JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY

SCHOOL OF BIOLOGICAL, PHYSICAL, MATHEMATICS AND ACTUARIAL SCIENCE

UNIVERSITY EXAMINATION FOR DIPLOMA IN APPLIED STATISTICS

2ND YEAR 1ST SEMESTER 2024/2025 ACADEMIC YEAR

MAIN CAMPUS

COURSE CODE: WAB 2216

COURSE TITLE: STATISTICAL MODELLING

EXAM VENUE: **STREAM:**

DATE: **EXAM SESSION:**

TIME: 2.00 HOURS

Instructions:

1. **Answer question one (compulsory) and any other three questions.**
2. **Candidates are advised not to write on the question paper.**
3. **Candidates must hand in their answer booklets to the invigilator while in the examination room.**

Section A

Question 1 [40 marks]

a) Define regression analysis and its primary purpose. (3 Marks)
b) List three application areas where regression analysis is commonly used. (3 Marks)
c) State and briefly explain the assumption of linearity in regression models. (3 Marks)
d) Define homoscedasticity and explain why it is important. (3 Marks)
e) List and explain two property of Least Squares Estimates. (4 Marks)
f) If the regression equation is y=2+3x, what is the predicted value of y when x=4x?

(4 Marks)

g) Given the predicted values y^= [2, 4, 6] and the actual values y= [3, 3, 8] calculate the residuals. (4 Marks)
h) Given the data points (1, 2), (2, 3), (3, 5), calculate the slope of the simple linear regression line. Show your working. (4 Marks)
i) A study finds a linear relationship between hours studied and exam scores. If a student studies for 5 hours, predict their score using a linear model y=10+8xy =. What is the predicted score? (3 Marks)
j) What role does ANOVA play in regression analysis? (2 Marks)
k) Describe how you would interpret an F-statistic value of 5.0 in the context of a regression analysis. (3 Marks0
l) In a regression analysis, if the total sum of squares (TSS) is 120 and the estimated sum of squares (ESS) is 90, what is the residual sum of squares (RSS)? (3 Marks)
m) What can a large residual indicate about a particular observation in your dataset?

(2 Marks)

n) If you have a dataset of 50 observations, what statistical test would you use to check for normality in the residuals, and how would you interpret the results? (3 Marks)

Question 2 [20 marks]

A real estate agency wants to analyse the relationship between the number of advertisements placed (in units) and the number of house sales (in units) in a particular region. They suspect that an increase in advertising may be positively associated with an increase in house sales.

The following table shows data for the last eight months on the number of advertisements placed and the number of houses sold:

Month	Advertisements (X)	House Sales (Y)

Month	Advertisements (X)	House Sales (Y)
1	10	15
2	15	18
3	12	14
4	20	22
5	25	24
6	30	30
7	35	28
8	40	36

Using this data, answer the following questions:

i. Calculate the mean of the number of advertisements placed (X) and the mean of the number of houses sold (Y). (5 Marks)

ii. Using the formula for Pearson's correlation coefficient, calculate the correlation coefficient r between advertisements placed and house sales. Show all steps in your calculation. (5 Marks)

iii. Interpret the value of the correlation coefficient in the context of this study. What does it imply about the relationship between advertising and house sales? (5 Marks)

iv. Based on the correlation coefficient calculated, discuss one practical recommendation the agency could consider in relation to their advertising strategy. (5 Marks)

Question 3 [20 marks]

a. Given a linear model $Yi = \beta_o + \beta_1 X_i + \varepsilon_i$, derive a formula for both β_0 and β_1

(10 Marks)

b. An expert on crime rates has collected the following information on five counties in one province.

County	Crime Rate (Y)	Poverty Rate (X)
1	10	5
2	19	7
3	20	11
4	16	8
5	15	9

Crime Rate (Y) stands for number of crimes per 10,000 people. Poverty Rate (X) is percent of families below poverty line. Sample statistics based on these data:

i) Find the regression line and interpret the intercept and slope coefficients

(4 Marks)

ii) Compute the coefficient of determination, R-squared. What does it mean? Make an inference about the crime rate in a county with a poverty rate of 10. [Use $\alpha = 0.05$ for the interval.] (3 Marks)

iii) A governor claims that based on the estimated slope of the regression, a one percentage point increase in the poverty rate is associated with less than two additional crimes per 10,000 people. Test this claim at a 5% significance level. (3 Marks)

Question 4 [20 marks]

i) What do you mean by one- way ANOVA? (2 Marks)

ii) What is the P-Value in ANOVA (2 Marks)

iii) What is ANOVA Test (2 Marks)

iv) Three different kinds of food are tested on three groups of rats for 5 weeks. The objective is to check the difference in mean weight (in grams) of the rats per week. Apply one-way A NOVA using a 0.05 significance level to the following data: (12 Marks)

Food 1	Food II	Food III
8	4	11
12	5	8
19	4	7
8	6	13
6	9	7
11	7	9

Question 5 (20marks)

The sales of a company (in million dollars) for each year are shown in the table below.

X (years)	2020	2021	2022	2023	2024
Y (years)	15	19	25	35	47

i. Find the least square regression line $y = a x + b$. (5 Marks)

ii. Use the least squares regression line as a model to estimate the sales of the company in 2030. (2 Marks)

iii. Calculate the Estimated Sum of Squares , Residual Sum of Squares and Total Sum of Squares (10 Marks)

iv. Calculate the Coefficient of Determination (3 Marks)

Question 6 [20 marks]

As part of an investigation into health service funding a working party was concerned with the issue of whether mortality rates could be used to predict sickness rates. Data on standardised mortality rates and standardised sickness rates were collected for a sample of 10 regions and are shown in the table below:

Region	Mortality rate m(per 1000)	Sickness rate s (per 1000)
1	125.2	206.8
2	119.3	213.8
3	125.3	197.2
4	111.7	200.6
5	117.3	189.1
6	100.7	183.6
7	108.8	181.2
8	102.0	168.2
9	104.7	165.2
10	121.1	228.5

Data summaries: $\sum m$ = 1136.1, $\sum m2$ = 129,853.03, $\sum s$ = 1934.2, $\sum s$ 2 = 377,700.62,

$\sum ms$ = 221,022.58

(i) Calculate the correlation coefficient between the mortality rates and the sickness rates (8 Marks)

(ii) Noting the issue under investigation, draw an appropriate scatter plot for these data and comment on the relationship between the two rates (3 Marks)

(iii) Determine the fitted linear regression of sickness rate on mortality rate and test whether the underlying slope coefficient can be considered to be as large as 2.0

(5 Marks)

(iv) For a region with mortality rate 115.0, estimate the expected sickness rate and calculate 95% confidence limits for this expected rate. (4 Marks)