MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

Monday, August 13, 2012: 9:00 AM - 3:00PM

Room: 1305

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 the answers to different questions on **separate sets of paper**. Write on **one side** of the sheet only.
- Put your code letter, **not your name**, in the upper right corner of each page.
- Return the examination 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. “Maximal heart rate ($\mathrm{HR}_{\mathrm{max}}$) is one of the most commonly used values in clinical medicine and physiology. For example, a straight percentage of $\mathrm{HR}_{\mathrm{max}}$ or a fixed percentage of heart rate reserve ($\mathrm{HR}_{\mathrm{max}}$ − heart rate at rest) is used as a basis for prescribing exercise intensity in both rehabilitation and disease prevention programs. Moreover, in some clinical settings, exercise testing is terminated when subjects reach an arbitrary percentage of their age-predicted maximal heart rate (e.g., 85% of $\mathrm{HR}_{\mathrm{max}}$). Maximal heart rate also is widely used as a criterion for achieving peak exertion in the determination of maximal aerobic capacity.” (Tanaka H, Monahan KD, Seals DR. Age-predicted maximal heart rate revisited. *J Am Coll Cardiol.* 2001 Jan;37(1):153-6.)

   Predicted maximum heart rate is also used by recreational athletes using heart rate monitors to guide their sub-maximal training, including trying to prevent over-exertion on “recovery” days. Determining maximum heart rate accurately is not easy, so a simple prediction equation is potentially useful. The best-known prediction equation for maximum heart rate is (220 − age) beats per minute (bpm).

   An experiment was conducted to ascertain the validity of the (220 − age) prediction equation. Healthy volunteers between the ages of 18 and 80 were recruited and had their maximum heart rate determined in an exercise physiology laboratory. Let N denote the sample size and A_i and H_i the age and $\mathrm{HR}_{\mathrm{max}}$, respectively, of volunteer i. Here are some summary statistics:

$$N = 40, \qquad \sum_i A_i = 1880, \qquad \sum_i A_i^2 = 101,374$$

$$\sum_i H_i = 7040, \qquad \sum_i H_i^2 = 1,249,420, \qquad \sum_i A_i \cdot H_i = 322,284.$$

   Assume that $\hat{\sigma}^2 = s_{y\cdot x}^2 = 123.74$. For any statistical tests requested below, conduct a two-sided test using $\alpha = 0.05$.

   (a) Write a linear model relating maximum heart rate to age. Explain the meaning of the parameters in your model.

   (b) Assuming the formula (220 − age) is correct, state the values of the parameters in your linear model and draw a rough sketch of the model, indicating how the parameters apply to the sketch.

   (c) Using the summary statistics provided, fit the model to the data.

   (d) Test whether the data are consistent with the age-related change in $\mathrm{HR}_{\mathrm{max}}$ implied by the formula (220 − age).

   (e) A 50-year-old was tested and found to have a $\mathrm{HR}_{\mathrm{max}}$ of 160. Test whether this $\mathrm{HR}_{\mathrm{max}}$ is significantly different from what would be expected for a 50-year-old based on the data from the experiment.

(f) The exercise physiologists also recorded the gender of the volunteers. Describe how to test whether the age-related change in $\mathrm{HR}_{\mathrm{max}}$ varies by gender. Your description should include information about any modifications needed to the model in part (a).

(g) After the model was fit, it was discovered that the data from an additional volunteer had been omitted.

i. If $A_{41} = 47$ and $H_{41} = 196$, calculate how your parameter estimates in part (c) would change.

ii. If $A_{41} = 70$ and $H_{41} = 161$, describe the likely effect on the parameter estimates in part (c). (Do not calculate the actual changes.)

iii. If $A_{41} = 18$ and $H_{41} = 170$, describe the directions in which the parameter estimates in part (c) would change. (Do not calculate the actual changes.)

