AFRICA UNIVERSITY
A United Methodist-Related Institution

"Investing in Africa's Future"

COLLEGE OF HEALTH, AGRICULTURE AND NATURAL SCIENCES

SPH 541 HEALTH STSTISTICS

SUPPLEMENTARY EXAMINATION

LECTURER: E. CHIKAKA

DURATION: (3 HRS)

INSTRUCTIONS

Answer **ALL** Questions in **Section A** and **ANY 3** questions from **Section B**

The mark allocation for each question is indicated at the end of the question

Credit will be given for logical, systematic and neat presentations.

SECTION A

QUESTION 1

a Discuss how the sample size of a study can influence the results of inferential statistical tests with regard to clinical significance. [3]

b Describe type I and type II errors, and how they are related to power and significance level of a study. [3]

c List and briefly discuss the criteria you would apply if you wanted to establish causality between an exposure and a disease. [3]

d Explain some of the issues to consider when estimating the required sample size for a study. [3]

QUESTION 2

A data set contains information about the hourly wage (in U.S. dollars) and the gender of each of the 534 workers surveyed in 1985, as well as information about each worker's age, union membership, and education level (measured in years of education). A linear regression analysis was performed to model hourly wage as a function of worker sex, number of years of education, and union membership. Below find the results from this regression:

Hourly Wage (in U.S. dollars)	Regression Coefficient	Standard Error of Regression Coefficient
sex (1 = Female, 0 = Male)	-1.9	0.4
union member (1 = yes, 0 = no)	1.9	0.5
years of education	0.76	0.08
Intercept	-0.3	1

a. Write down the regression equation [2]

b. Interpret the regression coefficients for sex, union member and years of education [6]

c. Using the above regression results, estimate the mean hourly wage in 1985 for male workers with a high school education (12 years of education), who were not union members. [2]

d. Give a 95% confidence interval for the mean difference in hourly wages for male workers with a high school education who were union members when compared to male workers with a high school education who were not union members. [4]

QUESTION 3

a. What are the assumptions underlying multiple linear regression analysis when one wishes to infer about the population from which the sample data have been drawn? How do you check for these assumptions? [10]

b. Explain the following terms:

i. Attributable Risk
ii. Odds
iii. Dummy Variable
iv. Relative Risk [8]

SECTION B

QUESTION 4

This study evaluated a maternal and child health program in which all pregnant women in the intervention area were monitored by village health workers. During the one-year study period data were collected during pregnancy, at delivery, and 28 days after delivery. Only single births of 28 weeks or more gestational age were included in the study. The primary outcomes of interest were birth weight and the infant's survival status at 28 days.

The Village Health Worker Programme involved the selection of two health workers from each village in the intervention area. Their function was to provide simple treatment such as iron supplementation for anaemia, to inform pregnant women of services available to them from the health centre, to monitor them throughout pregnancy, to identify high risk cases and arrange for necessary referrals. The pregnancy and delivery were monitored via the completion of the mother's card. This card contained demographic and medical data and information was recorded on the card throughout pregnancy together with the outcomes of delivery.

During the one-year study period there were 939 single deliveries in the intervention area and 944 in the control area. For the purpose of this exercise, records for 65 cases of perinatal death and 400 survivors were selected. The variables included in the dataset are:

Variable name	Interpretation	Codes
Status	infant's survival status at 28 days	0= alive 1= dead 1= abnormal
Birthwt	infants birth weight in grams	grams (cont. variable)
Mothage	mother's age in years	years (cont. variable)
DCHILD	no of live births now dead of mother	0 = none 1 = 1+

. logit status birthweight mothage dchild

```
Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -171.62235
Iteration 2:   log likelihood = -169.70456
Iteration 3:   log likelihood = -169.69649
Iteration 4:   log likelihood = -169.69649
Logit estimates                           Number of obs   =        465
                                          LR chi2(3)      =      36.86
                                          Prob > chi2     =     0.0000
Log likelihood = -169.69649               Pseudo R2       =     0.0980
------------------------------------------------------------------------------
      status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  birthweight |  -.0014929   .0003285    -4.54   0.000    -.0021368   -.0008491
     mothage |  -.0158665   .0287126    -0.55   0.581    -.0721421    .0404092
      dchild |   .7844647   .2196172     3.57   0.000     .3540228    1.214907
       _cons |   2.665993   1.138809     2.34   0.019     .4339674    4.898018
------------------------------------------------------------------------------
```

a. Using the logistic regression separately examine the relationship between each of the independent variables birth weight (as continuous), mothage and dchild and the outcome variable STATUS [6]

b. From the logistic regression models fitted in part (a) calculate the odds ratios corresponding to:
   (i) an increase in birth weight of 100grams
   (ii) an increase in mother's age of 5 years
   (iii) one additional dead child [6]
   Give an interpretation of each of these odds ratios [6]

a. Which of the following independent variables are statistically significantly related to infant's survival at 28 days after adjusting for the others? [2]
   *****Independent variables: birth weight, mother's age, number of dead children. (do not investigate interactions) *****

QUESTION 5

a. In a study conducted by Familiar et al (1989) two drugs for the treatment of peptic ulcer were compared. The percentage of ulcers healed by the drug pirenzepine was 85% and 70% of ulcers were healed by trithiozone.

