

UNIVERSITY EXAMINATIONS

SECOND SEMESTER 2023/2024 ACADEMIC YEAR

**SECOND YEAR EXAMINATION FOR THE DEGREE OF
BACHELOR OF SCIENCE (STATISTICS)**

STAT 221: REGRESSION AND CORRELATION ANALYSIS

STREAM: R

TIME: 2 HRS

DAY: MONDAY [11.30A.M – 1.30P.M] DATE: 15/04/2024

THIS QUESTION PAPER CONSISTS OF FIVE (5) PAGES

PLEASE DO NOT OPEN UNTIL THE INVIGILATOR SAYS SO

INSTRUCTIONS: ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS**QUESTION ONE (30 MARKS)**

- a) State any four assumptions of classical linear regression model **(4 Marks)**
- b) Briefly describe two ways in which regression analysis differ from analysis of variance
- c) Suppose $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (1)
is a simple linear model with $\varepsilon_i \sim N(0, \sigma^2)$ and let the errors ε_i associated with different observations ($i = 1, 2, \dots, N$) be independent.
- Explain why the model (1) is said to be simple **(2 Marks)**
 - Explain why the model (1) is linear **(2 Marks)**
 - What is the significance of ε_i in the model (1) **(2 Marks)**
 - Using the method of ordinary least squares (OLS), obtain estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the unknown parameter values β_0 and β_1 . **(10 Marks)**
- d) The following data give the annual income (in million shillings) and amounts (in million shillings) of life insurance policies for eight persons.

Annual income	42	58	27	36	70	24	53	37
Life insurance	150	175	25	75	250	50	250	100

Determine the least squares regression function for these data using matrix operation **(8 Marks)**

QUESTION TWO (20 MARKS)

The following data relate the demand for housing (Y) based on current interest rate, X_1 and the rate of unemployment, X_2 :

Units sold	Y	65	59	80	90	100	105
Interest rate (%)	X_1	9.0	9.3	8.9	9.1	9.0	8.7
Unemployment rate (%)	X_2	10.0	8.0	8.2	7.7	7.1	7.2

- Write down the multiple linear regression model for the data in matrix form. **(2 Marks)**
- Find $X^T X$, $(X^T X)^{-1}$ and $X^T Y$ **(6 Marks)**
- Estimate β . Hence report the fitted multiple regression model. Interpret your model parameter estimates. **(8 Marks)**
- Obtain Cov (β). **(4 Marks)**

QUESTION THREE (20 MARKS)

Consider the following simple linear regression model;

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (1) with $\varepsilon_i \sim N(0, \sigma^2)$ and let the errors ε_i associated with different observations ($i = 1, 2, \dots, N$) be independent. Given the estimates for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}, = \frac{1}{n} (\sum Y_i - \widehat{\beta}_1 \sum X_i) \quad \text{and} \quad \widehat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}, = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X^2 - (\sum X_i)^2}$$

Find,

- a) $E(\widehat{\beta}_0)$ (4 Marks)
- b) $E(\widehat{\beta}_1)$ (8 Marks)
- c) $\text{Var}(\widehat{\beta}_0)$ (4 Marks)
- d) $\text{Var}(\widehat{\beta}_1)$ (4 Marks)

QUESTION FOUR (20 MARKS)

- a) Define the following terms: (4 Marks)
 - i). a Type I error
 - ii). a Type II error
 - iii). Simple hypothesis
 - iv). Composite hypothesis

- b) calculate the correlation coefficient from the following data and interpret (8 Marks)

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	13

- c) Explain the difference between positive and negative correlation (2 Marks)
- d) A researcher is interested in determining whether there is a relationship between years of education (X) and income(Y). Consider the following annotated regression output for the data:



SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.860811139
R Square	0.740995817
Adjusted R Square	0.715095399
Standard Error	7.816452413
Observations	12

ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	1747.947383	1747.947383	28.60941509	0.000324168	
Residual	10	610.9692833	61.09692833			
Total	11	2358.916667				

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-11.02047782	8.909954058	-1.236872575	0.244393811	-30.87309606	8.832140427
X Variable 1	3.197952218	0.597884757	5.348776972	0.000324168	1.865781732	4.530122704

- i) Write down the regression function for this data. (2 Marks)
- ii) Interpret the regression model in question (i) above (2 Marks)
- iii) Interpret the correlation between education and income (2 Marks)

QUESTION FIVE (20 MARKS)

- a) The following data relate to the prices (Y) of five randomly chosen houses in a certain neighborhood, the corresponding ages of the houses (x_1), and square footage (x_2).

Price Y in thousands of dollars	Age X_1 in years	Square footage X_2 in thousands of square feet
100	1	1
80	5	1
104	5	2
94	10	2
130	20	3

Fit a multiple linear regression model $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$ to the above data using matrix method to estimate the regression parameters. (12 Marks)

- b) Explain the difference between simple and multiple regression (4 Marks)
- c) Using relevant illustrations, explain the difference between the following types of correlation, positive, negative and no correlation (4 Marks)

