



Kasarani Campus  
Off Thika Road  
P. O. Box 49274, 00101  
NAIROBI  
Westlands Campus  
Pamstech House  
Woodvale Grove  
Tel. 4442212  
Fax: 4444175

**KIRIRI WOMENS' UNIVERSITY OF SCIENCE AND TECHNOLOGY**  
**UNIVERSITY EXAMINATIONS, 2022/2023 ACADEMIC YEAR**  
**FOURTH YEAR, FIRST SEMESTER END OF SEMESTER EXAMINATIONS**  
**FOR THE DEGREE OF BACHELOR OF SCIENCE (MATHEMATICS)**

**KMA 321 - REGRESSION MODELLING I**

Date: 14<sup>th</sup> April 2022

Time: 11.30am - 1.30pm

**INSTRUCTIONS TO CANDIDATES**

**ANSWER QUESTION ONE (COMPULSORY) AND ANY OTHER TWO QUESTIONS**

**QUESTION ONE COMPULSORY (30 MARKS)**

- a) Suppose we have the following bivariate dataset:

(1, 3.1) (1.7, 3.9) (2.1, 3.8) (2.5, 4.7) (2.7, 4.5)

For these data  $\sum x = 10$ ,  $\sum y = 20$ ,  $\sum x^2 = 21.84$ ,  $\sum xy = 41.61$

- i) Determine the least squares estimates  $\hat{\alpha}$  and  $\hat{\beta}$  of the parameters of the regression line  $y = \alpha + \beta x$ .  
[4 Marks]
- ii) Draw in one figure the scatterplot of the data and the estimated regression line  $y = \hat{\alpha} + \hat{\beta} x$ .  
[2 Marks]
- b) Define what heteroscedasticity in regression analysis is and briefly explain how the problem can be solved.  
[3 Marks]
- c) Prove that in multiple regression, the estimator  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$  has a variance  $Var(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ .  
[4 Marks]
- d) The quality of primary schools in eight regions in the UK is measured by an index ranging from 1 (very poor) to 10 (excellent). In addition the value of a house price index for these eight regions is observed. The results are given in the following table:

| Region $i$                 | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | Sum  |
|----------------------------|-----|-----|-----|-----|-----|-----|-----|-----|------|
| School quality index $x_i$ | 7   | 8   | 5   | 8   | 4   | 9   | 6   | 9   | 56   |
| House price index $y_i$    | 195 | 195 | 170 | 190 | 150 | 190 | 200 | 210 | 1500 |

The last column contains the sum of all eight columns. From these values we obtain the following results:

$$\sum x^2 = 416 \quad \sum y^2 = 283,750 \quad \sum x_i y_i = 10,695$$

- i) Calculate the correlation coefficient between the index of school quality and the house price index. [4 Marks]
- ii) Fit a linear regression model to the data, by considering the school quality index as the explanatory variable. You should write down the model and estimate all parameters. [4 Marks]
- iii) Calculate the coefficient of determination  $R^2$  for the regression model obtained in part (ii). [3 Marks]
- iv) Provide a brief interpretation of the slope of the regression model obtained in part (ii). [2 Marks]
- e) Suppose the relationship between the independent variable *height* ( $x$ ) and dependent variable *weight* ( $y$ ) is described by a simple linear regression model with true regression line  $y = 7.5 + 0.5x$  and  $\sigma = 3$
- i) What is the interpretation of  $\beta_1 = 0.5$ ? [2 Marks]
- ii) If  $x = 20$  what is the expected value of  $Y$ ? [2 Marks]

## **QUESTION TWO (20 MARKS)**

- a) A linear regression model is given by the equation  $Y = X\beta + \epsilon$ , where  $Y$  is a  $(n \times 1)$  vector,  $X$  is a  $(n \times p)$  matrix,  $\beta$  is a  $(p \times 1)$  vector of constants and  $\epsilon$  is a  $(n \times 1)$  vector of random variables such that  $E(\epsilon) = 0$ ,  $Var(\epsilon) = \sigma^2 I_n$ ,  $\epsilon \sim N(0; \sigma^2 I_n)$ . Derive the least squares normal equations. [8 Marks]
- b) Heller Company manufactures lawn mowers and related lawn equipment. The managers believe the quantity of lawn mowers sold depends on the price of the mower and the price of a competitor's mower. We have the following data:

| Competitor's Price<br>$x_{i1}$ | Heller's Price<br>$x_{i2}$ | Quantity sold<br>$y_i$ |
|--------------------------------|----------------------------|------------------------|
| 120                            | 100                        | 102                    |
| 140                            | 110                        | 100                    |
| 190                            | 90                         | 120                    |
| 130                            | 150                        | 77                     |
| 155                            | 210                        | 46                     |
| 175                            | 150                        | 93                     |
| 125                            | 250                        | 26                     |
| 145                            | 270                        | 69                     |
| 180                            | 300                        | 65                     |
| 150                            | 250                        | 85                     |

- (i) Fit a multiple linear regression model of these data [8 Marks]
- (ii) What is the interpretation of the model fit in (i) [2 Marks]
- (iii) Use the model to predict quantity sold when  $x_{i1} = \$170$  and  $x_{i2} = \$160$  [2 Marks]

