Assignment #5 — Due Friday, March 2, 2007, by 4:00 P.M.

Turn in homework in lecture, discussion, or your TA's mailbox. Please indicate the discussion section you expect to attend to pick up this assignment.

311: Tues. 1:00–2:15 **312:** Wed. 2:30–3:45 **313:** Wed. 1:00–2:15

Many problems on this assignment require using the computer. Your turned in solutions should not include all of the computer output and graphs that you will produce. Write your solutions and include only sparingly computer output or graphs when necessary to support a point you are making in response to the problem question.

If a problem asks for a graph, provide it. If the problem asks for you to comment about a graph, you do not need to include the graph in your solution.

1. The file `glue.txt` contains a data set with the results of an experiment on the dry sheer strength (in pounds per square inch) of birch plywood, bonded with 5 different resin glues A, B, C, D, and E. Eight pieces of plywood were tested with each glue type. Let $\mu_A, \ldots, \mu_E$ be the unknown true population mean strengths for the corresponding treatments.

   Analyze the data with a linear model. Summarize the linear model using both the `summary` function in R and the `anova` function.

   (a) The `summary` function provides a p-value for each of several regression parameters. In each case, state the hypothesis that is being tested and provide an interpretation of the regression parameter in terms of the unknowm population means.
   (b) The ANOVA table has a single p-value. State the hypothesis that is being tested here. How does this hypothesis differ form the hypotheses in part (a)?

2. Polynomial regression is a special case of multiple regression where the explanatory variables are integer powers of a single explanatory variable. Typically, a model will only include a parameter with higher degree terms if all of the lower degree terms are included as well, even if these are not significant. For example, if X^3 is included, include X and X^2 too.

   To study the relation between the age of oat plants and their nickel and iron content when grown in sand, the relative absorption of the two metals was measured in terms of the Ni/Fe ratio in individual plants ranging in age from 4 to 73 days. The data are in the file `oat.txt` with columns for `age` and `ratio`. Examine a scatterplot of these data before proceeding with the analysis.

   (a) Fit a linear model, a quadratic model, and a cubic polynomial regression model for `ratio` as a function of `age`. *(R hint: The formula are:*
   `ratio ~ age`, `ratio ~ age + I(age^2)`, *and* `ratio ~ age + I(age^2) + I(age^3)`.
   Which of these three models is best? Explain briefly.
   (b) Which model is prefered by AIC and which is prefered by BIC?
   (c) Repeat (a) using a log transformation of `ratio` as the response variable. Which of these three models is best? Explain briefly.
   (d) Which model from (c) is prefered by AIC and which is prefered by BIC? Is the best model in (a) or the best model in (c) better? Provide reasons.
   (e) Explain why it is inapproriate to compare models in (a) and in (c) on the basis of AIC or BIC.

Work to do, but not turn in.

- Work through the advanced graphics R examples in sections 1.7, pages 31–33.
- Read Chapter 7 of the textbook.