Biostatistical Methods I (BIOS 5710)
Breheny

Assignment 11
Due: Thursday, December 8

1. Consider the `lipids` data on the course webpage, which contains measurements of triglycerides and LDL cholesterol for adult women in the 2005-2006 NHANES sample.

   (a) What is the correlation between triglycerides and LDL cholesterol? Include a 95% confidence interval.

   (b) Suppose a woman's LDL cholesterol levels are 1 SD above the mean. How many SDs above the mean would you expect her triglyceride levels to be?

   (c) Suppose a woman's triglyceride levels are 1 SD above the mean. How many SDs above the mean would you expect her cholesterol levels to be?

   (d) Suppose a woman's LDL cholesterol levels are 50 mg/dL above the mean. How many mg/dL above the mean would you expect her triglyceride levels to be?

   (e) Suppose a woman's triglyceride levels are 100 mg/dL above the mean. How many mg/dL above the mean would you expect her cholesterol levels to be?

2. In one study, the correlation between the educational level of the husbands and wives in a certain town was about 0.5; both averaged 12 years of schooling, with an SD of 3 years.

   (a) Predict the educational level of a woman whose husband has completed 18 years of schooling.

   (b) Predict the educational level of a man whose wife has completed 15 years of schooling.

   (c) From (a) and (b), it would seem that well-educated men marry women who are less well educated than themselves. But the women marry men with even less education. How is this possible?

3. Consider the crossover study of oat bran cereal and cholesterol levels that we have analyzed in previous assignments. Consider the question of predicting what an individual's cholesterol levels will be on the oat bran diet based on his cholesterol levels on the corn flake diet. Fit a regression line to the relationship between the two cholesterol levels.

   (a) What is the variance of individuals' cholesterol levels on the oat bran diet?

   (b) Consider the residuals (prediction errors) of the regression model (you can access these with `fit$residuals`). What is the variance of the residuals?

   (c) The quantity in (a) could be thought of as the total variability of cholesterol levels, and the quantity in (b) could be thought of as the variability in cholesterol levels that cannot be explained by knowing the individual's baseline cholesterol (treating the corn flake diet as baseline). What percent of the total variability can be explained by person-to-person variability, and what percent is due to other factors?

   (d) What is the correlation between individuals' cholesterol levels on the two diets?

   (e) If you square the quantity in (d), how does it relate to the quantity in (c)?

4. Prove the theorem on slide 18 of the regression notes. Hint: if we center x and y, what will happen to the intercept, α?

5. For the `genData` function that we used in Lab 12, show that it does, in fact, produce data with a correlation given by ρ. Specifically, letting X and Y denote the two columns of the matrix that the function returns, show that $\mathrm{E}(X) = \mathrm{E}(Y) = 0$, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, and that $\mathrm{E}(XY) = \rho$.

6. An investigator is exploring whether the expression levels of genes significantly differ between a sample of healthy individuals and a sample of individuals with Type 2 diabetes. He performs a separate t-test comparing the two samples for 5,000 different genes, and uses $\alpha = .05$ as his cutoff. His analysis identifies 411 genes as having different expression levels between the two samples.

   (a) The investigator reasons that because he carried out his t-tests using a type I error rate of 5%, he should expect about 5% of the 411 genes that he discovered to be type I errors. Is this reasoning correct or incorrect? If it is incorrect, what's wrong with it?

   (b) What is the investigator's false discovery rate?

7. To illustrate how multiple comparisons can produce significant associations with no clinical plausibility, a group of Canadian researchers conducted a study of the association between astrological signs and common reasons for hospitalization. They tested 24 such associations.

   (a) How many statistically significant findings (*i.e.*, with $p < 0.05$) would you expect the investigators to discover in their study?

   (b) If we apply the Bonferroni correction, what number should we compare our p-values to in order to maintain a 5% overall probability of making a single type I error?

   (c) The study obtained two "significant" findings: individuals born under Leo had a higher probability of gastrointestinal hemorrhage ($p = 0.0447$), while Sagittarians had a higher probability of humerus fracture ($p = 0.0123$) compared to all other signs combined. Are these findings statistically significant in light of the multiple comparisons that the investigators performed?

8. German researchers carried out a study of two different treatments for heart attacks in a randomized trial involving 421 patients suffering from acute myocardial infarctions. They performed hypothesis tests for 15 different cardiac outcomes.

   (a) In order to keep the overall probability of making a type I error at 5%, what significance level should they test each individual hypothesis at?

   (b) The hypothesis test for the most important outcome, mortality, was $p = .0095$. Is this statistically significant according to the cutoff you defined in part (a)?

   (c) Of the 15 hypotheses, 4 (including the test for mortality mentioned above) were significant at the level $\alpha = .01$. What is the false discovery rate associated with this α level?

   (d) The investigators conclude that there is a statistically significant difference in the mortality rates of the two treatments. Is this statement justified in light of the multiple comparisons that they have made?