MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

Friday, July 31, 2015: 9:00 AM - 3:00PM

Room: BCBS Auditorium

INSTRUCTIONS:

- This is an **OPEN BOOK** examination.
- Submit answers to **exactly 3** out of 4 questions. If you submit answers to more than 3 questions, then only questions 1-3 will be counted.
- Put answers to different questions on **separate sets of paper**. Write on **one side** of the sheet only, within the marked box.
- Put your code letter, **not your name**, on each page, and do that just before turning in your exam. Your code is highly confidential information. Sharing it with anyone is a violation of the Honor Code.
- Return the examination (questions and your answers) with a **signed Honor Pledge form**, separate from your answers.
- You are required to answer **only what is asked** in the questions and not to tell all you know about the topics.

1. An unmatched case-control study was conducted to investigate whether there is an association between illicit drug use and sudden infant death syndrome (SIDS). The cases were 50 mothers of SIDS victims. The controls were 50 women in the same age range as the cases. Each case was asked about illicit drug use in the month before her infant's death and each control was asked the same question about the month prior to the interview. One case and eight controls admitted having used illicit drugs in the time period of interest

   (a) Discuss the appropriateness of the choice of controls.

   (b) Test whether there is an association between illicit drug use and SIDS.

   (c) Calculate a measure of the strength of the association between illicit drug use and SIDS. (You do not need to provide an associated confidence interval.)

   (d) Estimate the proportion of women in this age range who use illicit drugs and give an associated 95% confidence interval.

   Suppose that instead of being an unmatched case-control study this had been an individually matched case-control study, with each case matched to a control woman of the same age who had an infant the same age as the case's infant was at the time of death. Assume the results were as previously.

   (e) Discuss the appropriateness of the choice of controls here relative to that of the unmatched case-control study.

   (f) Is there sufficient information available to estimate a measure of the strength of the association between illicit drug use and SIDS? If so, provide an estimate. If not, state what additional information is needed.

   The investigators were concerned about potential problems in their case-control study and so decided to conduct a large cohort study. They enrolled all consenting women who gave birth in their city over an extended period, followed them and their infants for one year and determined how many infants succumbed to SIDS (the age limit for classifying a death as SIDS is one year). They obtained urine samples from the women at the time of enrollment and tested these for illicit drugs. The table below cross-classifies women in terms of illicit drug use and whether their infant became a SIDS victim.

	SIDS case	Non-case
Illicit drugs detected	7	1008
Drugs not detected	21	3720

(g) Estimate the proportion of women using illicit drugs around the time of giving birth and provide an associated 95% confidence internval.

(h) Calculate a measure of the strength of the association between illicit drug use and SIDS and provide an associated 95% confidence interval.

(i) Give a possible explanation for the difference between the results in parts (c) and (h).

Points: (a) 2, (b) 5, (c) 1, (d) 5, (e) 2, (f) 2, (g) 3, (h) 3, (i) 2.

2. Diabetes mellitus is a group of metabolic diseases characterized by prolonged high blood glucose (blood sugar). In a non-diabetic person, the pancreas responds to elevated blood sugar by secreting insulin, a hormone that promotes absorption of blood glucose by the body. Diabetes occurs when this mechanism fails to adequately reduce blood glucose. There are two main types of diabetes. Type I diabetes is usually diagnosed in children and young adults and occurs when the pancreas is unable to produce the necessary insulin for glucose absorption. Type II diabetes begins, not with failed insulin production, but with insulin resistance, a failure of the cells to adequately respond to the insulin secreted by the pancreas. Type I diabetes has no known cause, and the primary causes of Type II diabetes are thought to be excessive body weight and low physical activity. Type II diabetes used to be found almost exclusively in adults, but with childhood obesity on the rise, it is becoming a health concern for children.

   A researcher at UNC has recently found that infant feeding practices - specifically, being breastfed as an infant - may be protective against the development of Type I and Type II diabetes. She has access to a large database of children age 10 to 13 years old with Type I and Type II diabetes, and for her next project she wishes to test the hypothesis that children with diabetes who were breastfed more as infants have better blood sugar control than those who were breastfed less. Blood sugar control (or glycemic control) is measured using hemoglobin A1c (HbA1c). HbA1c is reported as a percent, and higher values indicate higher blood sugar (i.e. poorer glycemic control) for the past 2–3 months. Breastfeeding is recorded as the duration (in months) of breastfeeding as an infant. The main study questions are whether longer breastfeeding duration is associated with better glycemic control and whether the relationship between breastfeeding duration and glycemic control is the same for Type I and Type II diabetics.

   (a) Provide a linear model that allows researchers to answer the questions outlined above. Clearly specify the dimensions and elements of $\mathbf{y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$. Your model should include an intercept.

