

BIO 226: Homework 1

Due: At Class 1:30pm on Tuesday, February 10, 2015

Purpose: To review linear regression methods and the use of SAS for these methods.

Instructions:

1. For each question requiring data analysis, support your conclusions by including only the relevant SAS output in your answer.
2. Include your SAS program/output as an Appendix to your solutions.
3. Homework to be turned in on the due date at the beginning of class. Late homework will not be graded unless you make prior arrangements with the Instructor.
4. You may (and are encouraged to) work with other students, but each student should turn in their own write-up of the homework including SAS programming and output that they have generated.

Assignment:

The dataset “hers.txt” available in the “Datasets” folder of the course webpage includes data from the Heart and Estrogen/Progestin Replacement study (HERS), a randomized clinical trial investigating the efficacy of hormone replacement therapy (HT) for secondary prevention of coronary heart disease. For each study subject, in addition to HT, investigators measured a set of known CHD risk factors, including LDL and HDL cholesterol levels, age, bmi, as well as other important subject characteristics, such as statin use. The dataset on the course webpage includes a row of data for each of 2747 subjects enrolled in the study. In each row, the variables are ln(LDL), age, BMI, and statin use (1=yes, 0=no).

1. Provide a plot of ln(LDL) against each of age, BMI, and statin use. Very briefly, provide a qualitative characterization of the associations that you see.
2. Use PROC REG in SAS to fit a regression model to describe each of the three associations plotted in response to question 1. Provide a table showing, for each parameter in the mean part of the model, the estimated parameter, standard error and p-value from a test that the true parameter is equal to zero. Provide a brief quantitative characterization of the two associations. [HINT: The parameters in the mean part of the model in the class notes are the β 's].
3. Use PROC REG in SAS to fit a single model describing the multivariable association between ln(LDL) and age, BMI, and statin use.

- a. Write down a complete algebraic definition of the model being fitted, including assumptions.
 - b. Based on your algebraic definition, show how $\ln(\text{LDL})$ is associated with age, BMI, and statin use.
 - c. Provide a table showing, for each parameter in the mean part of the model, the estimated parameter, standard error and p-value from a test that the true parameter is equal to zero. Provide a quantitative interpretation of the estimates and 95% confidence intervals for the parameters in this model.
4. Based on the results of the analyses conducted in response to questions 1 to 3, summarize the methods, results and interpretation of this study and analysis in a brief structured paragraph in a form that is informative and suitable for an abstract for submission to a conference (which has a 250 word limit). [HINT: It would be good to provide information about the association of LDL with each predictor, not $\ln(\text{LDL})$].
5. Now suppose interest focuses on whether BMI modifies the association between $\ln(\text{LDL})$ and statin use.
 - a. Extend the model in question 3 to address this question, and report the results. Assess whether the data provide strong evidence that BMI modifies the $\ln(\text{LDL})$: statin use association.
 - b. Explain why the main effect of statin use does not represent a meaningful effect in this model.
 - c. Construct a new variable (centered bmi) defined as $\text{bmi} - \text{mean}(\text{bmi})$, where this mean is taken over the entire dataset. Re-fit the model in this question replacing bmi with this centered version, and report the results. Explain why the main effect of statin use now represents a meaningful quantity.