Lecture 17

February 17, 2005

Under the condition that our model of a process is true, we've now examined three types of hypotheses we can test:

1. Overall regression: null is that all slopes = 0; no variables are useful predictors. Alternative is that at least one is. F-test.
2. slope = 0, assuming model is correct; alternative is that slope is not zero and therefore there is a relationship between that variable and the response, assuming other variables are included in the model. T-test.
3. some collection of slopes =0; some set of the variables are not useful predictors; alternative is that at least one of the set is useful. F-test.

We started to do an example of (3) above, but got sidetracked because of a "design feature" of R that is important to understand: R will delete missing values from the data set before doing the analysis. And this means that sometimes when variables go in and out, you are comparing different data sets. In theory, the total sum of squares (the total variation measured as the squared sum of the distances between each response and the average response) should be the same regardless of how many variables are included. But in practice, when including a variable with missing values, R (and many other packages, too) will drop those observations from all variables — including the response – and the result is a change in the total sums of squares.

So we'll review the problem last time with using the mussel data set without hte starfish (which are missing values). See the handout that accompanied lecture 15.

There are other tests, too: confidence intervals for the mean y for any given *set* of x values. For example, suppose we fit the model to predict mussel thickness on all of the variables. (Note: this is not the perfect model. We haven't checked whether the assumptiosn hold, whether the relationship is really linear, and some of the variables aren't even needed. Still, we'll proceed to illustrate the process.)

Our fit model is

$$y = -194.631 + 7.233x_1 + 10.225x_2 + 33.564x_3 - 4.080x_4$$

where y is the thickness of the mussel bed, x_1 is the level of the food supply at the site, x_2 was the water temperature, x_3 was the wave activity, ahd x_4 was the level of human use.

We might now ask two questions:

1. There's this coastal site in which the food supply is 2.0, the water temperature is 13.5 degrees, the wave activity is 3 and the human use is 4. What do we predict will be the mussel bed thickness of this site?
2. Among all coastal sites for which the food supply is 2, the water temp is 13.5, the wave activity is 3, and the human use is 4, what is the **mean** mussel bed thickness?

The first question is answered with a prediction interval, the second with a confidence interval. The prediction interval will be wider than the confidence interval, because there is more uncertainty in predicting for an individual site than for the mean of a large number of such sites.

Our first answer is $(-13.3, 97.8)$ and our second is $(21.9, 62.6)$. The first answer is rather unsatisfactory. There is so much variation that we simply can't make good predictions on individuals for this set of values. One hint: the width of the prediction and confidence intervals get narrower when the x values are closer to their means, and so if we had put more "typical" values in for the set of x values, we might be able to make decent predictions. On the other hand, there might just be too much variability, or perhaps our model is not correct.

Graphics

Together, these tools give us a way of selecting variables for a model and then using the model. NOte that we have yet to develop a coherent strategy for selecting variables for the model. Now I'd like to consider a graphical approach that helps guide you in choosing which should stay and which should go. These are Added Variable Plots. They provide a graphic way of understanding the effects of one variable on the response.

Let's consider that we had only temperature to use to predict mussel thickness. We can see in a plot that temperature does have a (semi) linear association with thickness. (plot thickness v. temp). Suppose we now decide to add food into the model. Food also has a (semi) linear relationship with thickness. (Plot thickness vs. food) But notice that food also has a linear(ish) relationship with temperature. (plot food vs. temperature). What this means is that if you already know temperature, you know something about the food supply. In fact, we can fit a regression line:

$$\text{food} = 12.9 - .7\text{temp}$$

. And r-squared is 55%. So if you tell me the temperature, I can come fairly close to predicting the food supply.

What use, then, is adding food into the model if we already have some of its information with temperature?

First, let's consider the model that uses only temperature to predict thickness:

$$y = 231.8 - 10.8\text{temp}$$

We can think of thickness as consisting of two parts. Part 1 is the part of thickness that can be explained by temperature. Part 2 is that part of thickness that was not explained by thickness.

Thickness equals (regression equation with temperature) plus (unexplained part having nothing to do with temperature).

Now if we are to add food supply to the model, we know that some of food supply helps us to explain thickness (we know this— or at least suspect it's true – because we've seen the scatterplot that shows an association between them.) But we also know that food explains some of temperature, too (again because we saw this in the scatterplot.)

So *some* of food goes to explain thickness, but some of it is redundant because it explains temperature. And we already know what temperature has to say about thickness.

So let's examine the part of the data that is temperature could not explain. That's the part we want to work on, after all. One approach we could do is

1. Fit the regresion line of temperature to predict thickness
2. Take the residuals (which contain all of the information that temp could not explain) and see how much of these can be predicted by food.

This *almost* works. Almost because we know that some of food explains thickness, but some explains temp, and we're putting both parts into this equation. Instead, we do this:

1. Fit the regression line of temp to predict thickness. Save the residuals. Call these "thickness without temperature"
2. Fit the regression line of temp to predict food. Save the residuals. These residuals contain everything in food that was not associated with temperature.
3. plot thickness without temperature (on the y axis) against food without temperature.

What we're doing is seeing just how good a job food does in predicting thickness after removing all effects of temperature.

If there's a pattern, well then food must still have some usefulness. Otherwise, none.

properties of the added variable plot

1. the estimated intercept in the added variable plot should be 0.

2. the estimated slope will be the same as the slope in the "full" model, when both temp and food are included.

For example, the full model is

$$y = 135.723 - 5.43\text{temp} + 7.469\text{food}$$

The regression of the thickness without temperature on food without temperature is

$$\text{thickness.notemp} = -2.8 \times 10^{-15} + 7.469\text{food.notemp}$$

.

To summarize, this plot helps us to see intuitively whether adding a variable is useful. In this case, we find that it is.