(b) For the model in part (a), provide the $\mathbf{C}$ and $\boldsymbol{\theta}_0$ matrices for testing the following hypotheses as $H_0 : \mathbf{C}\boldsymbol{\beta} = \theta_0$

i. Type of diabetes is unrelated to HbA1c.

ii. Number of months of breastfeeding is unrelated to HbA1c.

iii. In children who were breastfed for 6 months, HbA1c is 0.5 higher in children with Type II than Type I diabetes.

iv. The effect of diabetes type does not depend on the number of months of breastfeeding.

v. Children with Type I diabetes who were not breastfed have HbA1c that is no different from children with Type I diabetes who were breastfed for 3 months.

(c) Finding that HbA1c is lower in the Type II participants who were breastfed, the researcher wants to then determine whether the duration of breastfeeding is associated with HbA1c in the Type II participants. The design matrix contains only an intercept and the duration of breastfeeding. Here are some details that may be helpful:

$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 67 & 105 \\ 105 & 1113 \end{bmatrix}$

$\mathbf{X}'\mathbf{y} = 766.2$

Corrected Total Sum of Squares: 278.385

Sum of Squares for Error: 277.999

i. Compute the least square estimates of the model parameters.

ii. Test whether there is a linear association between duration of breastfeeding and HbA1c.

iii. Compute a 95% prediction interval for someone who was breastfed for 6 months.

Points: (a) 5, (b) 8, (c,i) 5 (c,ii) 4 (c,iii) 3.

3. Investigators are interested in testing hypotheses related to prevalence of cognitive impairment in an elderly population. To that end, 6456 participants in a biracial cohort of men and women aged 65–89 years underwent neurocognitive testing and were diagnosed on a continuum in order of increasing severity as cognitively normal, MCI (mild cognitive impairment) or dementia. A participant is considered to be cognitively impaired if diagnosed with MCI or dementia. Data was also collected on race (black/white), gender (male/female) and age in years.

   The following variables were created in a dataset called COHORT:
   DIAGNOSIS = N (normal), M (mild cognitive impairment), D (dementia);
   RACE = B (black), W (white);
   GENDER = F (female), M (male);
   AGE (years);
   COGIMP = 1 if DIAGNOSIS is M or D and COGIMP = 0 otherwise;
   DEMENTIA = 1 if DIAGNOSIS is D and DEMENTIA = 0 otherwise;
   MCI = 1 if DIAGNOSIS is M and MCI = 0 otherwise.

   (a) Using the given variable names, write down the algebraic expression for a model of the prevalence of cognitive impairment as a function of race, gender and age and assuming main effects only. Provide interpretation of the parameters.

   (b) Modify the model to test the hypothesis that the effect of race on prevalence of cognitive impairment differs between men and women. Provide interpretation of the relevant parameter.

   It turns out that the prevalence of dementia is of interest as well. A logistic regression model is run in SAS using the code provided. Use the provided output to answer the following questions.

   (c) Write down the algebraic expression for this model.

   (d) What is the probability of dementia for an black male who is 85 years old?

(e) Calculate appropriate measures of association, with 95% confidence intervals, for the effects of gender, race and 5 year increase in age on the prevalence of dementia. Summarize the results and conclusions of your analyses in 2–3 sentences.

(f) Now suppose that the investigators are interested in the prevalence of MCI. Can the same methods used to model dementia be used to model MCI? If so, describe (and justify) how it would be done. If not, support your answer.

Points: (a) 5, (b) 3, (c) 4, (d) 3, (e) 5, (f) 5.

```
proc means data=COHORT n mean std ndec = 1;
  class RACE GENDER DEMENTIA;
  var AGE;
proc logistic data=COHORT;
  class DEMENTIA RACE GENDER / param = ref;
  model DEMENTIA (event = ’1’) = RACE GENDER AGE;
```

```
The MEANS Procedure
Analysis Variable : AGE
RACE    SEX     DEMENTIA        N Obs     N      Mean     Std Dev
B       F       0                 938    938     74.3     5.0
                1                  95     95     79.4     5.5
        M       0                 449    449     74.1     5.0
                1                  48     48     76.9     5.2
W       F       0                2667   2667     75.2     5.2
                1                 100    100     79.0     5.2
        M       0                2060   2060     75.6     5.1
                1                  99     99     80.1     5.2
```

```
The LOGISTIC Procedure
Data Set                      WORK.COHORT
Response Variable             DEMENTIA
Number of Response Levels     2
Model                         binary logit
Number of Observations Used   6456

Response Profile
Ordered DEMENTIA        Total
Value                   Frequency
1       0               6114
2       1                342
Probability modeled is DEMENTIA=1.

Class Level Information
Class   Value   Design Variables
RACE    B       1
        W       0
GENDER  F       1
        M       0
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics
Criterion       Intercept       Intercept
                Only            and Covariates
```

```
AIC             2677.115        2420.768
SC              2683.888        2447.859
-2 Log L        2675.115        2412.768

Testing Global Null Hypothesis: BETA=0
Test                    Chi-Square      DF      Pr > ChiSq
Likelihood Ratio        262.3469        3       <.0001
Score                   279.6267        3       <.0001
Wald                    248.0622        3       <.0001

Type 3 Analysis of Effects
Effect  DF      Wald
                Chi-Square      Pr > ChiSq
RACE    1       83.6872         <.0001
GENDER  1       2.0511          0.1521
AGE     1       191.8514        <.0001

Analysis of Maximum Likelihood Estimates
Parameter     DF      Estimate    Standard   Wald        Pr > ChiSq
                                  Error      Chi-Square
Intercept     1      -14.5105     0.8501    291.3478     <.0001
RACE    B     1        1.0890     0.1190     83.6872     <.0001
GENDER  F     1       -0.1670     0.1166      2.0511     0.1521
AGE           1        0.1476     0.0107    191.8514     <.0001
```

