Lecture 7

January 25, 2005

Regression Diagnostics

Last time we ended by showing the relation of the square of the correlation coefficient to the regression:

$$r^2 = \text{SYY} - \text{RSS/SYY}$$

This term, r^2 is called the coefficient of determination, and is used to assess the "goodness of fit" of a regression, assuming that the relation is in fact linear. r^2 will vary between 0 and 1, and we usually multiply this by 100% and interpret it to mean the percent of variation explained by the regression line. To see what this means, lets examine a series of plots from linear to non-linear.

(plots and diagrams here. illustrate that SYY is total variation (roughly the vertical variation) and RSS is the variation of each point from the line. If there is no such variation, i.e the points fall perfectly on a line, then RSS is 0, SYY - RSS = SYY and the coefficient of determination is 1.

Part of the idea of this is that we can compare fits. All things being equal, the regression with the higher r-squared is "better". All things being equal.

Of course none of this matters if the fit is non-linear. The residuals themselves are worthy of study because they contain the "left over" information. If our regression line has explained things well, there shouldn't be any left-over information.

First, note that $\sum_i e_i = 0$ (show). This means a plot of the residuals against x should be centered on the line $e = 0$. So one of the first things to do after a regresison is plot the residuals and see if this is the case.

Let's look at examples.

1. Galton data. The R commands are in the handout for Lecture 6. You see that the residuals are patternless.

2. Tree data. Let's predict volume using height.

We see that the points are balanced about the $e = 0$ line, but that the scatter about the line increases as height increases. This isn't a big deal now, but will be later.

R-squared for this fit is 35.8%. Not real good. We know of at least one piece of information missing: the diameter. What happens if we plot residuals against diameter?

A very clear pattern develops. This tells us that some of the "left over" information is related to the diameter of the tree. Later, we'll see that this is an indication that our model should take into account diameter as well as height.

Let's fit a model just on diameter now.

3. Wine and mortality The residuals plot looks horrible. And the reason is that this, like the trees, is very non-linear. Unlike the trees, we have no clear-cut model to guide us. So what can we do to fit a better model?

One approach to "fixing" non-linearity is to do a transforming of the x or the y variable. There is a set of transformations called the "power transformations" They are

$-1/y$, $-1/\sqrt{y}$, log(y), $\sqrt{y}$, y.

The last is, of course, no transformation at all. In general, if the residual plot is convex up, move to a higher power of y. Convex down, move to a lower power. Another choice is to use the Box-Cox transform:

$$y(\lambda) = \frac{y^\lambda - 1}{\lambda}$$

for $\lambda \neq 0$ and

$$= log(y)$$

for lambda=0.

If the small values of a variable need to be spread, make λ smaller. If the large values need to be spread, make λ larger.

And experiment. R has a function that tries to estimate λ and provide the value that gives the most linear power. This is one of the "add on" functions. To access it, you have to add a "library" of functions called MASS: library(MASS) help(boxcox)

So lets look at mortality and wine again. The residual plot is convex down – suggests move to a lower power. Currently the power is 1, so the next down would be the square-root transform. Also, the plot itself suggests that we could make it more linear if we spread out the small wine values more. So we could look at the box-cox and try smaller vlaues of lambda (smaller than 1).

The boxcox function tells us to look at a value of λ between -1 and 1. This confirms our choice of the log for wine. We could then repeat this procedure to decide whether we want to also exponentiate mortality.

Let's just do that and fit this equation

$$E(log(y)) = a + blog(x)$$

The result is Log(predicted mortality)= 2.5555-.35556log(wine)

Exponentiate both sides: predicted mortality = exp(2.555-3.5556 log(wine)) = exp(2.55)exp(-3.555*log(wine)) = 12.807 (wine to the 1/3).

Influential Points

Regressions can be affected by exceptional points. It is possible to determine whether an outlier is influential simply by removing it and refitting the model and seeing for yourself. But there are other ways. One useful method is Cook's distance. We'll be able to explore this more when we study multiple regression, but for now, we'll have a simple version.

Let's define $\hat{y}_j$ to be the predicted value at x_j using all of the data. Now suppose we remove one of the observations, say the ith one. We can recompute the line, and figure out what the predicted value is at x_j. We'll call this $\hat{y}_{(i),j}$. Then for a given point that we want to remove, we can compute the distance between where we'd predict the line to be with it and without it for our data-set. Cook's distance is a weighted version of this:

$$D_i = \frac{1}{(RSS/(n-2)} \sum_{i}^{n} (\hat{y}_{(i),j} - \hat{y}_j)^2$$

We can compute this for every point in our data set. If D_i is big for a given point, this means it has a big influence on the line.

R does this fairly easily. Type plot(lmobject) and it gives you a series of 4 diagnostic plots. The first we've seen already (almost): a plot of the residuals against the fitted values (which will have the same shape as the residuals against the x), the next two ignore, and the last is a plot of Cook's distances:

We learn that the 18th point is very influential. Without it, we might have had a very different analysis. You'll notice a pattern in which the points at the edge tend to be more influential. This is not unusual.