Lecture 18

February 28, 2005

Today's topic is diagnostics. Diagnostics means techniques for checking for the extent to which the model is true. I shall remind you again of George Box's motto: "All models are wrong, some models are useful."

In practice this means that we can rarely get a perfectly fitting model, but we need to know how it doesn't fit, and know ways of improving it and for determining whether it's been improved.

For the most part, this is just built on what you've already seen for simple regression.

1. **Linearity**. The response must depend on the covariates in a linear way. This means that the response values sit on a plane (or hyper-plane if we have more than 3 dimensions) determined by the covariates. And this implies that the residuals should be distributed randomly above and below the hyper-plane with no discernible trend. If the truth is that the resopnse is NOT determined by a plane, but instead systematically deviates, then we'll see this in the residuals. The method is to plot the residuals against the fitted values — and see if we can see any trends.

   Why not plot against the predictor values? One reason is that it turns out that the residuals are actually correlated with each other, and that different residuals have different variances. However, the residuals are NOT correlated with the fitted values, so the plot of residuals against fitted values should show no trends. But a plot of residuals against predictors might, even if the linearity assumption is true.

   Another way around this is to instead look at "standardized residuals", which essentially divides all residuals by their standard deviation (and subtracts their expected value) so that they all have the same variance and are not correlated.

2. **constant variance** We assume that the variability in the residuals/errors does not depend on the values of the predictors. If this assumption is true, then the residuals plotted against the fitted values (the predicted values) should show a "band" of uniform (more or less) width. A "fan" shape is a sure sign that this assumption is violated, and the result will be incorrect p-values and biased confidence/prediction intervals.

3. **Normal distribution** The residuals should follow a normal distribution. We can see this by plotting them with a qq plot. The qq plot should be a straight line. This is the least important

assumption because, according to the Central Limit Theorem, if the sample size is sufficiently large, the sampling distribution of our estimates of the coefficients will be approximately normally distributed, and so our p-values and confidence intervals should be approximately correct. But one cannot be certain of just how good this approximation is.

4. **Independence** Observations must be independent of each other. Here are ways that this could be violated. In time-series we take observations in ordered time—for example daily temperatures, or daily ozone readings, or daily DJIA closings. Such things often follow cycles – if it is hotter than average today, we have some information that it will be hotter than average tomorrow too. If the DOW closes higher than it should today, well then depending on your economic theory it will likely close higher tomorrow too or will correct. Either way, today's observation tells you about tomorrow's or next week's. Dependence on time can sometimes be seen by plotting the residuals against the order in which they were collected. Another example: A study was done to compare the social economic status of kids in a school with their score on a standardized performance exam. The data consists of all children in a school for one particular year. It makes sense that within a classroom, the teacher might have an affect on this outcome. And so knowledge of one student's score in one classroom could tell you some information about another.

There are other things we should check that we might describe as sensitivity analysis. We're interesting in seeing how robust our model is to various data points, or slight changes in the data, or slight changes in our assumptions. One of these is "influence". Influence refers to particular data points that, if removed, could affect the data.

Influence

We already saw Cook's distance as a way of measuring influence. The basic idea is that we remove a point, and then we see how much each of our predicted values moved. We then summarize this into a statistic that tries to get at the "typical" movement caused by removing a point. We do this for each point in the data set.

In practice, the computer (or we) don't have to remove each point to compute Cook's distance. In practice, we get Cook's distance by examining something called the leverage.

Each point has a leverage, and the leverage measures how far away that point is from all of the other points. In a simple regression, the leverage of observation x_j is

$$h_{jj} = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

The average leverage is $(p+1)/n$ where p is the number of predictors. Thus, if a point has leverage close to 1, this means that our prediction at that point will be determined almost solely by that point – the rest of the data will have nothing to do with it. Values closer to 0 mean our prediction involves the other data, too.

Note that the above paragraph imples the leverages add up to the number of parameters in the model $(p + 1)$.

The leverage will be largest for those points that are furthest from the average and smallest for those closest to the average.

In multiple regression, the leverage comes from the hat matrix. Let's remind ourselves of what this is.

$$Y = X\beta + E$$

is the matrix version of our model where X is an $n \times (p+1)$ "design matrix". The coefficients are estimated by

$$\hat{\beta} = (X'X)^{-1}X'Y$$

and so the predicted values are

$$\hat{Y} = X(X'X)^{-1}X'Y$$

which we rewrite as

$$\hat{Y} = HY$$

where $H = X(X'X)^{-1}X'$ is the "hat matrix".

The diagonals of the hat matrix are (drum roll please) the leverages. For you math freaks, this is a project matrix, which tells us that we can think of the predicted values as the vector of Y projected onto the space determined by the X's.

And these brings us back to Cook's distance:

$$D_i = \frac{e_i^2}{p}\frac{h_i}{1-h_i}$$

where here e_i is the standardized residual we mentioned before. Here standardized means

$$e_i = \frac{(y_i - \hat{y}_i)}{\hat{\sigma}\sqrt{1-h_i}}$$

Again, we look for large Cook's distances. We don't necessarily act on this information, but it helps us to know whether any one point is having a large influence on our analysis.

Incidentally, this equation can be re-written as

$$D_i = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{p\hat{\sigma}^2}$$

where $\hat{y}_{(i)}$ means the fitted values with observation i removed.

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

Collinearity

A problem sometimes occurs if two predictor variables are linear combinations of each other. The worst case is if you put the same variable in twice. This means that the columns of X, the design matrix, are linearly dependent, which means it is not possible to take the inverse of $X'X$ which means the software crashes. In practice, it is rare to get two columns exactly equal to each other, so instead you end up with near-linear combinations, an dthe result is that you get "unstable" results. This means that you get very large standard errors, and also means that small changes in the data could result in drastically different slopes and also means that simply using a different type of computer might produce different slopes.

This is obviously not good. There are two ways of checking for it. One is to look at the scatterplot matrix of all predictors and be wary of any that appear to be highly correlated with each other. The other is to use added variable plots (which we talked about the week before last) as a means for checking for collinearity. If the added variable plot looks like a vertical line, then these two variables are linear combinations of each other.

Recall: an added variable plot is a three step process. First, fit the response against predictor 1 and save the residuals. Then fit predictor 2 against predictor 1 and save the residuals. Third, plot the first set of residuals against the second.

Quite often there will be some collinearity. One sign of this is if the values of the slopes changes when variables are added and removed. If the variables are linearly independent of each other, then the slopes will not change. If they do, then there is some dependence between them. We can live with moderate amounts.

What to do? Either remove one of the variables from the data set or you can do something called Principal Components Analysis.

Next:

1. Examples of diagnostics
2. Model Selection (AIC/BIC)
3. Validity analysis
4. Bootstrap