AFRICA UNIVERSITY
(A United Methodist-Related Institution)

"Investing in Africa's Future"
FACULTY OF HEALTH SCIENCES

FIRST SEMESTER EXAMINATIONS

COURSE CODE:	NSPH541
COURSE TITLE:	HEALTH STATISTICS
DATE:	MAY 2022
TIME:	3 hours

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: 20 marks

a) State the **four** main factors that influence sample size calculation in analytic studies. [4]

b) State the **three** methods that are used to assess for significance in hypothesis testing and describe how a decision is made for each method stated. [6]

c) In public health studies, researchers often use secondary data to answer public health research questions

i. Define what is meant by secondary data? [1]

ii. State **two** sources of secondary data [2]

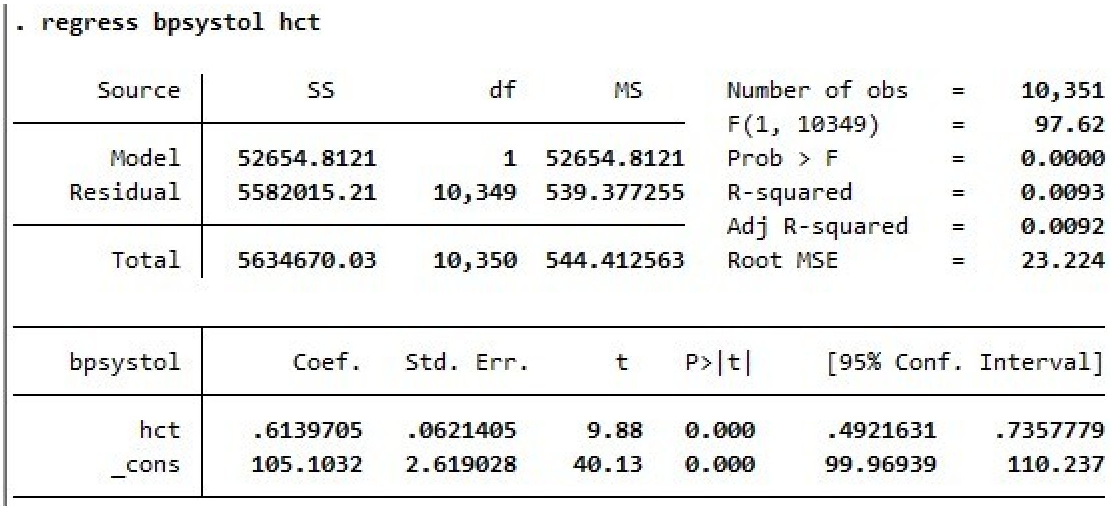
iii. State **two** advantages of using secondary data [2]

iv. State **two** disadvantages of using secondary data [2]

d) State **three** differences between logistic regression and linear regression [3]

QUESTION 2: 20 marks

A clinician wishes to assess the effect of hematocrit (htc) on patients' systolic blood pressure (bpsystol). The clinician consulted an MPH student on what analysis to perform and was advised to use a simple linear regression. Below are the results from the analysis performed by the clinician using STATA.



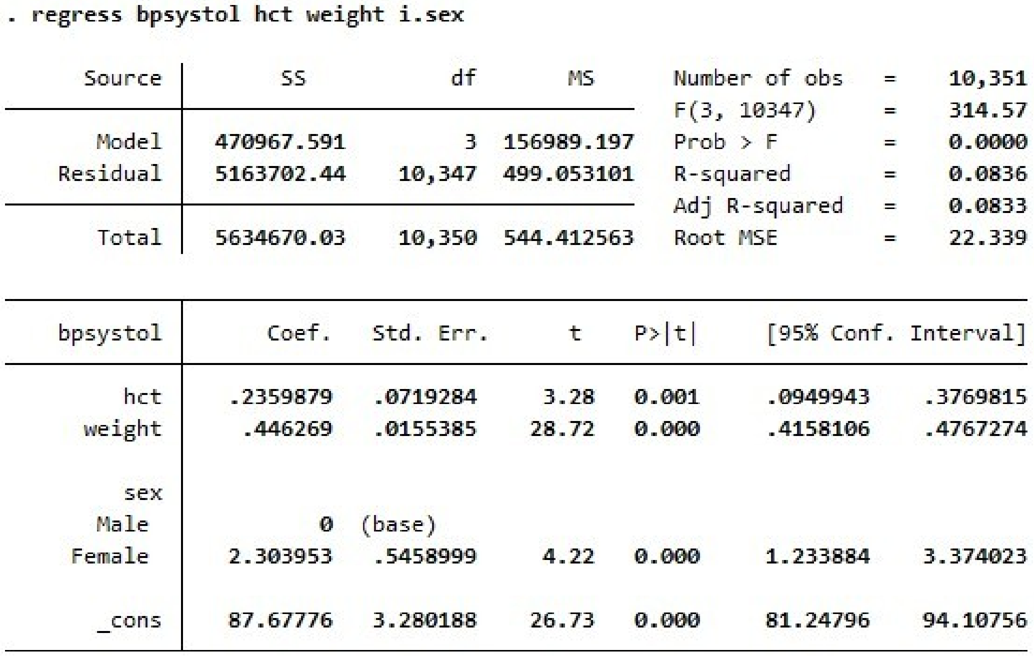
```
. regress bpsystol hct

      Source |       SS           df       MS      Number of obs   =    10,351
-------------+----------------------------------   F(1, 10349)     =     97.62
       Model |  52654.8121         1  52654.8121   Prob > F        =    0.0000
    Residual |  5582015.21    10,349  539.377255   R-squared       =    0.0093
-------------+----------------------------------   Adj R-squared   =    0.0092
       Total |  5634670.03    10,350  544.412563   Root MSE        =    23.224

------------------------------------------------------------------------------
    bpsystol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hct |   .6139705   .0621405     9.88   0.000     .4921631    .7357779
       _cons |   105.1032   2.619028    40.13   0.000     99.96939     110.237
------------------------------------------------------------------------------
```

