

**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)**

1. 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)
2. State the conditions under which you might consider using weighted least squares rather than ordinary least squares in simple linear regression. (4 marks)
3. Explain what is meant by a *transformation to stabilise variance*, and give an example of where this might be useful in linear regression. (3marks)
4. 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 $y=β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2 }+….β\_{p}X\_{p}+ε$ where the βi are model parameters and ε is an error term
5. 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)
6. 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)
7. 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)
8. Explain why an adjusted R2 value is often preferred to R2 when comparing models (2 marks)
9. Explain what is meant by influential observations and why they can be a problem. (3 marks)
10. 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.



You are given that
$\overbar{x}$= 1985, ∑( *x* − $\overbar{x}$)2 = 1500, ∑*y* = 1107, ∑*y*2 = 137233, ∑( *x* − $\overbar{x}$) *y* = 1200.

1. (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)

1. 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)



1. Use the analysis of part (b) to obtain point estimates of
2. the expected winning speed in 1985, (2 marks)
3. the expected winning speed in 2010, (1 marks)
4. the time by which a winning speed of 160 kph might be expected. (2 marks)
5. Mention any reservations you might have about your answers. (2 marks)

**Question THREE (20 MarkS)**

The accompanying edited computer outputsshow 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 | x*) = α + β *x*

Model 2: *E*(*Y | x*) = α + β *x* + γ *x2*

 Model 3: *E*(log10 *Y* | *x*) = α + β *x*

Use the output to answer the following questions.

1. 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)
2. In the light of Plot 1, comment on the adequacy of Model 1 for the data. (2 marks)
3. 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)
4. 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)
5. 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)
6. 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)**





**Model 2(edited output)**



**Model 3(edited output)**

****

****

**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).



1. Plot scatter diagrams of price against each of length and width. What do these graphs show? (5 marks)
2. 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)



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)**

1. 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),



The linear interpolant of the data can be written as

 

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



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)

1. 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)