Lecture 24

Multiple Linear Regression III - More Diagnostic Tools

In this lecture, we will discuss additional diagnostics for assessing the fit of a regression on multiple predictor variables and for figuring out how to adjust the model to improve the fit.

Partial Residuals

The scatter plot of the response against a particular predictor variable is confounded by the other predictors in the model, potentially obscuring the underlying pattern. Looking at a plot of the raw residuals against the predictor can be misleading as well since the mean residual is, by definition, zero, which causes many instances of curvature in the data to look quadratic. A partial residuals plots (or, perhaps better, the extension discussed below) can help in seeing the underlying relationship.

To form the partial residuals, you do the following:

- Run the full model regression. Store the residuals from this full model. Also, note the coefficient for the predictor variable of interest from this regression model. Call it $\hat{\beta}$.
- Form a column containing the residuals from above plus $\hat{\beta} \times X$ where X is the predictor of interest. These are the partial residuals. Plot the partial residuals against X 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 $\hat{\beta} \times X$ versus X.

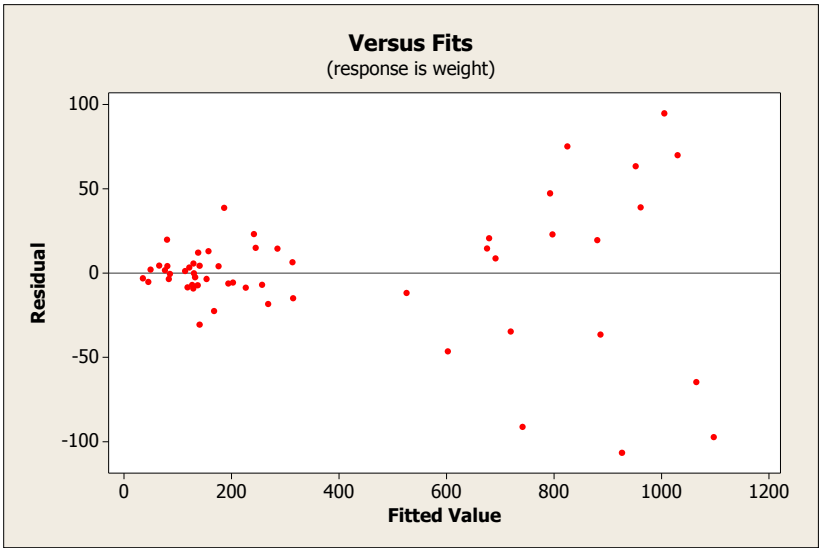
An extension that works better in some cases is to include the square of the variable of interest in the full model. That helps to capture some of the curvature even if the "right" nature of the curvature is not a square.

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

We will look at partial residual plots for predicting perch fish weight from length, height, and width in class. The curvature looks quadratic, and we will consider including a quadratic term in the model.

Plots of Residuals versus Fits

Generally, if there is heterogeneity of variance, then the variance will decrease or increase as Y decreases of increases. Thus, a standard diagnostic tool is a visual plot of the residuals versus the fits. After we include the quadratic term(s), a plot of the residuals versus the fits shows no quadratic pattern (curvature) anymore, but the variance does seem to increase as the fits increase.



What we should have done?

Heterogeneity of variance means we should have transformed the response variable Y. Let's try a log transform of weight for the fish. Does it work better?

What we should have REALLY done?

Weight should be log transformed. The data are fish measurements. Perhaps there is an allometric relationship? Thus, length and/or width and/or height should be log transformed as well. How many of the predictors do we need in the model?

Occam's Razor

Should we include all 3 predictor variables even when some are not significant? That depends on what you want from the regression. If you want a model, then probably not, citing Occam's Razor. If you want to know the effect of one predictor after controlling for the others, then definitely yes. Even if the others are not all significant, to control for them, they have to be in the model.

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 24

1. –

2. –