4. An experiment was conducted to study the effect of a new drug on weight gain in rats. Twenty rats were randomized to four doses, five rats to each dose level. The dose levels were 0, 1, 2 and 3 mg per day. The response is weight gain in grams over a one-month period. Four variables were defined as follows: d_i is 1 for dose i and $d_i = 0$ otherwise, $i = 0, 1, 2, 3$. The data and computer output are provided at the end.

   (a) The first linear model considered for the mean response was

   $$\alpha_0 d_0 + \alpha_1 d_1 + \alpha_2 d_2 + \alpha_3 d_3.$$

   Interpret the estimate of α_3 and its estimated standard error. Interpret also the corresponding p-value and comment on its relevance to this study.

   (b) Consider the following linear model for the mean response

   $$\mu + \gamma_1 d_1 + \gamma_2 d_2 + \gamma_3 d_3.$$

   Interpret γ_1. Estimate γ_1 and its standard error.

   (c) Test the null hypothesis that the four doses have the same mean weight gain, without assuming any specific dose-response relationship. Report the test statistic, its null distribution, the p-value and interpret the results.

   (d) Next, the following linear model for the mean was considered,

   $$\beta_0 + \beta_1 \text{dose}.$$

   Interpret the parameter β_1. Interpret also the corresponding p-value and comment on its relevance to this study.

   (e) Using the replicates (five rats at each dose), test the linearity assumption in the last part. Report the test statistic, its null distribution, the p-value and interpret the results.

(f) Provide a very brief summary of the above findings in a non-technical language for a medical journal.

(g) The person who conducted the experiment was initially out of town. But when she came back, further discussions revealed that there were only four rats in the experiment, one rat at each dose. At each dose, the data given are monthly weight gains (in grams) over a five-month period. Comment on whether and how this affects the results obtained in the previous parts.

Points: (a) 3, (b) 3, (c) 5, (d) 3, (e) 5 , (f) 3, (g) 3.

Data and computer output:

Dose	Weight gain	sum	sum of
(mg/day)	(g)		squares
0	96 97 98 107 113	511	52447
1	116 116 121 124 124	601	72305
2	128 132 144 146 149	699	98061
3	155 157 158 163 174	807	130483
	Total	2618	353296

```
proc reg data = A;
  model y = d0 d1 d2 d3 / noint;   * Note: y = weight gain (g);

Dependent Variable: y
Number of Observations Used          20

NOTE: No intercept in model. R-Square is redefined.

                              Sum of           Mean
Source             DF         Squares          Square    F Value    Pr > F

Model               4          352434           88109    1636.19    <.0001
Error              16      861.60000       53.85000
Uncorrected Total  20          353296


Root MSE              7.33826    R-Square     0.9976
Dependent Mean      130.90000    Adj R-Sq     0.9970
Coeff Var             5.60600

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

d0            1      102.20000        3.28177      31.14      <.0001
d1            1      120.20000        3.28177      36.63      <.0001
d2            1      139.80000        3.28177      42.60      <.0001
d3            1      161.40000        3.28177      49.18      <.0001
```

```
proc reg data = A;
  model y = dose;

Dependent Variable: y
Number of Observations Used          20

                             Sum of           Mean
Source              DF      Squares         Square    F Value    Pr > F

Model                1   9721.96000     9721.96000     199.35    <.0001
Error               18    877.84000       48.76889
Corrected Total     19        10600

Root MSE              6.98347    R-Square     0.9172
Dependent Mean      130.90000    Adj R-Sq     0.9126
Coeff Var             5.33497

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1      101.32000        2.61298      38.78      <.0001
dose          1       19.72000        1.39669      14.12      <.0001
```