   (i) How many subjects would be required for a randomised trial to have 80% chance of detecting this difference (85% and 70%) at the 0.05 (two-tailed) level of significance? [2]
   (ii) How many subjects would be required to increase the power to 90%? [2]
   (iii) Taking the healing rate for trithiozone to be 70%, what sample size would be required to detect an increase of 20% in the healing rates due to pirenzepine? (Use 80% power) [2]
   (iv) If only 40% of the subjects are to be given trithiozone, what sample size is now required to detect the difference described in part (i) [4]

b. The role of circulating catecholamine in essential hypertension has been a subject of intense study. You wish to carry out a study comparing serum catecholamine levels in normotensive patients and patients with essential hypertension. Previous studies have found mean serum catecholamine levels of 0.812mg/mL. (sd = 0.41) in normotensives
   (i) If the clinically important difference to be detected in catecholamine levels in hypertensive patients is an increase by 0.2 mg/mL. How many subjects would you sample? [2]
   (ii) How many subjects would you need if the difference to be detected was
      (a) 0.08 mg/mL
      (b) 0.065 mg/mL [4]
   (iii) The study to investigate catecholamine levels will be under taken on patients attending their general practitioner. Since 50 subjects will be selected for each GP enrolled in the study, the issue of clustering may be a problem. What sample size is required to detect a difference of 0.1 mg/mL with 80% power if the intra-cluster correlation coefficient is estimated to be:
      (a) 0.07
      (b) 0.20 [4]

QUESTION 6

The use of pesticides and their effect on health has been of interest for some time. A common pesticide used to treat termites is dieldrin, and all new houses must be pretreated with this chemical. A study was undertaken to examine factors associated with high levels of dieldrin in breastmilk. The independent variables were mother's age, suburb (old or new), and whether or not house has been treated for termites. The variables are defined as follows:

age	mother's age in years	
suburb	0 = lives in old suburb	1 = lives in new suburb
trtterm	house treated for termites	
	0 = not treated for termites	1 = treated for termites
hidiel	0 = low dieldrin level (≤0.009ppm);	
	1 = high dieldrin level (>0.009ppm)	

The STATA output for the analysis is shown below

logithidiel age suburb trtterm
iteration 0: log Likelihood = -59.440326
iteration 1: log Likelihood = -50.089793
iteration 2: log Likelihood = -49.703196
iteration 3: log Likelihood = -49.698005
iteration 4: log Likelihood = -49.698004

logit estimates	Number of Obs	=92
	Chi2(3)	=19.48
	Prob>chi2	=0.0002
Log likelihood= -49.698004	Pseudo R2	=0.1639

Hidiel	Coef.	Std Err .	Z	p>|z|	95% Conf.	Interval
Age	.0104068	.0735977	0.141	0.888	-0.133842	0.1546555
Suburb	1.356893	.5531554	2.453	0.014	0.2727286	2.441058
Trtterm	2.009632	.5721082	3.513	0.000	0.8883202	3.130943
_cons	-2.648924	2.144632	-1.235	0.217	-6.852325	1.554478

a. Which type of analysis has been conducted? [2]
b. Which variables are statistically significantly associated with high dieldrin levels in breast milk? Justify your answer. [6]
c. What does the coefficient of 1.36 for suburb mean [4]
d. calculate and interpret the odds ratio and 95% confidence interval for the relationship between treatment of termites with dieldrin and high dieldrin levels in breast milk [8]

QUESTION 8

The following are findings from a hypothetical study on the association between alcohol use and mortality from coronary heart disease among Zimbabwean doctors showing relative risks of death adjusted for age and smoking habits.

Drinking habits	relative risk	(95% confidence interval)
Non-drinker	1.0	
Occasional drinker	0.85	(0.4 – 0.9)
<2 drinks daily (beer/wine/spirits)	0.76	(0.5 – 1.1)
≥2 drinks daily (beer/wine/spirits)	0.67	(0.4 - 1.1)
Ex-drinker	1.59	(1.0 – 2.4)

a. Interpret the results (6)
b. What could explain the results in ex-drinkers (7)
c. What type of study was this? (2)
d. The authors adjust for age. Is this appropriate? Explain your answer (5)

QUESTION 8

Data from 47 patients receiving a non-depleted allogeneic bone marrow transplant were examined to see which variables were significantly related to survival time. Backward stepwise Cox regression analysis using diagnosis (CML or not), recipient's age and sex, donor's age and sex, whether the donor had been pregnant, the index of mixed epidermal cell-lymphocyte reactions and whether or not patient developed graft vs host disease (GvHD) to predict survival yields the following model:

Variable	**Regression Coefficient**	**Standard Error**
GvHD (0=No, 1= Yes)	2.4303	0.59898
CML (0=No, 1= Yes)	-2.8505	0.86095

a. What is the interpretation of the opposite signs for the regression coefficients? [3]
b. Calculate the relative risks of dying (hazard ratios) for the following patients relative to non-GvHD non-CML patients:
   (i) With GvHD but not CML
   (ii) CML but without GvHD
   (iii) CML and GvHD [9]
c. Calculate the 95% confidence interval for the hazard ratio associated with GvHD [6]
d. Comment on the reliability of the Cox regression model in view of the sample size (47) and the number of deaths (18). [2]