Common Regression Problems

Building a regression model is not the last step in the modeling process. You still need to check for common problems that are associated with regression, which include model misspecification, nonconstant variance, correlated error terms, influential observations, and multicollinearity.

The first three of these common regression problems can be identified using residual plots. Remember that residuals are the difference between each observed value of Y and its predicted value. Residual plots display the residual values on the Y axis. What the X axis displays depends on the type of residual plot. To validate the assumptions for general linear models, it is helpful to examine two kinds of residual plots. To start, you can look at a residual plot in which the X axis displays the predicted values. In addition, it can be helpful to look at a residual plot that has the values of a single predictor on the X axis. In the plot on the right, the predictor is represented by X_i. You interpret the two types of residual plots the same way.

Let's look at some examples. These residual plots show four different models fitted to four different sets of data. Your job is to analyze the shape of the residual values. What you ideally expect to see is a random scatter of the residual values above and below the reference line at 0, as in the graph in the top left corner. This indicates that the model is correctly specified and the model assumptions of linearity, equal variances, and independence are valid. However, if you see patterns or trends in the residual values, the assumptions might not be valid and the models might have problems.

Now consider the graph in the top right corner. Does this graph indicate a pattern? Yes, the residual values have a curved shape, which indicates model misspecification. The linearity assumption is being violated. How could you account for the curvature that is present in the data? A different model (for example, polynomial regression) might be needed.

Moving on to the graph in the lower left corner, is all well here? No, notice the increase in variability from left to right, which indicates nonconstant variance. The constant variance assumption is being violated. It can sometimes be corrected by a transformation of the response variable; natural log and square root transformations are very common.

Finally, the graph in the lower right corner indicates a cyclical shape. This pattern indicates correlated error terms; the observations are not independent, so the independence assumption is being violated. Cyclical patterns like this can appear when the predictor variable, X, is a measure of time. The residuals are autocorrelated, meaning correlated over time. More specifically, residuals are more correlated with observations that are nearer in time than with observations that are further away in time. Different modeling tools are needed to model data with correlated errors.

Influential observations might arise in both simple linear regression and multiple linear regression. An influential observation is an observation that is so far away from the rest of the data that it singlehandedly exerts influence on the slope of the regression line. An influential observation might be an indication of erroneous data. Another possibility is that the observation, though valid, might be unusual. Influential observations might affect your regression results, so it is important to identify them using statistics such as RSTUDENT residuals, DFFITS statistics, Cook’s D, and DFBETAS. You learn more about handling influential observations later in this course.

Multicollinearity can arise in multiple linear regression, and it occurs because two or more predictor variables used in the model are highly correlated. It is not a violation of the assumptions, but it can cause significant problems in modeling. You can use the variance inflation factor and other multicollinearity diagnostic statistics to identify multicollinear predictor variables.

Copyright © 2017 SAS Institute Inc., Cary, NC, USA. All rights reserved.