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

2. We want to study the relation between income, age, and education. Consider the model $y = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{education} + \epsilon$, where y is income, β_0 is intercept, β_1 and β_2 are the regression coefficients of age and education, respectively, and ϵ is a random variable that follows a normal distribution with mean 0 and standard deviation σ^2. The observed data come from a random sample of 20 individuals.

   (a) A matrix representation of this linear model is

   $$\mathbf{y}_{20\times1} = \mathbf{X}_{20\times3}\boldsymbol{\beta}_{3\times1} + \boldsymbol{\epsilon}_{20\times1},$$

   where $\mathbf{y} = (y_1, ..., y_{20})^T$, and the columns of $\mathbf{X}$ correspond to intercept, age, and education, in that order. Part of the data are listed below.

   $$\mathbf{y} = \begin{bmatrix} 106.4 \\ 125.3 \\ 129 \\ \vdots \\ 158.2 \\ 123.4 \\ 126.8 \end{bmatrix}, \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & 42 & 8 \\ 1 & 45 & 9 \\ 1 & 37 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 50 & 13 \\ 1 & 36 & 7 \\ 1 & 42 & 7 \end{bmatrix}.$$

   Assume that the incomes of two individuals, denoted by y_i and y_j, are independent given $\mathbf{X}$. What are the joint distributions of $\boldsymbol{\epsilon}_{20\times1}$ and $\mathbf{y}_{20\times1}$?

   (b) Consider a simple linear regression model of income versus age: $E(y) = \alpha_0 + \alpha_1\texttt{age}$. Let $\mathbf{x}_{20\times1} = (x_1, ..., x_{20})^T$ be a vector of the 20 observations of age. If $\sum_{i=1}^{20} y_i = 2301.6$, $\sum_{i=1}^{20}(x_iy_i) = 88328.1$, $\sum_{i=1}^{20} x_i = 757$, and $\sum_{i=1}^{20} x_i^2 = 29199$, calculate the least-squares estimates of α_0 and α_1.

   (c) Consider two simple linear regressions of income versus age and income versus education. The R^2 values for these two models are 0.67 and 0.71. Does this mean that age and education together explain more than 100% of the variance of income, and why?

   (d) For the multiple linear regression model

   $$E(y) = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{education},$$

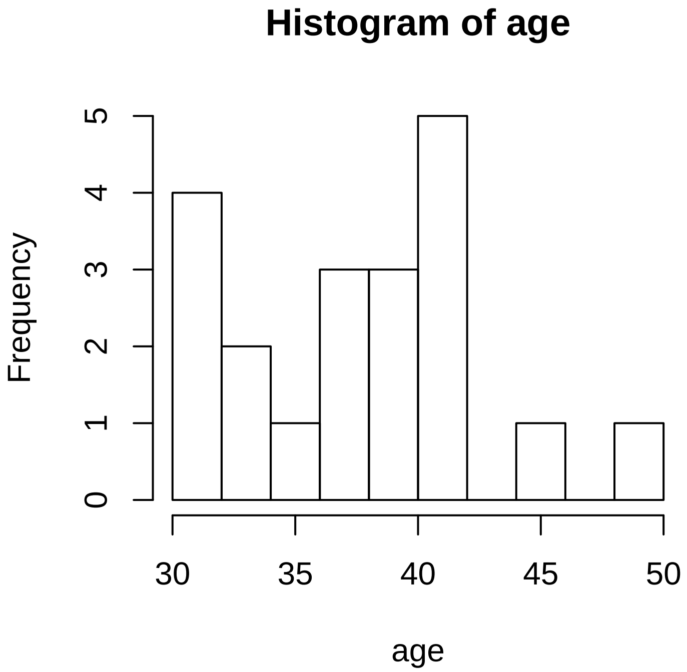
   we have $\hat{\sigma} = 0.5159$, and $(\mathbf{X}^T\mathbf{X})^{-1} =$

```
          intercept    age education
intercept     2.913 -0.098     0.111
age          -0.098  0.005    -0.013
education     0.111 -0.013     0.051
```

Test the hypothesis $\beta_0 = \beta_1 = \beta_2$ using the GLH (Generalized Linear Hypothesis) approach: $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0$. Write out the contrast matrix $\mathbf{C}$, calculate the test statistic and specify its null distribution and the corresponding degrees of freedom. You do not need to calculate the p-value.

