MACHAKOS UNIVERSITY

University Examinations 2022/2023 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR SUPPLEMENTARYSEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (STATISTICS AND PROGRAMMING)

BACHELOR OF SCIENCE (MATHEMATICS)

SMA 364: APPLIED STATISTICAL METHODS

DATE: 7/3/2023 **TIME: 2:00 – 4:00 PM**

INSTRUCTION:

Answer Question One and Any Other Two Questions

QUESTION ONE (30 MARKS)

a) Highlight the process of importing a data from excel sheet to R platform (4 marks)

b) Discuss any two test statistics highlighting when appropriate to apply (4 marks)

c) Distinguish the following terms as used in applied statistics

i. Confounding variable and Noise variable

ii. Univariate and Bivariate data dimensions (4 marks)

d) Highlight two assumptions underlying the regression diagnostic (4 marks)

e) Distinguish the following terms as they apply in data analysis

i. A variable and a case

ii. Parametric and Non-parametric test statistics

iii. Multicollinearity and outliers (6 marks)

f) A random sample of 400 persons was selected from each of three age groups and each person was asked to specify which of the three presidential candidates she/ he preferred. The results are shown in table 1:

Table 1

Age group	Presidential Candidate		
	A	B	C
Under 30	120	30	50
31 – 50	10	75	15
51 and above	10	30	60

Test the hypothesis that the populations were homogeneous with respect to the presidential candidate they preferred despite their age difference. (8 marks)

QUESTION TWO (20 MARKS)

a) Discuss briefly the differences between the following terms as they apply in statistics

i. Simple Linear and Multiple Linear Regression models (2 marks)

ii. Logistic regression and Multinomial regression (2 marks)

iii. One-way ANOVA and two-way ANOVA (2 marks)

b) The following are the results after conducting research on a micro finance company performance as affected by, wages, cost of capital and technology.

$\hat{Y} = -2.478 + 0.288x_1 + 0.758x_2 + 0.085x_3 + \mu$

Se (9.748) (0.110) (0.112) (0.168)

$R^2 = 0.928$ $n = 30$

Where Y is the microfinance performance as measured by return on assets

X_1 is wages

X_2 is cost of capital

X_3 is the technology

μ is the error term

i. State the hypotheses for the study (2 marks)

ii. Interpret the coefficient of determination. (2 marks)

iii. Interpret the results using regression coefficients (5 marks)

iv. Test the significance of the coefficients at $\alpha = 0.05$ level (5 marks)

QUESTION THREE (20 MARKS)

a) Highlight four ways of carrying out regression diagnostics (4 marks)

b) Explain the following terms as used in applied statistics

i. Fixed model effects

ii. Random model effects

iii. Mixed model effects (6 marks)

c) One type of lady's gel was placed at five different heights within the same week. Sales (in cartons of 12) at each level were recorded as summarized in table 1

Table 3:

Days	Height level Placed				
	2 feet	**3 feet**	**4 feet**	**5 feet**	**6 feet**
Monday	6	4	3	5	1
Tuesday	7	3	2	4	3
Wednesday	3	5	3	4	3
Thursday	8	7	4	5	5
Friday	7	8	8	4	8

i. Perform a one – way analysis of variance to test the hypothesis that the five different heights and days yielded the same average sales at α=0.05 (8 marks)

ii. Make statistical conclusion based on your findings in (i) (2 marks)

QUESTION FOUR (20 MARKS)

a) Discuss five ways of dealing with the problem of an outlier in a data set. (10 marks)

b) The data below shows the medical cost (y) in Kenya shillings ('000') per month for 10 randomly sampled patients over time (x) in years. If the two are assumed to relate in the form of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$.

Patients	A	B	C	D	E	F	G	H	I	J
Medical cost	49	37	33	11	10	29	44	52	69	71
Time in years	1	3	15	19	24	33	45	59	77	81

i. Write down all the steps for fitting in the non-linear regression connecting the medical cost and time in years in R (7 marks)

ii. Highlight three advantages of non-linear regression over the linear regression (3 marks)

QUESTION FIVE (20 MARKS)

a) Discuss three methods of variable selection for model building. (6 marks)

b) Discuss any three basic data types in R statistical software. (6 marks)

c) A contracting farming agency collected the data below according to the listed variables in order to predict the yield of biashara green grams in future:

X1	23	34	56	78	15	21	24
X2	34	56	78	79	90	65	14
X3	23	12	12	13	14	16	23
Y	4	5	12	23	4	5	8

Y= Yield per acre in kg '000'

X1= Amount of NPK fertilizer applied per plant in grams

X2= Amount of water applied per plant in litters for the season

X3 = Spacing in inches between two consecutives plants

The following is part of the output analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.625373
R Square	0.391092
Adjusted R Square	-0.21782
Standard Error	7.63801
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance*
Regression	3	112.411	37.47033	0.642284	0.637562
Residual	3	175.0176	58.33919		
Total	6	287.4286			

i. State the statistical model in this analysis. (2 marks)

ii. Interpret the finding in this analysis (3 marks)

iii. Explain the difference between R Square and R Adjusted (3 marks)