Assignment #4 — Due Friday, February 23, 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. A dairy scientist is interested in calibrating a new technique for determining mastitis cell counts by comparing it with a pre-existing technique. Ten independent samples of milk are used. Each sample is randomly divided into two subsamples, one to be analyzed by each technique. The recorded data consist of a mastitis cell count for each technique for each of the ten milk samples.

   The data are stored on the file `calib.txt` with the old technique measurements in the first column and the new technique measurements in the second column.

   (a) Fit the usual regression model: $y_i = \beta_0 + \beta_1 x_i + e_i$. Is it reasonable to conclude that β_0 is zero?
   (b) Fit the model $y_i = \beta_1 x_i + e_i$. Carry out the test $H_0 \colon \beta_1 = 1$ vs $H_A \colon \beta_1 \neq 1$. Hint: The test statistic for this hypothesis is of the form
   $$T = \frac{\hat{\beta} - 1}{\text{SE}(\hat{\beta})}$$
   You can use the `pt` function in R to find the p-value or consult a t table.
   (c) Based on the available data, how would you compare the new and the old technique?
   (d) For the model in (b), find $\sum r_i$ where r_i is the i^{th} residual, $r_i = y_i - \hat{y}_i$. Note that the sum of residuals need not be 0 in models where there is no intercept.

2. In the file `larch.txt` are data on 26 larch trees. The observation number is in column 1 `id`. The explanatory variables are in columns 2–5 and are the percent content of nitrogen `nitro`, the percent content of phosphorus `phos`, the percent content of potassium `potas`, and the percent content of residual ash `ash`. The percent content of the minerals is determined from dried needles from the tree. The response variable, tree height (`height`) is in column 6. The objective of the study is to relate tree height to the mineral composition of the needles. In other words, find a good model that explains `height` using one or more variables from among `nitro`, `phos`, `potas`, and `ash`. (For this assignment, do not consider interactions or higher order terms.)

   (a) Which model is prefered by AIC?
   (b) Which model is prefered by BIC?
   (c) Are all terms significant at the 5% level for best models by these criteria?
   (d) What unmeasured variables do you think might have the largest effect on tree height in this species?
   (e) Comment on the validity of the best models using mineral composition if there is substantial variation in one or more of of the unmeasured variables.

Work to do, but not turn in.

- Work through the R examples in sections 1.6, pages 25–30.
- Read Chapters 6 and 7 of the textbook.