Biostatistics 208 Final
Due 3/18/16, 5:00 PM

INSTRUCTIONS

This exam is open-book and open notes. Some questions may require that you refer to the text. You are expected to work entirely on your own. If you need help understanding the questions, please email `stephen.shiboski@ucsf.edu` .

Please email your completed exam to `biostat208hw@gmail.com` by 5:00 PM on Friday March 18, 2016.

QUESTION #1

The dataset `final-hersdata.dta` is a subset of data from the HERS trial.

Focus on the outcome variable `HDL` (high-density lipoprotein in *mg/dL*), and the following three predictors: physical activity `physact`, `age` and body mass index `BMI`. The variable `physact` is coded 1: much less active, 2: somewhat less active (labeled as "less" in the dataset), 3: about as active (labeled as "equally" in the dataset), 4: somewhat more active, 5: much more active.

Question 1.1: (***i***) How would you classify the data type of the predictor `physact`? (***ii***) Is a test for linear trend appropriate for this predictor? Why? (***iii***) Fit a regression model with `HDL` as the outcome and `physact` as a predictor. (***iv***) Perform a test for liner trend as described in section 4.3.5 of the textbook. (Evaluating departure from linear trend is optional.) (***v***) Write 2-4 sentences describing, as you would in a paper, the effect of physical activity on HDL levels.

Question 1.2: (***i***) Assume that it is reasonable to consider age and BMI as possible confounding influences on the relationship between physical activity and HDL. Adjust the relationship in Q1.1 for the variables `age` and `BMI` by including them as predictors in the regression model. (You may assume that these variables can be included as linear terms.) (***ii***) Interpret the results, and answer the following question: Are `age` and `BMI` confounders of the relationship between `HDL` and `physact`? (***iii***) Write 2-4 sentences describing the adjusted association between `HDL` and `physact`.

Question 1.3: (***i***) Based on the model in Q1.2, estimate the mean difference in `HDL` between a person who is "much more" active compared with someone who is "about as" active (adjusting for `age` and `BMI`). (***ii***) Give the 95% confidence interval for this effect

Question 1.4: (***i***) Based on the model in Q1.2, estimate the effect on mean `HDL` associated with a 10-year *decrease* in age (adjusting for `BMI` and `physact`). (***ii***) Give the 95% confidence interval for this effect.

Question 1.5: (***i***) Based on the model in Q1.2, estimate the effect on mean `HDL` of a 5-unit increase in `BMI` (adjusting for `age` and `physact`). (***ii***) Give the 95% confidence interval for this effect.

QUESTION #2

Use the `final-hersdata.dta` dataset again for this question.

Consider models for the relationship between `logHDL` (the (natural) log-transformed value of HDL) and the three predictors used in Question 1: `physact`, `age` and `BMI`.

Question 2.1: (***i***) *Fit* a regression model with `logHDL` as the outcome and `physact` as a predictor adjusting for `age` and `BMI`. (***ii***) Based on the model, determine the percentage increase in mean HDL for a person who is ”much more active” relative to someone who is ”about as active” (adjusting for `age` and `BMI`). (***iii***) Give the 95% confidence interval for this effect.

Question 2.2: (***i***) Based on the model in Q2.1, determine the percentage decrease in mean HDL associated with a 10-year decrease in age (adjusting for `BMI` and `physact`). (***ii***) *Give* the 95% confidence interval for this effect.

Question 2.3: (***i***) Based on the model in Q2.1, determine the percentage decrease in mean HDL associated with a 5-unit increase in `BMI` (adjusting for age and physical activity). (***ii***) Give the 95% confidence interval for this effect.

Question 2.4: How do the effects you found in Q2.2 and Q2.3 compare with the results of Q1.4 and Q1.5?

QUESTION #3

The dataset `bweight.dta` contains information on 40 fictional births. The variables include birth weight in grams (`birthwt`), maternal age (`momage`) and gestational age of the infant at birth (`gestationalage`) in weeks.

Question 3.1: Fit a regression for birth weight as a function of maternal and gestational age. Write a short description of the results

Question 3.2: Check the assumption of normally distributed errors in this dataset; does it appear reasonable?

Question 3.3: If not, perform an analysis that addresses the distributional assumption. Have your conclusions changed?

Question 3.4: Assess if there are any influential points and discuss their impact on the analysis.

QUESTION #4

The dataset `busulfan.dta` contains blood levels on busuflan in *ng/mL* in children undergoing bone marrow transplant. The research question is to summarize the relationship between the probability of engraftment and dose of busulfan. The variables in this dataset are the binary indicator `engraft` (1=engrafted, 0=not engrafted) and busulfan level (`dose`) in *ng/mL*.

Question 4.1: Fit a logistic regression model to look at the relationship described above. *Write* a 1-2 sentence summary of the effect of busulfan level (dose) on engraftment.

Question 4.2: For this question, assume that a linear relationship between the log odds of engraftment to `dose` is adequate. **(*i*)** What is the effect of a 50 *ng/mL* increase in busulfan levels on the odds of engraftment (give estimate with 95% confidence interval)? **(*ii*)** What is the effect of a 100 *ng/mL* increase in busulfan level? (Again, provide an estimate and confidence interval.)

Question 4.3: Using the model fitted in the previous question, calculate the **(*i*)** log odds, **(*ii*)** the odds and **(*iii*)** the probability of engraftment for busulfan levels of 200 *ng/mL*, 450 *ng/mL* and 600 *ng/mL*.

Question 4.4: Use the model fitted in question 4.2 to estimate the *relative risk* and *risk difference* comparing engraftment probability in individuals with bulsulfan level of 600 *ng/mL* with the corresponding probability in individuals with 200 *ng/mL* of bulsulfan. Provide 95% confidence intervals for the two estimates.

QUESTION # 5

The dataset `casecontrol.dta` is from a case-control study conducted on post-transplant diabetes in children who underwent a kidney transplant. The major interest in the study was whether the post-transplant anti-rejection medication increased diabetes risk. The variables in this dataset are `case` which gives case/control status (1=case, 0=control); `meds`, a binary indicator of whether cyclosporin A (CSA) or tacrolimus (FK506) was used as the basis of the anti-rejection regimen (1=FK506, 0=CSA). In addition, the dataset includes each patients's age at transplant `agetx`. As discussed in the last lecture, *logistic regression* is the appropriate method of regression for data from case-control studies, and this method should be used for the questions below.

Question 5.1: Write a sentence summarizing the crude (or unadjusted) effect of medication (comparing patients receiving FK506 to those receiving CSA) on the odds of post-transplant diabetes.

Question 5.2: Write a sentence summarizing the crude effect of age at transplant on the odds of post-transplant diabetes. (You can assume that the relationship between the log outcome odds and age is linear).

Question 5.3: Fit a model including an interaction between age and medication. What do you conclude from the results?

Question 5.4: Use the model fitted in the previous question to estimate the OR evaluating the effects of FK506 use, in patients with an age of (*i*) 5.0, (*ii*) 10.0 and (*iii*) 15.0.

Question 5.5: Obtain the OR for a 5-year increase in age (*i*) among children treated with CSA and (*ii*) among children treated with FK506.

Question 5.6: Write 2-4 sentences summarizing the results above, as you would in a paper.