TECHNICAL UNIVERSITY OF MOMBASA

FACULTY OF APPLIED AND HEALTH SCIENCES

DEPARTMENT OF MATHEMATICS & PHYSICS

UNIVERSITY EXAMINATION FOR:

BMCS

AMA 4411: REGRESSION MODELLING

END OF SEMESTER EXAMINATION

SERIES:APRIL2016

TIME:2HOURS

DATE:20May2016

Instructions to Candidates

You should have the following for this examination

-Answer Booklet, examination pass and student ID

This paper consists of Choose No questions. AttemptChoose instruction.

Do not write on the question paper.

Question ONE (30 MarkS)

(a) Write down the model for, and standard assumptions of, simple linear regression analysis. State a condition under which the method of least squares is equivalent to the method of maximum likelihood for estimating the regression coefficients. (4 marks)

(b) State the conditions under which you might consider using weighted least squares rather than ordinary least squares in simple linear regression. (4 marks)

(c) Explain what is meant by a *transformation to stabilise variance*, and give an example of where this might be useful in linear regression. (3marks)

(d) In many instances of linear modelling, a response variable y might be dependent on more than one predictor variable. Thus a set of variables xi (i = 1, 2, …, p) could be used to predict y through the general linear model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ + \cdots . \beta_p X_p + \varepsilon$ where the βi are model parameters and ε is an error term

(i) Write down the equivalent matrix formulation of the model, and state the form of the least squares estimators for the parameters in the model. (3 marks)

(ii) These least squares estimators have some very useful properties. State the properties they possess irrespective of the distribution of the errors. State the extra properties they possess if the errors are independent and normally distributed. (3 marks)

(e) Explain why highly dependent predictor variables can cause problems in fitting MLR models. What methods can be used to try to overcome such problems? (3 marks)

(f) Explain why an adjusted R^2 value is often preferred to R^2 when comparing models (2 marks)

(g) Explain what is meant by influential observations and why they can be a problem. (3 marks)

(h) Briefly discuss the relative merits of forward selection and backward elimination as applied to model selection in multiple linear regression. (5 marks)

Question TWO (20 MarkS)

The Devon Motor Racing Grand Prix takes place every five years. Winning average lap speeds (in kilometres per hour) in the last nine events are shown in the table below.

Year x	1965	1970	1975	1980	1985	1990	1995	2000	2005
Speed y	109	114	116	117	114	127	131	138	141

You are given that

$\bar{x} = 1985$, $\sum(x - \bar{x})^2 = 1500$, $\sum y = 1107$, $\sum y^2 = 137233$, $\sum(x - \bar{x})\, y = 1200$.

(a) (i) Plot these data and comment on their suitability for simple linear regression analysis. (4 marks)

(ii) Fit a simple linear regression model and state its equation. Also compute the total sum of squares and regression sum of squares for this regression, and deduce the error mean square. (6 marks)

(b) It is later noted that driving conditions in 1985 were affected by a freak thunderstorm which caused partial flooding of the track. The 1985 values were therefore omitted and the regression was recalculated. Results are shown in the computer output below. Compare this analysis with your own results and say with reasons which you regard as the more satisfactory. (3 marks)

```
The regression equation is    y = - 1464 + 0.800 x

Predictor       Coef   SE Coef        T       P
Constant    -1463.87     95.60   -15.31   0.000
x            0.80000   0.04816    16.61   0.000

S = 1.86525   R-Sq = 97.9%   R-Sq(adj) = 97.5%

Analysis of Variance
Source           DF       SS       MS        F       P
Regression        1   960.00   960.00   275.93   0.000
Residual Error    6    20.87     3.48
Total             7   980.87
```

(c) Use the analysis of part (b) to obtain point estimates of

(i) the expected winning speed in 1985, (2 marks)

(ii) the expected winning speed in 2010, (1 marks)

(iii) the time by which a winning speed of 160 kph might be expected. (2 marks)

(iv) Mention any reservations you might have about your answers. (2 marks)

Question THREE (20 MarkS)

The accompanying edited computer outputs show analyses of three different regression models for the progress in sales in hundreds (y) of a newly developed electronic component over time in months (x) since the launch of this product. These regression models may be written as shown below.

Model 1: $E(Y \mid x) = \alpha + \beta x$

Model 2: $E(Y \mid x) = \alpha + \beta x + \gamma x^2$

Model 3: $E(\log 10\ Y \mid x) = \alpha + \beta x$

Use the output to answer the following questions.

