W1-2-60-1-6

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

**UNIVERSITY EXAMINATIONS 2019/2020**

**MASTER OF SCIENCE IN EPIDEMIOLOGY**

**TPH3101: EPIDEMIOLOGIC METHODS II**

**DATE: FEBRUARY 2020** **TIME: 3 HOURS**

**INSTRUCTIONS: ANSWER ANY FOUR QUESTIONS**

1)

(a) State the appropriate multivariable regression technique that would be used to analyse each of the following studies:

(i) A cross-sectional study of the association between wine-drinking and the presence or absence of coronary artery disease. *(2 Marks)*

(ii) A prospective cohort study of the risk factors for Alzheimer's disease. *(2 Marks)*

(iii) A cross-sectional study of the association between three behaviors (fat intake, smoking, and exercise) and body-mass index. *(2 Marks)*

(iv) A case-control study of the utility of beta-carotene in preventing breast cancer. *(2 Marks)*

(v) A retrospective cohort study of the risk associated with breastfeeding for development of breast cancer. *(2 Marks)*

(b) A randomized placebo-controlled trial of a newly developed therapy to prevent recurrent breast cancer is performed. The 1200 subjects recruited are all high-risk women having had a prior history of breast cancer and unilateral mastectomy. The outcome is development of cancer in the remaining breast. All women are followed for 2 years. Follow-up is complete and no deaths occur. The following STATA output represents the results of this study showing the age-specific measures of association (for the <65 year old and >65 year old groups) between the new therapy and development of breast cancer in these women:

```
. cs  disease therapy, by(age)
```

| age | RR | [95% Conf. | Interval] | M-H Weight |
|---|---|---|---|---|
| < 65 years old | .2 | .0449546 | .8897864 | 5 |
| >= 65 years old | .9833333 | .7018033 | 1.3778 | 30 |
| Crude | .8714286 | .6301911 | 1.205012 | |
| M-H combined | .8714286 | .630477 | 1.204465 | |

Test of homogeneity (M-H) chi2(1) = 4.228 Pr>chi2 = 0.0398

(i) Does the new agent appear effective in the prevention of breast cancer recurrence? *(4 Marks)*

(ii) An *a priori* hypothesis of the trial was that the new drug would be more effective in younger women. Is that true from the results (i.e. is there evidence of effect modification by age)? *(6 Marks)*

(iii) Is there evidence of confounding by age? *(3 Marks)*

(iv) Outline two other confounding variables which you might consider in this study? *(2 Marks)*

2) The Heart and Oestrogen/progestin Replacement Study (HERS) was a randomized, double-blind, placebo-controlled trial designed to test the efficacy and safety of oestrogen plus progestin therapy for prevention of recurrent coronary heart disease (CHD) events in women. The participants were postmenopausal women with a uterus and with CHD (as evidenced by prior myocardial infarction). Among the risk factors measured were blood glucose levels (mg/dl), exercise (if the participant exercised or not), age (years), average number of alcoholic drinks per week (avgdrpwk), systolic blood pressure (sbp) and body mass index (bmi). The following is a STATA output of the simple regression model on the association between exercise (predictor) and blood glucose level (outcome) in these women:

```
. regress glucose exercise
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 2298.41225 | 1 | 2298.41225 |
| Residual | 567636.29 | 5461 | 103.943653 |
| Total | 569934.702 | 5462 | 104.345423 |

| | |
|---|---|
| Number of obs | = 5463 |
| F( 1, 5461) | = 22.11 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0040 |
| Adj R-squared | = 0.0039 |
| Root MSE | = 10.195 |

| glucose | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| exercise | -1.333987 | .2836851 | -4.70 | 0.000 | -1.890123 | -.7778509 |
| _cons | 97.11728 | .1756762 | 552.82 | 0.000 | 96.77288 | 97.46168 |

(a) Briefly describe the relationship between exercise and blood glucose using the results in this model. *(3 Marks)*

(b) Explain the meaning of the following outputs in the model:

(i) Exercise (P>| t | = 0.000) *(2 Marks)*

(ii) Coefficient (_cons = 97.11728) *(2 Marks)*

(iii) Probability (Prob > F = 0.0000) *(2 Marks)*

(c) Which two ways could be used to evaluate the fit of this model? *(4 Marks)*

(d) What assumptions were made in using the method of least squares to estimate the population line in this model? *(4 Marks)*

The following is a STATA output of the multiple regression model for the association between blood glucose level (outcome) and exercise and several other predictor variables:

