Lecture 12, Multiple Regression

February 9, 2005

The outline will be as follows:

1. an example of a multiple regression analysis
2. mathematical notation (matrices)
3. equations of estimates, predictions
4. interpretations and assumptions
5. diagnostics

First some loose ends.

1. In a simple regression, when asked "is there a relation between x and y", look at the slope. If the slope is 0, the answer is no. In a random sample, of course, the estimated slope will be non-zero even if the population slope is 0. To see if the estimated slope is so large it would be unreasonable to believe that the population slope is 0, look at the p-value or do a confidence interval. Don't look at r-squared. r-squared tells you if, assuming there is a linear relationship, the relationship is weak or strong.
2. When we talk about a linear model, we mean that we are working under the belief that the relationship between the mean and the predictor is linear. The lowess line shows us what the "local" averages are for different values of x, but the lowess is "model free". It doesn't tell us about a relationship. It does help us pick out major trends, but we should be cautious because it also varies a lot from sample to sample.
3. the great thing about regression is that it provides us a way of predicting mean y values even for x's we did not observe. How? well, if the model really is linear, then we can predict values "in between" observations. Is this right? Well, that's why we try so hard to make sure the model is truly linear.

Off the coast of Austrlia, abalone were potentially being harvested before they reach maturity. This is bad because it means the population might diminish and could go extinct. To know whether or not an abalone can be harvested, therefore, one needs to know its age.

One method has taken advantage of the fact that abalone shells grow about one ring per year after the first 15 years. So the age is approximately number of rings plus 15. However, sometimes they grow more than one a year, sometimes less. So there should be other factors that could be used to help us predict age. This data set is supposed to be useful for predicting age (nrings + 15) based on a variety of variables.

1. rings is the count of rings
2. length is mm longest shell measurement
3. diameter is mm measured perpendicular to length
4. height mm height of abalone with meat in the shell
5. whole is grams weight of the whole abalone
6. shucked is gram weight of just the meat
7. viscera is gram weight of meat after bleeding
8. shell is gram weight of shell after drying
9. infant is 1 if an infant.

1. print the first few rows of the data set. What we do see?
2. histogram of age. What do we see?
3. use the plot(dataframe) command. What do we see?
4. use pairs command on a sub-matrix of the data.
5. plot height and rings (sqrt rings). Interpret
6. plot sqrt(rings) against rest and interpret

does it make sense to say, "among abalone's of same diameter, if length increases....."?

Now there are various diagnostics we can use to try to improve the fit. We want to examine any influential points, for example. And if making inferences

Strategy for datasets with small number of predictors:

1. examine the distribution of each predictor variable. Look for high skew, outliers.

2. examine scatterplot matrix. look for non-linearity
3. try transformations to fix linearity and skew
4. look for pairs of predictors that seem highly correlated. Maybe we don't need both?
5. fit model
6. evaluate fit
7. try again?