MACHAKOS UNIVERSITY

University Examinations 2017/2018

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE IN MATHEMATICS

BACHELOR OF SCIENCE IN STATISTICS & PROGRAMMING

SMA 364: APPLIED STATISTICAL METHODS

DATE: 6/12/2017 **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. Simple Linear and Multiple Linear Regression
   ii. Parametric and Non-parametric tests
   iii. 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) Explain briefly the following terms as used in applied statistics (4 marks)

i. Confounding variable

ii. Noise variable

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:

	Presidential Candidate			
Age group	A	B	C	Total
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								
							95% C.I.for EXP(B)	
Step1[a]	B	S.E.	Wald	df	Sig.	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		

i. By citing the reasons highlight the variables that contributed significantly to the prediction of the relationship status of a student.
ii. Interpret the betas, Exp(β) and the 95%CI for Exp(β) of the variables highlighted in (i)
iii. Fit in the regression equation using only the significant variables. (12 marks)

QUESTION THREE (20 MARKS)

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

b) Explain the following terms as they apply in data analysis and applied statistics

i) Fixed model effects

ii) Random model effects

iii) Mixed model effects (6 marks)

c) 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

i) Perform a one – way analysis of variance to test the hypothesis that the five different heights yielded the same average sales at α=0.01

ii) 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.

ii. Highlight three advantages of non-linear regression over the linear regression (10 marks)

QUESTION FIVE (20 MARKS)

a) Discuss three ways of dealing with the problem of an outlier in a data set (6 marks)

b) Plains view operates hotels in 11 cities of medium size in Africa. The management is considering an expansion into other cities of medium size and wishes to investigate whether the number of tourists visiting per annum (Y) in a city can be predicted from the number of graduates with bachelor of tours and guide certificates in the city(X_1), the disposable personal income in the city(X_2), the level of political temperatures and internal conflicts within the country categorized as high "H" or low "L" (X_3) and then the security levels in the cities(X_4). Categorized as "K" terrorist unlikely, "L" terrorist likely and "M" terrorist most likely, Returns and incomes are expressed in thousands of dollars. Data on these variables for the year 2016 for the 11 cities in which Plains is now operating are shown in the table below:

City	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
X1	69	45	91	49	47	66	50	52	48	38	88
X2	16.7	16.8	18.2	16.3	17.3	18.2	15.9	17.2	16.6	16	18.3
X3	H	H	L	H	L	L	H	L	H	H	L
X4	K	K	K	M	M	M	M	L	L	L	L
Y"000"	174.4	164.4	244.2	154.6	181.6	207.5	152.8	163.2	145.4	137.2	241.9

Using the data above answer the following

i. Capture and save the data in excel and then import and save it as a R file

ii. Using R software.

1. Fit a univariate linear regression model for each regressor to explain y
2. Fit a multiple linear regression model using x1, x2, x3 and x4 to explain y

iii. Write the equation of the best fit and interpret the coefficients. (14 marks)