(e) What is the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$?

(f) While discussing the linear regression analysis results with your collaborator, he pointed out that the distribution of age, which is shown in the figure below, is not normal. He is concerned that this may violate the normality assumption in your regression analysis. How do you answer him?



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

3. We compare the treatment effects of a certain medication (a particular drug) and counseling on Major Depression Disorder (MDD). To develop personalized treatment, we ask the question whether the efficacy of a treatment is related to age and gender. Denote the treatment efficacy of the i-th individual by y_i, where larger values of y_i indicates more effective treatments. We consider an ANCOVA model $E(y) = \beta_0 + \beta_1\texttt{treatment} + \beta_2\texttt{age} + \beta_3\texttt{gender} + \beta_4\texttt{treatment} \times \texttt{age} + \beta_5\texttt{treatment} \times \texttt{gender}$. Suppose that the sample size is 100, with 25 subjects for each treatment and gender combination. A summary of the values of the covariates is listed in the following table.

Variable	Possible Values	Explanation
treatment	0 or 1	0: medication, 1: counseling
gender	0 or 1	0: female, 1: male
age	15-72	age in years

(a) Write the relation between age and treatment efficacy for each combination of treatment and gender, in terms of β_j's for $j = 0, 1, ..., 5$, in the following format,

Treatment	*Gender*	*Relation*
medication	female	
medication	male	
counseling	female	
counseling	male	

(b) The following regression estimates were obtained:

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       10.446306   0.459411  22.738  < 2e-16
treatment          0.891509   0.606388   1.470 0.144848
age               -0.020759   0.010614  -1.956 0.053453
gender             1.509582   0.302991   4.982 2.85e-06
treatment:age     -0.004731   0.013695  -0.345 0.730534
treatment:gender   1.529518   0.426799   3.584 0.000539
```

Explain the meaning of these regression coefficients. To answer the question "whether we should use different treatments for female versus male, given that they are of the same age", what hypothesis should we test? Specify the test-statistic, its null distribution, and the degrees of freedom.

(c) Are the p-values from the t-tests added-in-order tests or added-at-last tests? Could the model specified in part (b) be simplified, if so, which terms can be removed?

(d) Based on the model in part (b), fill the following ANCOVA table.

```
Source           DF  TypeI SS Mean Sq  F value    Pr(>F)
treatment        __  ______    ______  48.5821  4.321e-10
age              __  ______    ______  0.6181   0.4337367
gender           __  ______    ______  113.8701 < 2.2e-16
treatment*age    __  ______    ______  0.3580   0.5510739
treatment*gender __  ______    13.649  _______  _______
Residuals        __  ______     _____
```

What is the F-statistic, and its degrees of freedom, for testing

$$E(y) = \beta_0+\beta_1\texttt{treatment}+\beta_2\texttt{age}+\beta_3\texttt{gender}+\beta_4\texttt{treatment}\times\texttt{age}+\beta_5\texttt{treatment}\times\texttt{gender}$$

against

$$E(y) = \beta_0 + \beta_1\texttt{treatment} + \beta_2\texttt{age} + \beta_3\texttt{gender}?$$

(e) Let μ_m and μ_c be the mean values of treatment efficacy for medication and counseling when the patient is female at age 33. Test the null hypothesis $H_0 : \mu_m = \mu_c$, write H_0 in terms of β_j's for j=0, 1, ..., 5, the contrast matrix, and the degrees of freedom. Is the conclusion a function of age? Under which situations is the conclusion independent of age?

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

4. Data from a British survey examining smoking habits of twelve year old boys are presented below. Study participants were asked to report their current smoking status and whether or not they have had a cough on the day of the survey. Analyses of these data conducted using SAS are also provided below.

   (a) Two separate logistic regression models are specified in the appendix below. Provide the algebraic expressions for each of these two models as well as interpretations of the parameters.

   (b) Interpret the model with "smoke" as a covariate but *without* an intercept term.

   (c) Give one advantage and one disadvantage for each of these two logistic regression models when using them to model the effect of smoking on the prevalence of coughing.

   (d) For each model, estimate the probability of coughing for occasional smokers.

   (e) Based on your preferred model, what can we conclude about the effect of smoking on the prevalence of coughing? Write 2 or 3 sentences summarizing your results.

Points: (a) 6, (b) 3, (c) 5, (d) 6, (e) 5.

Appendix

Data and analysis results: Cough and cigarette smoking in twelve year old boys.

	Non-smoker	Occasional smoker	Regular smoker	Total
No cough	1037	977	92	2106
Cough	266	395	80	741
Total	1303	1372	172	2847