(i) In the output for Model 1, the p-values for the partial t tests for the slope and intercept parameters are missing. Making any necessary assumptions, use the available information to test these parameters for statistical significance at the 5% level. (3 marks)

(ii) In the light of Plot 1, comment on the adequacy of Model 1 for the data. (2 marks)

(iii) Test the significance of the coefficient of x in Model 2 and compare the outcome with the result of your test for the coefficient of x in Model 1. Interpret the statement "R-Sq = 98.1%" in the output for Model 2. (3 marks)

(iv) With reference to Plot 2, what standard assumption about the distribution of the error term may be called into question in Model 2? (2 marks)

(v) Critically compare Models 1 and 3 with regard to their success in fitting the data. Why does the total sum of squares for Model 3 differ from the total sum of squares for Models 1 and 2? (5 marks)

(vi) Use each of these models to give a point estimate of sales 10 months after launching the product. State with reasons which of the three estimates you think is the best. (5 marks)

Model 1(edited output)

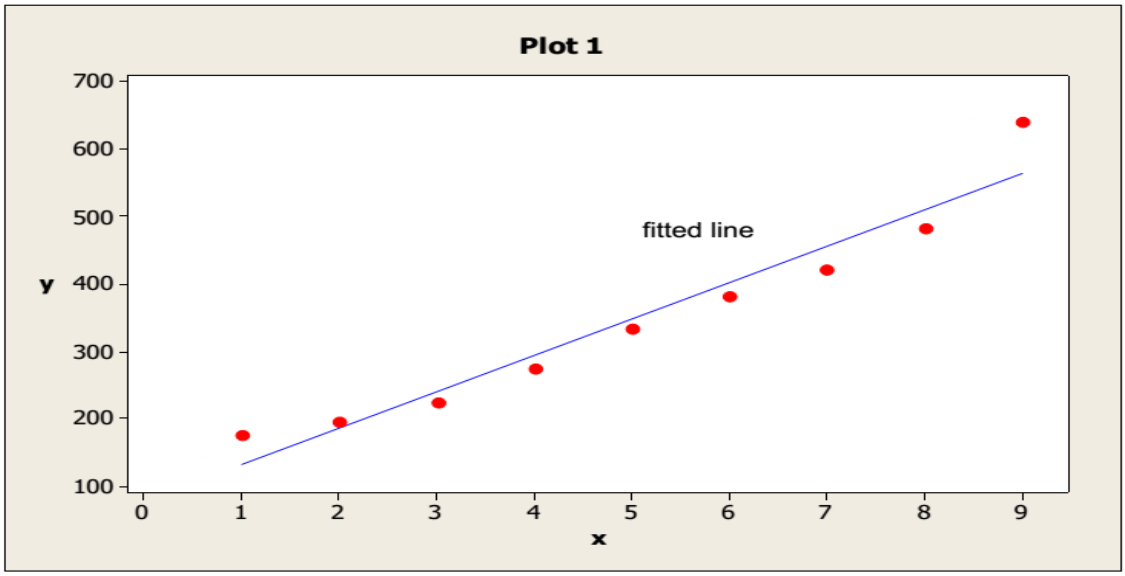
```
The regression equation is    y = 78.3 + 54.0 x

Predictor      Coef    SE Coef
Constant      78.33      29.01
x            54.000      5.155


S = 39.9285   R-Sq = 94.0%   R-Sq(adj) = 93.1%


Analysis of Variance

Source            DF       SS        MS         F        P
Regression         1   174960    174960    109.74    0.000
Residual Error     7    11160      1594
Total              8   186120
```



Model 2(edited output)

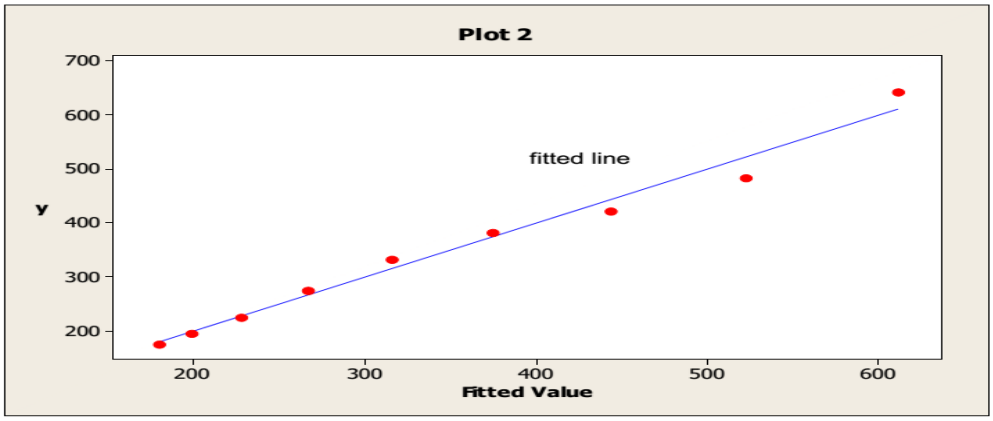
```
The regression equation is     y = 170 + 4.0 x + 5.00 x-sq

Predictor      Coef    SE Coef       T        P
Constant     170.00      30.56    5.56    0.001
x              4.00      14.03    0.29    0.785
x-sq          5.000      1.368    3.65    0.011

S = 24.0139    R-Sq = 98.1%    R-Sq(adj) = 97.5%

Analysis of Variance

Source            DF       SS       MS        F       P
Regression         2   182660    91330   158.38   0.000
Residual Error     6     3460      577
Total              8   186120
```



Model 3(edited output)