```
. regress glucose exercise age avgdrpwk bmi
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 37754.454 | 4 | 9438.6135 |
| Residual | 528940.191 | 5434 | 97.3390119 |
| Total | 566694.645 | 5438 | 104.210122 |

| | |
|---|---|
| Number of obs | = 5439 |
| F( 4, 5434) | = 96.97 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0666 |
| Adj R-squared | = 0.0659 |
| Root MSE | = 9.8661 |

| glucose | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| exercise | -.6451864 | .2775 | -2.32 | 0.020 | -1.189197 | -.1011753 |
| age | .0535494 | .0203083 | 2.64 | 0.008 | .0137369 | .0933619 |
| avgdrpwk | .1291425 | .0349691 | 3.69 | 0.000 | .060589 | .197696 |
| bmi | .5063333 | .0266872 | 18.97 | 0.000 | .4540156 | .5586509 |
| _cons | 79.07654 | 1.677771 | 47.13 | 0.000 | 75.78744 | 82.36564 |

(e) Briefly describe the relationship between exercise and blood glucose using the results in this multiple regression model? *(4 Marks)*

(f) Why do you think the coefficient of exercise reduced from -1.3 in the simple regression model to -0.6 in the multiple regression model? *(4 Marks)*

3) In designing epidemiological studies:

(a) Explain why sample size calculation is important. *(6 Marks)*

(b) Briefly describe the two major classical approaches to sample size calculations in the design of quantitative studies. *(6 Marks)*

(c) What is the *p*-value, the significance level and the power of a test? *(6 Marks)*

(d) You want to carry out a study in Narok County to measure the prevalence of hypercholesterolemia (which is high fat content in the blood with total cholesterol of above 5.2mmol/l) in men aged 18 to 50 years with the hypothesis that high red meat intake might lead to high cholesterol levels in this population. A previous study done in Nairobi had reported that 15% of men aged 18 to 50 years had hypercholesterolemia. Suppose that the level of precision you require is such that the 95% confidence interval should be no wider than 5% (i.e. 0.05) (which means that the margin of error ($m$) should be no more than 0.025). Calculate the sample size you will require using the following formula: *(7 Marks)*

$$n = \left\{\frac{z_{\alpha/2}}{m}\right\}^2 \times \hat{p}(1-\hat{p})$$

4) Among women, there is convincing evidence that human herpesvirus 8 (HHV-8), the causative agent of Kaposi's sarcoma, is transmitted by injection drug use (i.e. sharing of needles), but it is controversial as to whether it is sexually transmitted. *Goedert et al. (AIDS 17:425, 2003)* evaluated the association between syphilis antibody status (a marker of sexual activity) and HHV-8 infection status among women in a cross-sectional analysis. Because of the concern about potential confounding by injection drug use, they performed the following analysis where they stratified participants into one of two categories of injection drug use: a) women who reported a history of injection drug use and/or who tested positive for hepatitis C virus (HCV)

infection; or b) women with no self-reported history of injection drug use and who tested negative for hepatitis C virus injection. *(Note: hepatitis C virus is known to be easily transmitted by injection drug use and hence is used as an objective and reasonably sensitive and specific surrogate of prior injection drug use behaviour).*

**Crude Data**

| | HHV-8 present | HHV-8 absent |
|---|---|---|
| Syphilis | 9 | 78 |
| No syphilis | 37 | 753 |

**Stratum Specific Data**

History of injection drug use

| | HHV-8 present | HHV-8 absent |
|---|---|---|
| Syphilis | 7 | 34 |
| No syphilis | 11 | 152 |

No history of injection drug use

| | HHV-8 present | HHV-8 absent |
|---|---|---|
| Syphilis | 2 | 44 |
| No syphilis | 26 | 601 |

(a) Calculate the crude odds ratio and the two stratum-specific odds ratios *(show your work).* *(9 Marks)*

(b) Calculate the Mantel-Haenszel adjusted odds ratio *(use the formula below and show your work).* *(5 Marks)*

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{N_i}}{\sum \frac{b_i c_i}{N_i}}$$

The following STATA output is the Mantel-Haenszel adjusted odds ratio model generated from this study's data:

```
. cc hhv8 syphilis, by(inj_drug)

Injection Drug U |      OR    [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------
              NO |   1.050699   .1171107   4.434032   1.699851
             Yes |    2.84492    .863028   8.689501   1.833333
-----------------+-------------------------------------------
           Crude |   2.348233   .9590632   5.183353
     M-H combined |   1.981702   .8880265   4.422326
-----------------+-------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     1.20  Pr>chi2 = 0.2733

                Test that combined OR = 1:
                        Mantel-Haenszel chi2(1) =     2.76
                                         Pr>chi2 =   0.0967
```

(a) Would you conclude that injection drug use is a confounder in the relationship between human herpesvirus 8 (HHV-8) infection and Syphilis? Why or why not? *(5 Marks)*

