Biostatistics 208 Final
Due 3/23/18, 5:00 PM

INSTRUCTIONS

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

Please submit your completed exam to the course web site by 5:00 PM on Friday March 23, 2018.

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 linear trend. (Evaluating departure from linear trend is optional.) (***v***) Write 2-4 sentences describing, as you would in a paper, the estimated 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 the results consistent with `age` and `BMI` having a confounding influence on the relationship between `physact` and `HDL`? (***iii***) Write 2-4 sentences describing the adjusted association between `physact` and `HDL`.

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

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. (Note that since this dataset is simulated, some of the values may fall outside realistic ranges.)

Question 2.1: Fit a regression for birth weight as a function of maternal and gestational age. Write a short description of the results

Question 2.2: Check the assumption of normally distributed residuals in this dataset; does it appear reasonable?

Question 2.3: Provide an alternate analysis that addresses violations of assumption of normality assessed in the previous question. Have your conclusions changed?

Question 2.4: Assess if there are any influential points and discuss their impact on the analysis.

QUESTION #3

This question is based on a subset of variables from the Western Collaborative Group Study (WCGS) data. The dataset `final-wcgs.dta` includes the following variables measured at the enrollment visit: `dibpat,` a binary indicator of a type-A behavior (with type-B as the reference category); `smoke`, a binary indicator of reporting smoking at least one cigarette per day; `bmi` (body mass index); `chol`, total serum cholesterol in mg/dl; `age` (years), `ncigs` (number of cigarettes smoked per day), and `npacks`, a categorical indicator based on `ncigs` of number of packs smoked per day (with category levels corresponding to 0, < ½, ½-1, >1 – 2, >2 packs per day). The outcome variable `chd69` is a binary indicator of a coronary heart disease (CHD)-related event occurring during study follow-up.

The goal is to investigate smoking, as measured by the binary indicator `smoke`, as a risk factor for CHD, given the available predictor variables.

Question 3.1: Write a sentence summarizing the crude (or unadjusted) effect of smoking on CHD, using the binary indicator `smoke` of baseline smoking status.

Question 3.2: Fit a logistic or log relative risk regression model to estimate the effect smoking on CHD adjusted for the possible confounding influence of other enrollment variables. Focus again on the `smoke` indicator as the predictor of primary interest, and on providing an estimate that addresses the possible causal effect of any enrollment baseline smoking status on CHD risk. Assume also that any included non-categorical predictors can be adequately modeled by linear terms, and that interactions can be ignored. Also, ignore the alternative measures of smoking. Answer the following questions:

(i) Describe how you selected adjustment variables for your model including a discussion of what you assumed about confounding or mediating roles of the variables considered. Also, identify your chosen measure of the effect of smoking. Note that a conditional estimate (e.g. adjusted regression coefficient), or a marginal estimate (e.g. a marginal odds ratio or relative risk) is acceptable as an effect measure.

(ii) Present the estimated adjusted effect measure with associated 95% confidence intervals.

(iii) Provide a brief paragraph summarizing your findings, and describing issues with and/or limitations of your analysis.

Question 3.3: Fit an additional regression model including any variables adjusted for in the previous analysis, and also including a two-way interaction between the binary indicators `smoke` and `dibpat`.

(i) Summarize evidence for (multiplicative) interaction from your fitted model.

(ii) Present the estimated associations between the `smoke` variable and CHD for both type A and B behavior categories (including 95% confidence intervals).

(iii) Provide a summary of evidence additive interaction between `smoke` and `dibpat`, using the relative excess risk for interaction (*RERI*) measure, and give the accompanying 95% confidence interval and *p*-value. (Note that if you used logistic regression, you may assume that the outcome is sufficiently rare that the odds ratio-based version of *RERI* is acceptable.)

QUESTION #4

This question is based on the Western Collaborative Group Study (WCGS) dataset `final-wcgs.dta` considered in the previous question, and focuses on a subsidiary analysis of the relationship between smoking and CHD risk based on the quantitative variable `ncigs` (number of cigarettes smoked per day) and the categorical indicator `npacks` of number of packs smoked per day.

Question 4.1: Fit a logistic regression model for the relationship between the CHD outcome `chd69` and `ncigs`, including `age` (as a linear term) and behavior pattern (`dibpat`) as adjustment variables.

- *(i)* Summarize evidence for a linear association between the log odds of CHD and `ncigs`.
- *(ii)* Discuss any limitations of the linear model fitted in part (*i*). (Consider the relative frequency of observed values of `ncigs` in your answer.)
- *(iii)* Conduct additional analyses to evaluate the plausibility of a nonlinear association between `ncigs` and the log odds of CHD, and write a brief summary of your findings.

Question 4.2: Fit a logistic regression model for the relationship between the CHD outcome `chd69` and the categorical variable `npacks`, again including `age` and `dibpat` as adjustment variables.

- *(i)* Summarize the overall effect of smoking (as measured by `npacks`) on the odds of CHD using a likelihood ratio test.
- *(ii)* Perform a test for linear trend of the association between and the log odds of CHD and `npacks`, and summarize the results.
- *(iii)* Discuss briefly the limitations/ merits of the analysis conducted in question 4.2 compared to the analysis conducted in question 4.1. You may also suggest alternative analyses.