### QUESTION THREE (20 MARKS)

- a) An actuary is fitting the following simple linear regression model  $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ , with  $E(\epsilon_i) = 0$ ,  $V(\epsilon_i) = \sigma^2$ ,  $i = 1, 2, \dots, n$  and the errors  $\epsilon_i$ ,  $\epsilon_j$  for  $i \neq j$  are uncorrelated.
- (i) Show that the least squares estimators of  $\beta_0$  and  $\beta_1$  are unbiased. [9 marks]
- (ii) Find the variance of  $\hat{\beta}_1$ . What implication does this have on the choice of design points? [6 marks]
- b) Show that the least squares estimation of  $\beta$  for the multiple linear regression model  $\mathbf{y} = \mathbf{X}\beta + \epsilon$ , is  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$  assuming  $(\mathbf{X}'\mathbf{X})$  is a non-singular matrix. [5 marks]

### QUESTION FOUR (20 MARKS)

- a) It is thought that a plumber charges £22 per hour plus an administrative charge of £15 per call-out. A sample of eight invoices was obtained corresponding to jobs with durations of 1 hour, 2 hours... 8 hours. For each invoice the total cost of the job was noted with the following results:

|                     |    |    |    |    |     |     |     |     |
|---------------------|----|----|----|----|-----|-----|-----|-----|
| Time $x$<br>(hours) | 1  | 2  | 3  | 4  | 5   | 6   | 7   | 8   |
| Cost $y$ (£)        | 40 | 50 | 81 | 89 | 122 | 128 | 151 | 179 |

$$\sum(x - \bar{x})^2 = 42$$

$$\sum(y - \bar{y})^2 = 16492$$

$$\sum(x - \bar{x})(y - \bar{y}) = 826$$

The following model is used to represent the data

$$y_i = a + bx_i + e_i$$

Where  $y_i$  ( $i = 1, 2, \dots, 8$ ) are the costs,  $x_i$  ( $i = 1, 2, \dots, 8$ ) are the fixed times and  $e_i$  ( $i = 1, 2, \dots, 8$ ) are independent errors with a  $N(0, \sigma^2)$  distribution.

- i) Derive the formulae for the least square estimators of  $a$  and  $b$ . [6 Marks]
- ii) How would your answer to part (i) have differed if you had been asked to find the maximum likelihood estimators. [2 Marks]
- iii) Calculate the regression coefficients  $\hat{a}$  and  $\hat{b}$ . [4 Marks]
- iv) Carry out a test to establish whether or not the slope in the model agrees with the suggested £22 per hour. [4 Marks]
- v) Calculate a 90% confidence interval for the:
- a. Average cost of a job lasting 4 hours [2 Marks]
- b. Cost of an individual job lasting 6 hours [2 Marks]

**QUESTION FIVE(20 MARKS)**

- a) The data given in the following table are the numbers of deaths from AIDS in Australia for 12 consecutive quarters starting from the second quarter of 1983.

|                                |   |   |   |   |   |   |   |    |    |    |    |    |
|--------------------------------|---|---|---|---|---|---|---|----|----|----|----|----|
| Quarter<br>(i):                | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8  | 9  | 10 | 11 | 12 |
| Number of<br>deaths ( $n_i$ ): | 1 | 2 | 3 | 1 | 4 | 9 | 1 | 23 | 31 | 20 | 25 | 37 |

- i) Draw a scatterplot of the data. [2 Marks]
- ii) Comment on the nature of the relationship between the number of deaths and the quarter in this early phase of the epidemic. [2 Marks]
- b) A statistician has suggested that a model of the form

$$E[N_i = \gamma i^2]$$

might be appropriate for these data, where  $\gamma$  is a parameter to be estimated from the above data. She has proposed two methods for estimating  $\gamma$ , and these are given in parts (a) and (b) below.

- i) Show that the least squares estimate of  $\gamma$ , obtained by minimizing

$$q = \sum_{i=1}^{12} (n_i - \gamma i^2)^2 \text{ , is given by}$$

$$\bar{\gamma} = \frac{\sum_{i=1}^{12} i^2 n_i}{\sum_{i=1}^{12} i^4} \quad [4 \text{ Marks}]$$

- ii) Show that an alternative (weighted) least squares estimate of  $\gamma$ , obtained by minimizing

$$q^* = \sum_{i=1}^{12} \frac{(n_i - \gamma i^2)^2}{i^2} \text{ , is given by}$$

$$\bar{\gamma}^* = \frac{\sum_{i=1}^{12} n_i}{\sum_{i=1}^{12} i^2} \quad [2 \text{ Marks}]$$

- iii) Noting that  $\sum_{i=1}^{12} i^4 = 60710$  and  $\sum_{i=1}^{12} i^2 = 650$  , calculate  $\bar{\gamma}$  and  $\bar{\gamma}^*$  for the above data. [4 marks]
- c) To assess whether the single parameter model which was used in part (b) is appropriate for the data, a two parameter model is now considered. The model is of the form

$$E[N_i = \gamma i^\theta] \quad \text{for } i = 1, \dots, 12.$$

- i) To estimate the parameters  $\gamma$  and  $\theta$ , a simple linear regression model

$$E[y_i] = \alpha + \beta x_i$$

is used, where  $x_i = \log(i)$  and  $y_i = \log(N_i)$  for  $i = 1, \dots, 12$ . Relate the parameters  $\gamma$  and  $\theta$  to the regression parameters  $\alpha$  and  $\beta$ . [3 Marks]

- ii) The least squares estimates of  $\alpha$  and  $\beta$  are 0.6112 and 1.6008 with standard errors 0.4586 and 0.2525 respectively (you are not asked to verify these results).

Using the value for the estimate of  $\beta$ , conduct a formal statistical test to assess whether the form of the model suggested in (b) is adequate. [3 Marks]