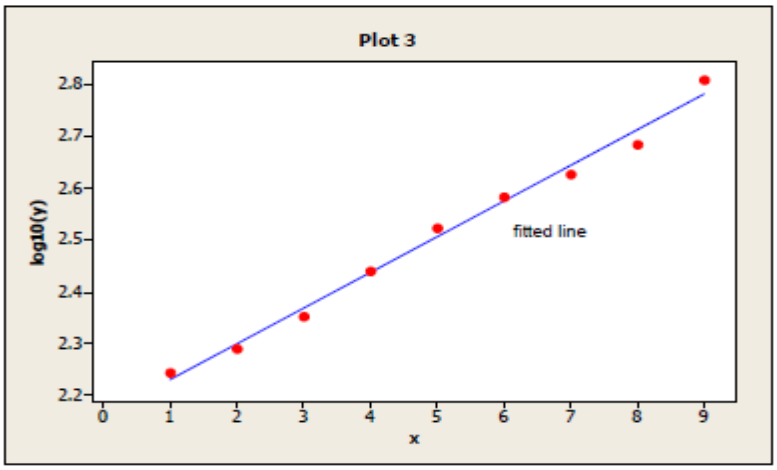
```
The regression equation is    log10(y) = 2.16 + 0.0689 x

Predictor        Coef     SE Coef          T        P
Constant      2.16086     0.01422     151.93    0.000
x             0.068910    0.002527     27.26    0.000


S = 0.0195777    R-Sq = 99.1%    R-Sq(adj) = 98.9%


Analysis of Variance

Source            DF        SS         MS          F       P
Regression         1    0.28492    0.28492     743.35   0.000
Residual Error     7    0.00268    0.00038
Total              8    0.28760
```



Question FOUR(20 Marks)

(a) Two explanatory variables are used to predict a dependent variable *Y*. Write down a multiple linear regression model which can be used as a basis for the analysis, and explain the meanings and properties of the terms in the model. (4 marks)

(b) The data in the following table show the values of price *Y* (£) for individually patterned Persian carpets of length *x*1 (cm) and width *x*2 (cm).

y	14	20	37	36	31	42	54	64	38	66	64	77	79	93	119	135
x_1	120	120	120	120	150	150	150	150	180	180	180	180	240	240	240	240
x_2	60	80	100	120	75	100	125	150	90	120	150	180	120	160	200	240

(i) Plot scatter diagrams of price against each of length and width. What do these graphs show? (5 marks)

(ii) A multiple regression model of price on length and width was fitted to the data given in the table. Edited computer output of the results is as follows. (11 marks)

Predictor	Coef	SE Coef	T	P
Constant	-52.671	5.34500	-9.85	0.000
Length	0.32356	0.04250	7.61	0.000
Width	0.44383	0.04012	11.06	0.000

S = 5.32611 R-Sq = 97.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	17045.2	8522.6	300.43	0.000
Residual Error	13	368.8	28.4		
Total	15	17413.9			

Interpret these results fully, in terms that a non-statistician would understand. Write down the fitted regression equation of *Y* on *x*1 and *x*2, and use it to predict the price of a similar carpet of length 200 cm and width 150 cm. To what extent would you rely on the model to predict the prices of carpets of dimensions outside the sizes observed in the above table (for example, much smaller carpets)? (11 marks)

Question FIVE(20 Marks)

(a) The setting for the weight sequence is *{Whi* (2)(*x*)*}*. Consider linear interpolation between two successive observations (*Xi*−1, *Yi*−1) and (*Xi*, *Yi*) with (*X*0, *Y*0) = (0, *Y*1),

$$g_i(u) = \frac{Y_i - Y_{i-1}}{X_i - X_{i-1}}(u - X_{i-1}) + Y_{i-1}, \quad i = 1, \ldots, n.$$

The linear interpolant of the data can be written as

$$G_n(u) = \sum_{i=1}^{n} g_i(u)\ I(X_{i-1} \le u < X_i).$$

Clark (1980) suggested convoling this linear interpolant with a kernel function with bandwidth *h*,

$$\begin{aligned}
\hat{m}(x) &= \int K_h(x-u)G_n(u)du \\
&= \sum_{i=1}^{n} \int_{X_{i-1}}^{X_i} K_h(x-u)g_i(u)du \\
&= \sum_{i=1}^{n} \int_{X_{i-1}}^{X_i} K_h(x-u)du Y_{i-1} \\
&\quad + \sum_{i=1}^{n} \int_{X_{i-1}}^{X_i} K_h(x-u)(u-X_i)du \frac{Y_i - Y_{i-1}}{X_i - X_{i-1}}.
\end{aligned}$$

Show that if the *x*-variables are equispaced on [0, 1], that is, *Xi* = *i/n* then the last term converges in probability to zero. (10 marks)

(b) Discuss the behavior of the kernel estimator when a single observation moves to a very large value, that is, study the case (*Xi*, *Yi*) → (*Xi*, *Yi* ±*c*) with *c* → ∞ for a fixed *i*. How does the curve change under such a distortion? What will happen for a distortion in *X*-direction (*Xi*, *Yi*) → (*Xi* ± *c*, *Yi*)? (10 marks)