a) Write down the simple linear regression equation and interpret the intercept and slope coefficients. [1+2+2=5]

b) Comment on the variability of hematocrit (htc) in explaining the patients' systolic blood pressure (bpsystol) based on the model results. [2]

c) Predict the systolic blood pressure (bpsystol) of the person with a hematocrit (htc) value of 47.98 [2]

The clinician had information on patients' weight and sex so the researcher went on and fitted a multiple linear regression model after fitting the simple linear regression model. The results of the analysis are shown below:



```
. regress bpsystol hct weight i.sex

      Source |       SS           df       MS      Number of obs   =    10,351
-------------+----------------------------------   F(3, 10347)     =    314.57
       Model |  470967.591         3  156989.197   Prob > F        =    0.0000
    Residual |  5163702.44    10,347  499.053101   R-squared       =    0.0836
-------------+----------------------------------   Adj R-squared   =    0.0833
       Total |  5634670.03    10,350  544.412563   Root MSE        =    22.339

------------------------------------------------------------------------------
    bpsystol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hct |   .2359879   .0719284     3.28   0.001     .0949943    .3769815
      weight |    .446269   .0155385    28.72   0.000     .4158106    .4767274
             |
         sex |
       Male  |          0  (base)
     Female  |   2.303953   .5458999     4.22   0.000     1.233884    3.374023
             |
       _cons |   87.67776   3.280188    26.73   0.000     81.24796    94.10756
------------------------------------------------------------------------------
```

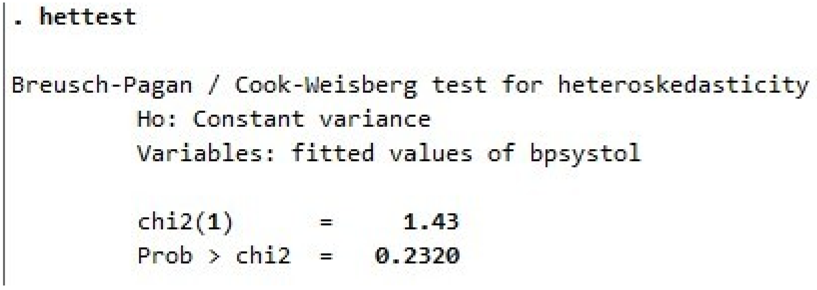
d) State the advantage (s) of fitting a multiple linear regression over simple linear regression. [2]

e) Interpret the sex coefficient fully [3]

f) Determine if weight and sex are possible confounders of the relationship between systolic blood pressure (bpsystol) and haematocrit (htc) for these patients [2]

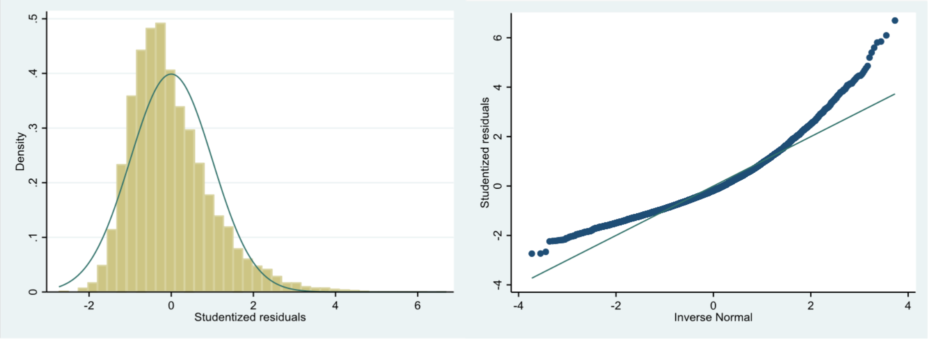
g) The following diagnostic assessments were done on the adjusted model.

i. State the assumption being tested and comment if valid or not [2]



```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of bpsystol

         chi2(1)      =     1.43
         Prob > chi2  =   0.2320
```

ii. State the assumption being tested and comment if valid or not [2]