(b) One interpretation of the data is that there is effect modification by injection drug use status on the association between syphilis serostatus and HHV-8 serostatus (i.e. that there is truly an interaction between injection drug use and sexual activity in the transmission of HHV-8). However, it is not clear as to what the biologic mechanism would be for how injection drug use modifies someone's susceptibility to become infected with HHV-8 via sexual transmission. Assuming that there is no selection bias in the study and that any misclassification of any of the stated variables is non-differential (if it exists at all); please describe two additional interpretations/explanations of these data/model *(i.e. state two explanations as to why the apparent presence of effect modification is false)*. *(6 Marks)*

5)

(a) In a study of the determinants of serum glucose levels in a population of women, four variables were found to be significantly associated with impaired glucose tolerance (serum glucose above 7.8mmol/L) in multiple logistic regression analysis. The following results were obtained when all four variables were included in the model with impaired glucose tolerance as the outcome variable:

| Variable | Logistic regression coefficient | Standard Error |
|---|---|---|
| Body mass index (kg/m$^2$) | 0.077 | 0.032 |
| Waist/hip ratio (ratio increase of 1 unit) | 3.625 | 1.670 |
| Diabetogenic drugs (yes=1; no=0) | 0.599 | 0.302 |
| Regular exercise (yes=0; no=1) | 1.664 | 0.740 |

(i) Carefully study and state in words the meaning (interpretation) of the logistic regression coefficients for waist/hip ratio (3.625) and regular exercise (1.664). *(6 Marks)*

(ii) Calculate the odds ratio and the corresponding 95% confidence interval for body mass index and Diabetogenic drugs? { 95%CI (logOR) = logOR± [1.96 x SE(logOR)]} *(6 Marks)*

(b) Road traffic injuries have become one of the most important non-infectious causes of morbidity and mortality in Kenya. You are designing a proposal with the main aim of finding out the most important risk factors of road traffic injuries in drivers in Nairobi. You decide to recruit motor vehicle drivers coming to accident and emergency departments of 3 major public hospitals in Nairobi.

(i) Outline 5 major factors you might consider for inclusion in the logistic regression model of risk factors of road traffic injuries. *(5 Marks)*

(ii) Which factor would you consider as your main predictor variable and why? *(2 Marks)*

(iii) Assuming that you are carrying out a case-control study, what will be your control group and why? *(2 Marks)*

(iv) Outline 2 methods you would use for selecting predictors for inclusion into your logistic regression model. *(4 Marks)*

6) Your colleague undertakes a case-control study to determine the association between "statins" (drugs that are conventionally used to treat high cholesterol) and breast cancer. She finds that when she stratifies the data into premenopausal and postmenopausal age-groups (at 50 years of age), she cannot interpret the resulting stratified odds ratios. She brings the data to you for interpretation. You enter the data into STATA and generate the following model:

```
cc cancer statins, by(agegroup)

     Age-Group |      OR       [95% Conf. Interval]    M-H Weight
---------------+-------------------------------------------------
 Premenopausal |  1.539906     .5121943    4.629709      2.547847
 Postmenopausal |  .4587156     .2257189    .9322212        11.995
---------------+-------------------------------------------------
         Crude |  .6519244     .3672991    1.15711
   M-H combined |  .6481357     .364752     1.151687
---------------------------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     3.287  Pr>chi2 = 0.0698
```

a) What is your interpretation of this data? *(7 Marks)*

b) What are your conclusions regarding the relationship between statins and breast cancer. *(3 Marks)*

You peruse through the results section of your colleague's thesis and you find the following table about other potential risk factors of breast cancer (table below):

| Factor | Odds ratio for development of breast cancer |
|---|---|
| Age at menarche | |
| <12 years | 1.0 (reference) |
| 12 years | 0.98 |
| 13 years | 0.98 |
| Number of live births | |
| 0 | 1.0 (reference) |
| 1 | 0.95 |
| 2-3 | 0.85 |
| 4 | 0.75 |
| Education | |
| <High school | 1.0 (reference) |
| High school | 1.92 |
| College | 3.48 |

Your colleague had made some statements in her thesis regarding these potentials risk factors for breast cancer. Indicate whether each of the following statements in her thesis is **true** or **false** and justify your answer.

c) Based on this table, age at menarche is unlikely to be a meaningful confounder of the relationship between "statins" and breast cancer. *(5 Marks)*

d) Number of live births is unlikely to be a meaningful confounder of the relationship between "statins" and breast cancer. *(5 Marks)*

e) Education is definitely a confounder of the relationship between "statins" and breast cancer. *(5 Marks)*