Stat 215b (Spring 2004): Lab 2

B. M. Bolstad
bolstad@stat.berkeley.edu

Due Feb 26, 2004 by 2:30pm

In this lab we shall investigate another linear model. The data maybe found in the file `bonding.dat` on the class webpage. The focus for this lab will be on hypothesis testing, confidence intervals and model selection. In particular this data relates to data from an experiment on a particular wire bonding process. The goal is to study how bond sheer strength depends on four variables of interest. A total of 30 observations were made. The first column gives the dependent variable Bond Sheer Strength (S). The second through fifth give the independent variables of interest: Force (F), Ultrasonic Power (P), Temperature (T) and Time (D) in that respective order.

First consider fitting a linear model relating strength to the four other variables. That is

$$S_i = \beta_0 + \beta_1 F_i + \beta_2 P_i + \beta_3 T_i + \beta_4 D_i + \epsilon_i \tag{1}$$

Be sure to carefully state your assumptions. Report your parameter estimates, their standard errors etc. Analyze the residuals from your fitted model, comment on how they relate to your model assumptions. How about any outliers?

Give a 95% confidence interval for $2\beta_1 + \beta_2 + 4\beta_3$. Comment on any assumptions you may have made.

Consider the following three hypotheses

- H_0: The model in (1) is correct.
- H_1: $\beta_4 = 0$
- H_2: $\beta_2 - \beta_1 = 0$

Carry out tests of H_0 against H_1 and H_2. Be clear to state what test statistic you are using, any p-value and interpret you results. As always make clear what it is you are assuming. Note that you do not need to do the simultaneous test here, but you should comment on the issue.

Find an appropriate 95% simultaneous confidence intervals for β_4 and $2\beta_1 + \beta_3$. Make sure to clearly explain how you calculated these. How else could you do this (Hint: think about an ellipse)? Explain how you could answer the previous hypothesis testing question in this context.

Now consider fitting a more flexible model including quadratic terms (including the "cross-product" terms. Try to explain the advantages and drawbacks of including more parameters in the model. One method of determining a model is to use backward stepwise deletion (note we are going to restrict ourself to hierarchical models, a later lab will consider model selection in more detail using AIC etc). The procedure is basically as follows.

1. Determine the terms in your model and fit using least squares
2. Look at the P - Values for each term. If all P values are 0.05 or less no variable deletion is required and you may stop, otherwise continue
3. If the term with the largest p value is a quadratic or cross term remove it and go back to the first step, other wise continue
4. If the term with the largest p value is a linear term delete it only if it is not involved in any quadratic or cross product terms. Otherwise look for the term with the next largest p-value and and repeat the previous step. Continue until either a term is deleted in which case return to the first step or that all p-values are 0.05 or less.

Consider the suggested full model and carry out the backwards deletion. Report your final model. Test your final model from this method against your linear model from the first part of this lab, be sure to explain the test statistic you are using.

Consider what might happen if you treated the measurements of the independent variables as factors rather than just numerical values. Explain why you might do this (ie what sort of evidence might you expect to see, in residuals or otherwise). What sort of drawbacks can you imagine for treating the data like this? Note that you are not required to fit the model, however if it helps your understanding go ahead and try it.