Lecture 11

February 7, 2005

cleaning up from last time

Last time we talked about doing inference with our linear model. And there were lots of assumptions that go into this. These assumptions need to be checked to the extent possible.

We're still concerned that the linear model be sound, and a plot of the residuals against the x variable or against the fitted values ($\hat{y}$'s) will help here.

This same plot is a helpful way to make sure the variance of the errors is constant across x. There should be no "fan" shapes in the residual plot.

We're still concerned with influential points, and so an examination of the Cook's distances is still helpful.

Now we're also concerned about normality, and a qqnorm plot of the residuals will help us assess whether or not the residuals are normally distributed.

Even if not, regression is pretty robust to departures from normality. This is because it uses linear estimators (linear functions of the response variable) and therefore the Central Limit Theorem helps out. There are "robust" techniques and techniques for dealing with different distributions. In particular, if your responses are 0's and 1's, or counts, there are other assumptions and approaches we can use.

Independence is harder to check. Sometimes you can see this by plotting the residuals against the order they were collected to see if there are patterns.

Properties of LS Estimates or "Why Least Squares"? Turns out that thsi method of finding estimators produces some nice properties. (Strictly speaking, we did something called ordinary least squares. more on that later.) Al of these properties assume the relation between mean and x is truly linear and variance is constant.

1. estimates are consistent; roughly, as sample size increases, difference between estimates and population values decrease.

2. invariance to scale. The estimates change in a predictable way if units are changed (say from metric to english)
3. Gauss-markov theorem: these estimates have the smallest variance among all linear, unbiased estimates (linear in y).

We'll leave "ordinary" LS behind when we have non-constant variance.

Multiple Regression Introduction

Load the trees data into R.

Fit a linear model using height to predict Volume. What's r-squared? This means that we've explained about 36% of the variation, which means there's more explaining to do.

Now plot the residuals against Diameter. What do you see? What does this mean?

Earlier we explored using a linear model to predict Volume from Diameter. We found that a straight-forward linear model ($E(V|d) = \beta_0 + \beta_1 d$) was not a good fit. We found a better fit using the square of the diameter: $E(V|d) = \beta_0 + \beta_1 d^2$. Define a new variable diam2 that is equal to the square of the diameter. Fit it, plot residuals against diam2. And comment. What's r-squared?

NOte we explained much more variation. A bit alarming is there is a potential "fan" in the residual plot.

Now plot the residuals against height. Note that although we've explained 95% of the variation, there is still a dependence on Height.

So let's resort to mathematical theory. Math tells us that for a cylinder $V = \pi(d/2)^2 h$. This suggests that $\log(V) = \log(\pi/4) + 2d + \log(h)$. So create some new variables: lv is log of the Volume, and lh is log of the height. Now we fit a model that says that log V is a function of both diameter and height: $lv \sim d + lh$.

First, lets examine the diagnostic plots. Do a plot command with the linear model as input. What patterns do you see in the residuals? Why do we plot against the fitted values and not the predictors?

Look at the summary statistics (summary(linearmdl)). What do the p-values tell us? How do we interpret them?

Let's back up and consider our model.

We're now making things more complicated:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{2i} + \ldots \beta_p x_{pi} + \epsilon_i$$

The first subscript is used to keep track of the different predictor variables. There are p of them. the second subscript, the i, keeps track of individual observations.

Our data are now more complicated. Instead of being just pairs (x_i, y_i), they are now entire rows. We think of the data as a matrix in which the rows represent individuals and the columns represent variables.

With two variables, we tried to find the best straight line. With more variables, we're now trying to fit the best hyper-plane. In the two-predictor case: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, our best fit "line" will in fact be a plane.

The geometry is slightly more difficult to visualize (and becomes even more difficult when we add more variables.) But much of the intuition from simple regression carries over.

But think about what a plane looks like. When we now talk about a "one unit increase in x_1", what does this mean? It means that we're moving in a line parallel to the x2 axis. So what we're saying is that x2 is not changing, while we see how the plane responds as x1 changes.

this is key to interpreting the coefficients: the slope of x1 represents the mean response as *all other variables are held constant.*

So let's go back to our estimated tree equation:

$$\log(V) = -4.00 + .145d + 1.2\log(h)$$

What this says is that if we compare trees whose diameter differs by 1 unit, then *as long as we consider trees of the same height*, the log ov the volum inreases by .145. So for example, among all trees that are 70 feet tall, if the diameter differs by 1 foot, the volume is greater, on average, by .145.

What we want to do now is find some techniques for helping us to visualize this.