



MACHAKOS UNIVERSITY

University Examinations 2018/2019

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

THIRD YEAR SECOND SEMESTER EXAMINATION FOR
BACHELOR OF SCIENCE IN STATISTICS & PROGRAMMING
BACHELOR OF SCIENCE IN MATHEMATICS

SMA 364: APPLIED STATISTICAL METHODS

DATE: 23/4/2019

TIME: 2:00 – 4:00 PM

INSTRUCTION:

*Answer Question **ONE** which is compulsory and any other **TWO** Questions*

QUESTION ONE (COMPULSORY) (30 MARKS)

- a) Distinguish the following terms as they apply in data analysis
- i) A variable and a case
 - ii) Simple Linear and Multiple Linear Regression
 - iii) Parametric and Non-parametric tests
 - iv) Multicollinearity and outliers (6 marks)
- b) Highlight the process of importing a data from excel sheet to R platform (4 marks)
- c) Discuss any two methods of variable selection to a model fitting. (4 marks)
- d) The data below is a summary of slim possible finalists' weight difference in kilograms.

Participant	A	B	C	D	E	F	G	H
Weight Before	105	160	175	143	156	127	95	100
Weight after	85	124	172	123	111	139	99	77

Test the hypothesis that on average the exercise did not result to any significant weight loss

(6 marks)

- e) Distinguish the following terms as used in applied statistics
- i) Confounding variable and Noise variable
 - ii) Univariate and Bivariate data dimensions (4 marks)
- f) Highlight three assumptions underlying the least-squares regression (6 marks)

QUESTION TWO (20 MARKS)

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

Age group	Presidential Candidate			Total
	A	B	C	
Under 30	120	30	50	200
30 – 44	10	75	15	100
45 and above	10	30	60	100
Total	140	135	125	400

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

The table below is part of an output of data analyzed whose response variable was whether or not (1=yes,0=no) a student was in a relationship (y), the predictor variables included the students age (x1), gender (x2), fee balance (x3) in Ksh'000',family size (x4) and the religion coded such that (1=catholic,2=protestant,3=muslim,4=hindu) (x5)

Variables in the equation								
Step1 ^a	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Family-size	1.338	.494	4.419	1	.036	.354	.135	.932
Age	1.197	.120	2.710	1	.043	.821	.649	1.038
Fee bal	.201	.173	1.342	1	.247	1.222	.870	1.716
Religion			1.085	3	.781			
Religion(1)	-.073	1.477	.002	1	.961	.930	.051	16.801
Religion(2)	1.468	1.743	1.710	1	.047	4.342	.143	132.234
Religion(3)	-19.860	11.714	.000	1	.999	.000	.000	.
Gender(1)	.645	1.343	.231	1	.631	1.906	.137	26.476
Constant	8.329	4.262	3.819	1	.051	412.576		

- By citing the reasons highlight the variables that contributed significantly to the prediction of the relationship status of a student.
- Interpret the betas, $\text{Exp}(\beta)$ and the 95%CI for $\text{Exp}(\beta)$ of the variables highlighted in (i)
- Fit in the regression equation using only the significant variables. (12 marks)

QUESTION THREE (20 MARKS)

- Highlight four ways of carrying out regression diagnostics (4 marks)
- Explain the following terms as used in applied statistics
 - Fixed model effects
 - Random model effects
 - Mixed model effects (6 marks)
- One type of ladies gel was placed at five different heights within the same season. Sales at each level were recorded as summarized in table 3

Table 3:

Height level Placed				
2 feet	3 feet	4 feet	5 feet	6 feet
26	46	35	55	41
27	39	42	46	39
32	35	37	49	37
38	37	43	45	35
37	48	38	42	38

- Perform a one – way analysis of variance to test the hypothesis that the five different heights yielded the same average sales at $\alpha=0.01$
- Which heights differed significantly and by how much (10 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) Fit in the non-linear regression connecting the medical cost over time and write down all the steps and the non linear equation.
- ii) Highlight three advantages of non-linear regression over the linear regression (10 marks)

QUESTION FIVE (20 MARKS)

- a) Distinguish the following terms as they apply in exploratory data analysis
- Exploratory data analysis (EDA) and Response Surface Methodology (RSM)
 - Saturated and parsimonious model (6 marks)
- b) Afrifix is a NGO assisting the slums poor in the major cities in Africa. The management wishes to investigate whether the poverty index can be predicted from the family size, the family head; age, gender, formal education level and the city. The following is a sample of the family information and their poverty status.

City (X1)	A	B	C	D	F	G	H
Age(X2)	33	57	32	67	27	66	50
Gender(X3)	F	F	M	F	M	F	M
Educ (X4)	basic	sec	tertiary	basic	basic	tertiary	sec
Family size(X5)	K	K	K	M	M	M	M
Poverty Status (y)	Better	Better	Poor	Poor	Poor	Better	poor

Using the data above answer the following

- Capture and save the data in excel and then import and save it as a R file
- Using R software.
 - Fit a univariate linear regression model for each regressor to explain y
 - Fit a multiple linear regression model using x_1, x_2, x_3, x_4 and x_5 to explain y
- Write the equation of the best fit and interpret the coefficients. (14 marks)