SECTION B

QUESTION 3: 20 marks

a) A researcher from the cancer registry in Zimbabwe set to conduct a study to determine if smoking was a risk factor for lung cancer. A sample size of 160 record reviews was retrieved from the registry and the data is summarised in the following 2x2 table:

	Lung Cancer		
Current Smoker	**Yes**	**No**	**Totals**
Yes	a	30	b
No	25	c	45
Totals	d	50	160

i. Fill in the missing values in the results table [2]

ii. State the appropriate measures of association for this study [1]

iii. Calculate the measure of association [2]

iv. Using a level of significance (α) of 0.05, determine if smoking is significantly associated with having lung cancer. [5]

b) A study on the low birth weight of babies born at Hospital A collected the following data.

```
Contains data from https://www.stata-press.com/data/r16/lbw.dta
obs:            189                          Hosmer & Lemeshow
data  vars:            11                           15 Jan 2018
05:01
                                                                             _
                    storage   display    value variable name   type    format     label      variable label

id              int     %8.0g              identification code low             byte
%8.0g                birthweight<2500g age              byte    %8.0g               age
of mother lwt             int     %8.0g              weight at last menstrual period
race            byte    %8.0g      race        race smoke          byte    %9.0g
smoke      smoked during pregnancy ptl               byte    %8.0g
premature labor history (count) ht              byte    %8.0g               has history
of hypertension ui              byte    %8.0g              presence, uterine
irritability ftv             byte    %8.0g              number of visits to physician
during 1st trimester
```

```
bwt             int     %8.0g                 birthweight (grams)
```

The researcher aimed to determine the factors associated with low birth weight at this hospital and the following analysis was performed.

```
. logistic low lwt i.race i.smoke ht, base

Logistic regression                                Number of obs     =
189
                                                   LR chi2(5)        =
26.40
                                                   Prob > chi2       =
0.0001
Log likelihood = -104.13523                        Pseudo R2         =
0.1125
-------------------------------------------------------------------------
            low | Odds Ratio   Std. Err.       z     P>|z|      [95% Conf.
            Interval]
-------------------------------------------------------------------------

lwt             |            .9822801         .0066765             -2.63
            0.009     .9692812    .9954533


race            |
white           |
                |        1  (base)
      black     |   3.622729    1.889602      2.47    0.014       1.303324
            10.06977
      other     |   2.570489    1.088183      2.23    0.026       1.121168
            5.89333


smoke           |
nonsmoker       |
                |        1  (base)
     smoker     |   2.919794     1.13138      2.77    0.006       1.366229
            6.239948
                |   5.743724    3.967166      2.53    0.011       1.483448
ht          22.23898
          _cons |   1.417217    1.309925      0.38    0.706       .2315665
            8.673551
```

i. State the differences between this type of analysis and linear regression [3]

ii. Interpret the effect of smoking on childbirth weight [3]

iii. Interpret the effect of weight at last mensural period (lwt) on child-birth weight [2]

iv. State what additional analysis the researcher could have done. [*Hint: Link your answer to the variables in the dataset*] [2]

QUESTION 4: 20 marks

An MPH student at Africa University set to determine the effect of two drugs for the treatment of peptic ulcers. Previous studies reported that the percentage of ulcers healed by the pirenzepine drug was 65% while the trithiozone drug healed 55% of ulcers.

a) How many participants would be required for a randomised trial to have 80% power of detecting this difference between the two drugs (65% and 55%) at the 0.05 (twotailed) two-tailed level of significance? [6]

b) How many participants would be required to increase the power to 90%? [3]

c) With reference to the sample size in (a) and (b), comment on the relationship between sample size and power of the study [3]

d) A new study was published and reported that the healing rate for trithiozone was 70% and the healing rate due to pirenzepine was 20% higher. The study considered recalculating the sample size using the new information. What sample size would be required to detect an increase of 20% in the healing rate at a 5% level of significance and 80% power? [5]

e) Taking into account the 10% attrition rate, what is the final sample size the student should use from (d)? [3]

QUESTION 5: 20 marks

a) Describe type I and type II errors, and how they are related to the power and significance level of a study. [5]

b) Describe the steps for hypothesis testing giving examples to explain your points [10]

c) State the two types of hypothesis a researcher can perform [2]

d) A researcher carried out a hypothesis testing and got a test statistic value of which was greater than the critical value. What decision should the researcher make? [3]

QUESTION 6: 20 marks

a) State **three** differences between linear regression and survival analysis [3]

b) State one assumption that is made in survival analysis and describes how it is assessed. [3]

c) 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 Occasional drinker	1 0.65	 0.55-0.75
<2 drinks daily	0.76 0.98	0.70-0.82 0.91-1.05

≥2 drinks daily	1.56	1.26-1.86
Ex-drinker		

i. What is the study design conducted [2]

ii. Interpret the results fully [8]

iii. The authors adjusted for age and smoking habits. Explain the epidemiological concept this study was trying to account for how to outline the effects of not accounting for such issues in epidemiological studies [4]