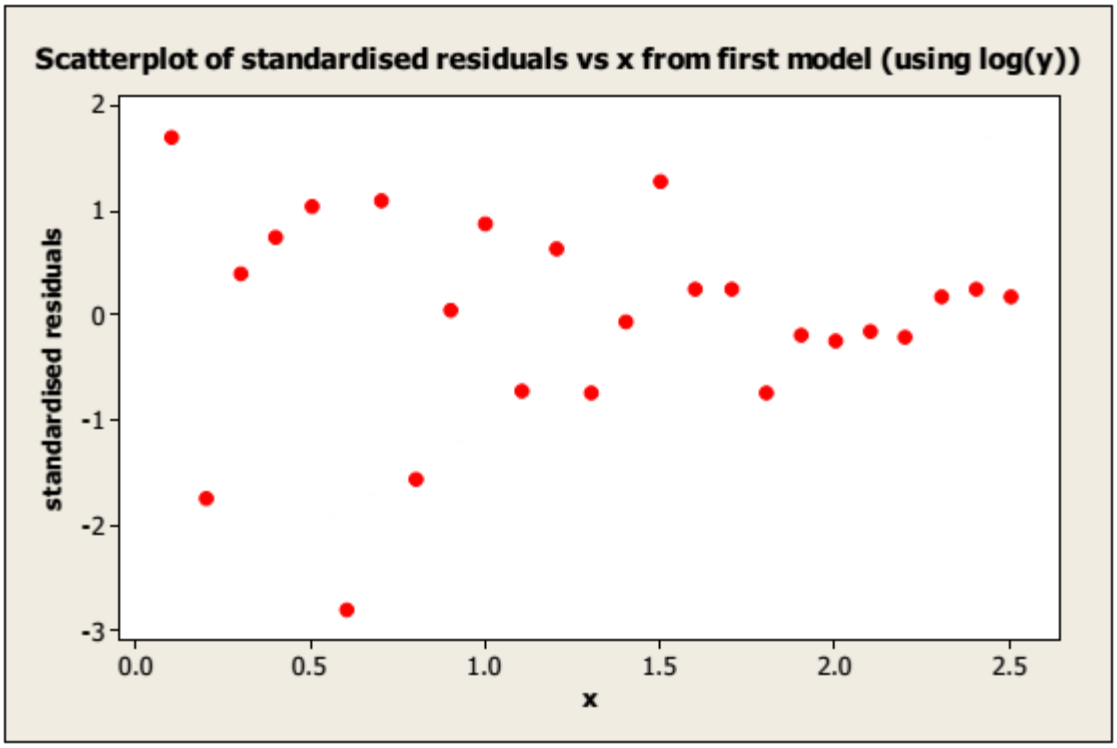
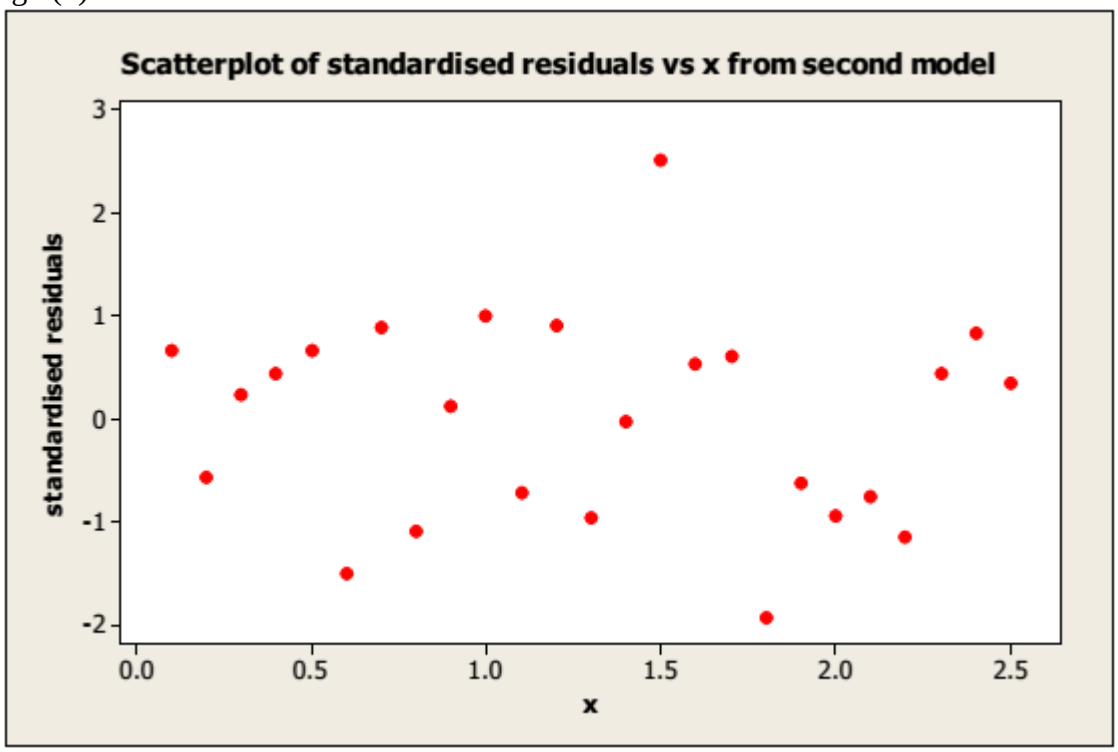

Scatterplot of standardised residuals vs x from first model (using log(y))
standardised residuals
2
1
0
-1
-2
-3
0.0
0.5
1.0
1.5
2.0
2.5
x

Fig.2(b)


Scatterplot of standardised residuals vs x from second model
standardised residuals
3
2
1
0
-1
-2
0.0
0.5
1.0
1.5
2.0
2.5
x

Question 5 (20 Marks)

(a) Briefly describe the advantages and disadvantages of the backward elimination method of model selection in multiple linear regression (4 marks)

(b) A dataset of 13 observations contains four predictor variables (X1, X2, X3, X4) and one response variable (Y), and the following statistics are available.

Variables in linear model	Model Sum of Squares
X1, X2, X3, X4	2667.90
X2, X3, X4	2641.95
X1, X3, X4	2664.93
X1, X2, X4	2667.79
X1, X2, X3	2667.65
X3, X4	2540.00
X2, X4	1846.88
X2, X3	2300.30
X1, X4	2641.00
X1, X3	1488.70
X1, X2	2657.90
X4	1831.90
X3	776.40
X2	1809.40
X1	1450.10
Total Sum of Squares	2715.76

Apply a backward elimination method to select the set of predictor variables that you consider "best" model the data. (8 marks)

(c) Explain how your method of model selection might be different if you knew what the predictor variables were. (4 marks)

(d) A journal gives the following advice to authors.
"Automated stepwise techniques often produce wildly unreliable results. This includes not only forward and backward automated selection but also 'best subset' approaches. Manuscripts that employ these techniques will not be considered unless the model is supported by a validation procedure."

(i) Do you agree with the first sentence? Justify your answer.
(ii) Suggest possible validation procedures that could be used for the model chosen in part (ii).
(4 marks)