AMREF INTERNATIONAL UNIVERSITY

SCHOOL OF PUBLIC HEALTH

DEPARTMENT OF COMMUNITY HEALTH

MASTER OF PUBLIC HEALTH

END OF SEMESTER EXAMINATION DECEMBER 2023

MAP 716: BIOSTATISTICS II AND COMPUTTING
DATE: 11th December 2023
TIME: Three Hours **Start:** 1600 Hours **Finish** 1900 Hours

INSTRUCTIONS

1. This exam is marked out of 100 marks
2. This Examination comprises TWO Sections
   Section A: Compulsory Question (25 marks)
   Section B: Long Answer Questions (75 marks)
3. All questions in Section A are compulsory and Answer any THREE questions in Section B
4. This online exam shall take 3 Hours
5. Late submission of the answers will not be accepted
6. Ensure your web-camera is on at all times during the examination period
7. No movement is allowed during the examination
8. Idling of your machine for 5 min or more will lead to lock out from the exam
9. The Learning Management System (LMS) has inbuilt integrity checks to detect cheating
10. Any aspect of cheating detected during and or after the exam administration will lead to nullification of your exam
11. In case you have any questions call the invigilator for this exam on Tel. 0722840012 and or the Head of Department on Tel +254 727239519
12. For adverse incidences please write an email to: amiu.examinations@amref.ac.ke and jarim.omogi@amref.ac.ke

SECTION A: COMPULSORY (25 Marks)

Question 1

a) Describe with appropriate examples the application of logistic regression model (**8 marks)**

b) A research interviewed university students and asked if they have ever driven after drinking. They also were asked, "How many days per month do you drink at least two beers?" In the following discussion, $\pi\pi$ = the probability a student says "yes" they have driven after drinking. This is modeled using X = days per month of drinking two beers. Results were as follows.

- Drink drive yes 122, no -127 (daysbeer)
- Gender 1-male, 0-female

Using logistic regression model the output from analysis is as shown below. Study the output and answer the following questions

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	Upper
Constant	-1.7736	0.2945	-6.02	0.000			
DaysBeer	0.18693	0.03004	6.22	0.000	1.21	1.14	1.28
Gender							
male	0.6172	0.2954	2.09	0.037	1.85	1.04	3.31

```
Log-Likelihood = -139.981
Test that all slopes are zero: G = 65.125, DF = 2, P-Value = 0.000
```

i. Write the regression equation **(4 marks)**

ii. Interpret the results from the analysis **(4 marks)**

iii. Why was the logistic regression preferred for this analysis, explain (3 marks)

iv. Write a detail summary of findings **(6 marks)**

SECTION B: SELECT ANY THREE (3) QUESTIONS -75 MARKS

Question 2

a) What is the function of ANOVA in regression analysis**?** (2 marks)

b) Despres et al. point out that the topography of adipose tissue (AT) is associated with metabolic complications considered as risk factors for cardiovascular disease. It is important, they state, to measure the amount of intra abdominal AT as part of the evaluation of the cardiovascular-disease risk of an individual. Computed tomography (CT), the only available technique that precisely and reliably measures the amount of deep abdominal AT, however, is costly and requires irradiation of the subject. In addition, the technique is not available to many physicians. Despres and his colleagues conducted a study to develop equations to predict the amount of deep abdominal AT from simple anthropometric measurements. Their subjects were **109** men between the ages of 18 and 42 years who were free from metabolic disease that would require treatment. Among the measurements taken on each subject were deep abdominal AT obtained by CT and waist circumference as shown in Table 1 (shows data for aonly16 of the 109 patients).

