Print

Residual Analysis and Outliers

We conduct a residual analysis to verify that the conditions for drawing inferences about the coefficients in a linear model have been met. Recall that, if a linear model makes sense, the residuals will have a constant variance, be approximately normally distributed (with a mean of zero), and be independent of one another over time.

In the Impurity example, we've fit a model with three continuous predictors: Temp, Catalyst Conc, and Reaction Time.

In the residual by predicted plot, we see that the residuals are randomly scattered around the center line of zero, with no obvious non-random pattern. The normal quantile plot of the residuals gives us no reason to believe that the errors are not normally distributed.

Because our data are time-ordered, we also look at the residual by row number plot to verify that observations are independent over time. This plot also does not show any obvious patterns, giving us no reason to believe that the model errors are autocorrelated. Note that a formal test for autocorrelation, the Durbin-Watson test, is available. But this discussion is beyond the scope of this course.

One limitation of these residual plots is that the residuals reflect the scale of measurement. The standard deviation of the residuals at different values of the predictors can vary, even if the variances are constant. So, it's difficult to use residuals to determine whether an observation is an outlier, or to assess whether the variance is constant.

An alternative is to use studentized residuals. A studentized residual is calculated by dividing the residual by an estimate of its standard deviation.

The standard deviation for each residual is computed with the observation excluded. For this reason, studentized residuals are sometimes referred to as externally studentized residuals. Studentized residuals are more effective in detecting outliers and in assessing the equal variance assumption. The Studentized Residual by Row Number plot essentially conducts a t test for each residual. Studentized residuals falling outside the red limits are potential outliers.

This plot does not show any obvious violations of the model assumptions. We also don't see any obvious outliers or unusual observations.

Let's take a closer look at the topic of outliers, and introduce some terminology. An observation is considered an outlier if it is extreme relative to the other response values. In contrast, some observations have extremely high or low values of the predictor variable, relative to the other values. These are referred to as high leverage observations. The fact that an observation is an outlier or has high leverage isn't necessarily a problem in regression. But some outliers or high leverage observations exert influence on the fitted regression model, biasing our model estimates.

Take, for example, a simple scenario with one severe outlier. This observation has a much lower Yield value than we would expect, given the other values and Concentration. The regression model for Yield as a function of Concentration is significant, but note that the line of fit appears to be tilted towards the outlier. We can see the effect of this outlier in the residual by predicted plot. The center line of zero does not appear to pass through the points.

For illustration, we exclude this point from the analysis, and fit a new line. Note the change in the slope of the line. The slope is now steeper. An increase in the value of Concentration now results in a larger decrease in Yield.

Also, note the change in the fit statistics. RSquare increased from 0.372 to 0.771, and Root Mean Square Error improved, changing from 1.15 to 0.68. Much more of the variation in Yield is explained by Concentration, and as a result, model predictions will be more precise.

In this example, the one outlier essentially controlled the fit of the model. It's easy to visualize outliers using scatterplots and residual plots. But how do we determine if outliers are influential?

A statistic referred to as Cook's D, or Cook's Distance, helps us identify influential points. Cook's D measures how much the model coefficient estimates would change if an observation were to be removed from the data set. There is one Cook's D value for each observation used to fit the model.

The higher the Cook's D value, the greater the influence. Generally accepted rules of thumb are that Cook's D values above 1.0 indicate influential values, and any values that stick out from the rest might also be influential.

For our simple Yield versus Concentration example, the Cook's D value for the outlier is 1.894, confirming that the observation is, indeed, influential. Returning to our Impurity example, none of the Cook's D values are greater than 1.0. So, we can conclude that no one observation is overly influential on the model.

What do we do if we identify influential observations?

These observations might be valid data points, but this should be confirmed. Sometimes influential observations are extreme values for one or more predictor variables. If this is the case, one solution is to collect more data over the entire region spanned by the regressors. There are also robust statistical methods, which downweight the influence of the outliers, but these methods are beyond the scope of this course.

In the next video, we see how to perform a residual analysis and evaluate outliers in JMP.

Close