JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY

SCHOOL OF MATHEMATICS AND ACTUARIAL SCIENCE

UNIVERSITY EXAMINATION FOR DEGREE OF BACHELOR OF SCIENCE ACTUARIAL

SPECIAL RESIT 2 2020/2021 ACADEMIC YEAR

MAIN REGULAR

COURSE CODE: SAC 104

COURSE TITLE: Linear Models And Forecasting

EXAM VENUE: **STREAM:**

DATE: EXAM SESSION:

TIME: 2.00 HOURS

Instructions:

1. **Answer question 1 (Compulsory) and ANY other 2 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.**

QUESTION ONE: COMPULSORY

i. $Define\ \ the\ following\ terms\ and\ state\ their\ relevance\ in\ \ modern\ statistical\ theory\ and$ practice

(a) Regression Analysis
(b) Correlation Analysis
(c) Coefficient of determination
(d) Scatter plots
(e) Linear Forecasting [10mks]

ii. State four assumptions of Ordinary Least Squares Method [4mks]

iii. Prove that $\hat{\beta}$ is unbiased estimator of β [4mks]

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

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

a. Find the least square regression line y = a x + b. [5mks]

b. Use the least squares regression line as a model to estimate the sales of the company in 2012. [3mks]

c. Calculate the Residual Sum of Squares [3mks]

d. Calculate the Coefficient of Determination [3mks]

QUESTION TWO

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 10,000)*	*Sickness rate s (per 1,000)*
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:

$\sum m = 1136.1$, $\sum m^2 = 129{,}853.03$, $\sum s = 1934.2$, $\sum s^2 = 377{,}700.62$, $\sum ms = 221{,}022.58$

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

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

(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]

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

[Total 20]

QUESTION THREE (20marks)

a) Given a linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, derive a formula for both β_0 and β_1 [9]

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:

$$\sum X = 40 \quad \sum X^2 = 340 \quad \sum Y = 80 \quad \sum Y^2 = 1342 \quad \sum XY = 666$$

i) Find the regression line and interpret the intercept and slope coefficients (5marks)

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 α = 0.05 for the interval.] (3marks)

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. (3marks)

QUESTION FOUR

The effectiveness of teaching Methodology X_1 *and* X_2 was being tested on a group of JOOUST Actuarial Science students and the following results were obtained

% Effectiveness, y	X_1	X_2
92.5	50.9	20.8
94.9	54.1	16.9
89.3	47.3	25.2
94.1	45.1	49.7
98.9	37.6	95.2

The Data summery is as below

$\sum X_1 = 235,$ $\sum X_1^2 = 11{,}202.68,$
$\sum X_2^2 = 12{,}886.42,$ $\sum YX_1 = 22{,}028.78,$

$\sum YX_2 = 19870.22$ $\sum X_2 = 207.8,$ $\sum X_2X_1 = 8985.96$

(i) Using the Multiple Linear least squares regression Model:

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

(a) Show that the least Squares estimate of α , β_1 *and* β_2 satisfy

$$\sum Y_i = n\alpha + \beta_1 \sum X_{i1} + \beta_2 \sum X_{i2}$$

$$\sum Y_i X_{i1} = \alpha \sum X_{i1} + \beta_1 \sum X_{i1}^2 + \beta_2 \sum X_{i2}X_{i1}$$

$$\sum Y_i X_{i1} = \alpha \sum X_{i2} + \beta_1 \sum X_{i1}X_{i2} + \beta_2 \sum X_{i2}^2$$

(b) Hence or otherwise use the Matrix Method to find their Algebraic Values and also use the data values to find the Numerical values

[**TOTAL 20]**

QUESTION FIVE

(i)A sample of 20 claim amounts (£) on a group of household policies gave the following data summaries:

$\sum X$= 3,256 and $\sum X^2$ = 866,600.

(a) Calculate the sample mean and standard deviation for these claim amounts. **[3]**

(b) Comment on the skewness of the distribution of these claim amounts, giving reasons for your answer. **[3]**

(ii) Consider the following two random samples of ten observations which come from the distributions of random variables which assume non-negative integer values only.

Sample 1: 7 4 6 11 5 9 8 3 5 5

Sample mean = 6.3, sample variance = 6.01

Sample 2: 8 3 5 11 2 4 6 12 3 9

Sample mean = 6.3, sample variance = 12.46

One sample comes from a Poisson distribution, the other does not.
State, with brief reasons, which sample you think is likely to be which. **[4]**

(iii) The sample correlation coefficient for the set of data consisting of the three pairs of values $(-1, -2)$, $(0, 0)$, $(1, 1)$ is 0.982. After the x and y values have been transformed by particular linear functions, the data become:
$(2, 2)$, $(6, -4)$, $(10, -7)$.
Calculate the correlation coefficient for the transformed data. **[4]**

(i) An investigation concerning the improvement in the average performance of female track athletes relative to male track athletes was conducted using data from various international athletics meetings over a period of 16 years in the 1950s and 1960s. For each year and each selected track distance the observation y was recorded as the average of the ratios of the twenty best male times to the corresponding twenty best female times.

The data for the 100 metres event are given below together with some summaries.

Year t:	1	2	3	4	5	6	7	8
Ratio y:	0.882	0.879	0.876	0.888	0.890	0.882	0.885	0.886
Year t:	9	10	11	12	13	14	15	16
Ratio y:	0.885	0.887	0.882	0.893	0.878	0.889	0.888	0.890

$\Sigma t = 136$, $\Sigma t^2 = 1496$, $\Sigma y = 14.160$, $\Sigma y^2 = 12.531946$, $\Sigma ty = 120.518$

(a) Verify that the equation of the least squares fitted regression line of ratio on year is given by:
$y = 0.88105 + 0.000465t$. [4]

(b) Calculate the standard error of the estimated slope coefficient in part (b). [6]