TABLE 1 Waist Circumference(cm), X, and Deep Abdominal AT, Y, of 109 men

Subject	X	Y	Subject	X	Y	Subject	X	Y
1	74.75	25.72	38	103.00	129.00	75	108.00	217.00
2	72.60	25.89	39	80.00	74.02	76	100.00	140.00
3	81.80	42.60	40	79.00	55.48	77	103.00	109.00
...	...	...	...	...	...	...	...	...
...	....	...	...	...	...	...	...	...
...	...	...	...	...	...	.	...	...
35	102.00	127.00	72	101.00	154.00	109	108.50	208.00
36	94.50	121.00	73	97.00	100.00			
37	91.00	107.00	74	100.00	123.00			

A simple regression between Deep Abdominal (Y) and Waist Circumference (X) was performed. Partial STATA output shown below depict the analysis.

```
. regress y x

      Source |       SS       df       MS              Number of obs =     109
-------------+------------------------------           F(  1,   107) =  217.35
       Model |  237574.528     1  237574.528           Prob > F      =  0.0000
    Residual |  116955.976   107  1093.04651           R-squared     =  0.6701
-------------+------------------------------           Adj R-squared =  0.6670
       Total |  354530.504   108  3282.68985           Root MSE      =  33.061

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   3.459007   .2346231    14.74   0.000     2.993894     3.92412
       _cons |  -215.9887   21.79316    -9.91   0.000    -259.1911   -172.7863
------------------------------------------------------------------------------
```

i. What are the hypotheses associated with this test? (4 marks)

ii. What is the result of this test (2 marks)

iii. What conclusion do you reach with respect to the existence of a linear association between Deep abdominal and waist circumference? (3 marks)

iv. Write down the estimated regression line. (2marks)

v. What is the estimate of the slope of the line? (2marks)

vi. What is the test that is associated with this estimate and what are the hypotheses associated with it? (2marks)

vii. State the result of this test and your conclusion in terms of the problem. (4 marks)

viii. What is the implication of the estimated slope? (2 marks)

ix. What is the implication of the intercept? (2 marks)

Question 3

The following data illustrates the response times of two groups of 18 students subjected to a battery of cognitive tests. The tests were administered to random sample of 10 and 8 students and the reaction times recorded as shown below;

Student No.	1	2	3	4	5	6	7	8	9	10
TEST 1 (hr.)	18.5	16.0	22.7	17.6	18.9	25.5	16.5	24.2		
TEST 2 (hrs.)	12.6	14.5	20.5	10.7	15.9	19.6	12.0	15.5	11.9	14.9

We would like to determine if the two tests differ in cognitive response times.

a) Develop a null and alternative hypothesis for this data (4 marks)

b) Use the Mann-Whitney U test to test your hypotheses (12 marks)

c) What conclusions do make from your analysis? (5 marks):

d) Why was the Mann-Whitney U-test preferred for this data (4 marks)?

Question 4

The table below shows the results from a multiple regression analysis carried out to assess what factors can predict barriers to mammography screening in women. The predictors were: Age in years, Education (years of formal education) and Race (1=African, 0= Other).

Overall model: $F = 16.58$, $df = 3$ and 1035, $p < .0001$

$R2 = .046$ Adjusted $R2 = .043$

Least squares estimates of parameters

Variable	df	Regression Coefficient	Standard Error	t-value	Pr>|t|
Intercept	1	26.63	2.277	11.69	<.0001
Age	1	0.087	0.026	?	0.0024
Education	1	-0.265	0.088	?	0.0011
Race	1	3.218	0.559	?	<.0001

---

a) What hypothesis is being tested by the F-value? **(2 marks)**

b) What is the conclusion from this test? **(3 marks)**

c) What is the interpretation of the R-square **(3 marks)**

d) What is the difference between the R-square and the adjusted R square **(2 marks)**

e) Fill in the missing t-value's **(3 marks)**

f) State the null and alternative hypothesis being tested by the t-value associated with Age **(3 marks)**

g) What is the conclusion for this test? **(4 marks)**

h) What is the interpretation of the intercept? **(2.5 marks)**

i) What is the interpretation of the coefficient associated with education? **(2.5 marks)**

Question 5

a. Explain in detail the application of principal components analysis and use appropriate examples (7 Marks)

b. Explain in detail the application of factor analysis using appropriate examples (8 marks)

c. Factor analysis involves finding estimates of factor loading and their communalities. Describe the approach used in achieving loadings and communalities **(10 marks)**

Question 6

Explain the following concepts is used in multivariate techniques

a) Non-linear regression function (6 marks)
b) Heteroscedasticity (5 marks)
c) Dependence of error terms (5 marks)
d) Outliers and influential observations (5 marks)
e) Censoring (4 marks)