```
data X1;
        input smoke y n;

        if smoke = 1 then smoke1 = 1; else smoke1 = 0;
        if smoke = 2 then smoke2 = 1; else smoke2 = 0;

        cards;
        0 266 1303
        1 395 1372
        2 80 172
        ;

        proc logistic data=X1;
                title 'Model 1';
                model y/n = smoke;
        run;

        proc logistic data=X1;
                title 'Model 2';
                model y/n = smoke1 smoke2;
        run;

---------------------------------------------------------------------

                             Model 1

                      The LOGISTIC Procedure
                         Model Information
          Data Set                       WORK.X1
          Response Variable (Events)     y
          Response Variable (Trials)     n
          Model                          binary logit
          Optimization Technique         Fisher's scoring

              Number of Observations Read           3
              Number of Observations Used           3
              Sum of Frequencies Read            2847
              Sum of Frequencies Used            2847

                          Response Profile

                 Ordered     Binary          Total
                   Value     Outcome     Frequency

                       1     Event             741
                       2     Nonevent         2106

       Testing Global Null Hypothesis: BETA=0
```

```
      Test                 Chi-Square       DF     Pr > ChiSq

      Likelihood Ratio        58.7323        1         <.0001
      Score                   59.4697        1         <.0001
      Wald                    58.1672        1         <.0001

             Analysis of Maximum Likelihood Estimates

                               Standard          Wald
Parameter    DF     Estimate      Error    Chi-Square    Pr > ChiSq
Intercept     1      -1.3954     0.0654      455.3236        <.0001
smoke         1       0.5415     0.0710       58.1672        <.0001

---------------------------------------------------------------------

                             Model 2

                      The LOGISTIC Procedure
                         Model Information
          Data Set                      WORK.X1
          Response Variable (Events)    y
          Response Variable (Trials)    n
          Model                         binary logit
          Optimization Technique        Fisher's scoring

              Number of Observations Read          3
              Number of Observations Used          3
              Sum of Frequencies Read           2847
              Sum of Frequencies Used           2847

                          Response Profile

                 Ordered     Binary           Total
                   Value     Outcome      Frequency

                       1     Event              741
                       2     Nonevent          2106
```

```
                Testing Global Null Hypothesis: BETA=0

        Test                   Chi-Square       DF     Pr > ChiSq

        Likelihood Ratio          61.0130        2         <.0001
        Score                     64.2467        2         <.0001
        Wald                      61.3785        2         <.0001

                   Analysis of Maximum Likelihood Estimates

                                         Standard          Wald
  Parameter     DF     Estimate          Error      Chi-Square    Pr > ChiSq
  Intercept      1      -1.3604         0.0687        391.8359        <.0001
  smoke1         1       0.4548         0.0910         24.9879        <.0001
  smoke2         1       1.2209         0.1676         53.0575        <.0001
```