



MACHAKOS UNIVERSITY

University Examinations 2020/2021 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR SPECIAL /SUPPLEMENTARY EXAMINATION FOR
BACHELOR OF SCIENCE IN STATISTICS AND PROGRAMMING

SMA 364: APPLIED STATISTICAL METHODS

DATE: 23/3/2021

TIME: 2.00-4.00 PM

INSTRUCTIONS TO CANDIDATES

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

QUESTION ONE (30 MARKS)

- a) Distinguish the following terms as they apply in data analysis
- A variable and a case
 - Simple Linear and Multiple Linear Regression
 - Parametric and Non-parametric tests
 - Multicollinearity and outliers (6 marks)
- b) Discuss any two methods of variable selection to a model fitting. (4 marks)
- c) Write the R script to test the hypothesis that $H_0=70$ against $H_0\neq 70$ using the randomly sampled values $A=\{70,80,55,60,75,65,85,60,60,50\}$ (4 marks)
- d) Eight upper primary school pupils visiting the school clinic had their age and weight taken as summarized in the table below.

Student	A	B	C	D	E	F	G	H
Age (years)	9	10	11	12	13	13	14	14
Weight (kg)	25	23	24	26	36	35	35	34

Write the R script for executing the following

- i. The average weight
 - ii. Scatter diagram for age and weight
 - iii. The correlation between age and weight
 - iv. The regression equation weight as a function of age
 - v. Box plots for age and weight in a common axis (10 marks)
- e) Distinguish the following terms as they apply in data analysis
- i. Saturated and parsimonious model
 - ii. Confounded variable and noise variable
 - iii. Univariate and Bivariate data dimensions (6 marks)

QUESTION TWO (20 MARKS)

- a) 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	113	114	116	116	115	115	84

Table 1: Sample t-test output

	<i>t</i>	<i>df</i>	<i>Sig.(2-tailed)</i>	<i>Mean difference</i>	<i>95% CI of the difference</i>	
					<i>Lower</i>	<i>Upper</i>
<i>Attendance</i>	2.8	7	0.023	25.38	4.64	46.11
	9					

- i. State the hypothesis for the above scenario
 - ii. Write the R script that executed the given output
 - iii. Based on the sample t-test output make statistical conclusion
 - iv. Interpret the 95% CI lower and upper difference values (12 marks)
- b) 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)

Variables in the equation								
Step1 ^a	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Family-size	1.33	.49	4.41	1	.036	.354	.135	.932
Age	1.19	.12	2.71	1	.043	.821	.649	1.038
Fee (balance)	.20	.17	1.34	1	.247	1.222	.870	1.716
Religion			1.08	3	.781			
Religion(Protestants)	-.07	1.48	.01	1	.961	.930	.051	16.801
Religion(Catholics)	1.46	1.74	1.71	1	.047	4.342	.143	132.234
Religion(Muslims)	-19.86	11.71	.00	1	.999	.000	.000	.
Gender(Male)	.65	1.34	.23	1	.631	1.906	.137	26.476
Constant	8.32	4.26	3.81	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, $\text{Exp}(\beta)$ and the 95%CI for $\text{Exp}(\beta)$ 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 two ways of carrying out regression diagnostics (2 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) The weight , length(inch) and the gender of 12 children born on 1st January 2020 in a certain hospital in Nairobi was recorded as below

Child	A	B	C	D	E	F	G	H	I	J	K	L
Weight	2.9	4.3	5.2	3.5	3.7	2.5	4.0	3.9	4.3	5.1	3.2	4.1
Gender	F	F	M	M	F	F	M	F	F	M	M	F
Length	8	9.2	10.2	7.7	7.9	6.5	7.6	8.1	8.0	9.4	7.2	8.3

Write down the R script for determining

- Average weight and length of the 12 children
- Middle weight and height
- The box plot for weights per gender in a single axis
- Histogram of the weight distribution
- A correlation linear graph for weight and length

(12 marks)

QUESTION FOUR (20 MARKS)

- Discuss five ways of dealing with the problem of an outlier in a data set (10 marks)
- 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 $y = \beta_0 + \beta_1x_1 + \beta_2x_2^2 + \beta_3x_3^3$

Participants	A	B	C	D	E	F	G	H	I	J
Medical cost	34	45	32	11	10	36	57	77	81	90
Time in years	1	4	16	26	24	43	49	59	67	70

The non- linear regression equation output

<i>Variables</i>	<i>Coefficients</i>	<i>Std. error</i>	<i>t-value</i>	<i>Pr(> t)</i>
<i>intercept</i>	45.6471302	8.5582253	5.334	0.00108 **
<i>x</i>	-2.9540370	1.2151151	-2.431	0.04535 *
<i>X^2</i>	0.0939042	0.0418626	2.243	0.05980 .
<i>X^3</i>	-0.0006124	0.0003905	-1.568	0.16083
<i>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</i>				
<i>Residual standard error: 10.09 on 7 degrees of freedom</i>				

Multiple R-squared: 0.9016, Adjusted R-squared: 0.8595

F-statistic: 21.38 on 3 and 7 DF, p-value: 0.0006682

- i. Using the data and the output above write the R script for fitting in the age as a function of the medical cost
- ii. Explain the significance of each coefficient and the model (10 marks)

QUESTION FIVE (20 MARKS)

- a) Discuss three ways of dealing with the problem of an outlier in a data set (4 marks)
- b) A MSc student on the main causes of basic mathematics among BSc undergraduate students. Below is a sample of his factors and students he interviewed. The response variable is the score categorized as either passed or failed against; gender, residence categorized as either inside the university or outside, age, and entry marks.

Score	P	F	P	F	P	P	F	P	F	P	F
Age	23	26	21	24	22	23	25	22	24	20	26
Gender	F	M	F	F	F	M	F	F	M	M	M
Residence	in	out	in	out	out	in	in	in	out	in	out
Entry	61	54	58	45	56	55	64	62	60	54	49

Using the data above write the R script to answer the following

- i. Fit a simple logistic regression model using age to explain score
- ii. Fit a multiple logistic regression model using age and gender to explain score
- iii. Fit a multiple logistic regression model using age, gender, residence and entry to explain score
- iv. By combining all the independent variable into a design matrix (X)
Compute the coefficients of the best fit (16 marks)