Lecture 26

More Regression Diagnostics

In this lecture, we will discuss diagnostics for assessing the fit of a regression on multiple predictor variables, diagnostics for finding problem points for the regression, and techniques for dealing with these problem points.

Partial Residuals

A partial residual plot can help you determine the form of the curvature in your data because just looking at the scatter plot of the response against the predictor is confounded by the other predictors in the model. Partial residual plots are not necessarily clearer than looking at the plot of the residual versus the predictor of interest, but sometimes they are useful and we will discuss an extension below which sometimes improves matters.

Consider the problem of predicting oxidant levels from weather data. The data are at:
http://mypage.iu.edu/∼ehouswor/Fall2004/Math467/L21.html

To determine the form of the nature of the curvature in the humidity data, one would form the partial residual plot as follows:

- Run the Full model regression (using wind speed, temperature, humidity, and possibly insolation.) Store the residuals from this full model. Also, note the coefficient for humidity from this regression model. Call it $\hat{\beta}$.
- Form a column containing the residuals from above plus $\hat{\beta}\times$humidity. These are the partial residuals. Plot the partial residuals against humidity and look to see if you can tell the pattern of the curvature. Sometimes it is useful to include the best line through the data: that would be the plot of $\hat{\beta}\times$humidity versus humidity.

An extension that works better in some cases is to go ahead and include the square of the variable you are interested in exploring in the full model. That helps to capture some of the curvature even if the "right" nature of the curvature is not a square.

- Run the Full model regression (using wind speed, temperature, humidity, and humidity^2.) Store the residuals from this full model. Also, note the coefficients for humidity and humidity^2 from this regression model. Call them $\hat{\beta}_h$ and $\hat{\beta}_{h^2}$.
- Form a column containing the residuals from above plus $\hat{\beta}_h\times$humidity plus $\hat{\beta}_{h^2}\times$humidity^2. Plot these partial residuals against humidity and look to see if you can tell the pattern of the curvature.

We will look at these in class with this set of pollution data.

Cook's Distances

We will also look at the Cook's Distances for the pollution data based on models with and without insolation. There are unusual collections of predictors that become obvious under this diagnostic. We will discuss what the interpretation of the data should be in light of this observation.

Problem Points

What should you do with unusual or problematic data points? That depends on what you want to do with the regression analysis. If you are trying to answer a specific question with your analysis, then you should make sure that your problem points are not changing your fundamental conclusions. If you are simply reporting relationships and have no reason to believe that your problem point is wrong (a data entry error, for instance), then you should probably include your problem point in the analysis. If your problem point is a problem because it's collection of predictor variables is highly unusual, it is reasonable to omit it and say that your regression is valid only in the smaller range and combination of predictor variables.

Exercises for Lecture 26

1. –
2. –