Evaluating Model Fit

With the assumptions checked, let's turn our attention to the fit of our model. A well-fitting regression model results in predicted values close to the observed data values.

To evaluate model fit, you can assess the diagnostic plots created by ODS Graphics in the output of PROC REG. For models that are fit with PROC GLMSELECT, you can use the OUTDESIGN= option to create an input data set for PROC REG. These plots include:

- plots of residuals and studentized residuals versus predicted values
- “residual-fit spread” (or R-F) plots
- plots of the observed values versus the predicted values
- partial regression leverage plots

Let's examine each of these plots in detail. You can use plots of residuals versus the predicted values to visually assess the goodness of fit of the model. You can also use the plot of RSTUDENT residuals versus predicted values to visually assess model fit. The RSTUDENT residual is calculated as the residual divided by the standard error estimated with the current observation deleted. In both of these plots a random scatter of points about a zero reference line indicates a good model fit. You can also use RStudent residuals to identify influential observations. You'll learn more about this later in this lesson.

The Residual-Fit Spread Plot provides a visual summary of the amount of variability accounted for by a model. The plot consists of two panels. The left panel shows the quantile plot of the predicted values minus their mean. The right panel is a quantile plot of the residuals. For comparison purposes, the scales are identical on both plots. This enables you to compare the spread of the fit minus mean plot to the spread of the residual plot. Because it is the spreads of the distributions that are of interest, the fitted values minus their overall mean are graphed.

Using centered data (fit minus mean value) instead of raw values enables you to compare the spread of variables that have different means. If the range of the fit-mean plot is substantially larger than the range of the residual plot, then the model explains most of the variability in the data. Otherwise, the model does not fit the data well and most of the variation in the data remains unexplained.

In order to evaluate the “explanatory power” of the model, you interpret the R-F plot together with the adjusted R-square statistic. Plots of the observed values versus the predicted values also provide a visual tool for examining how close the fitted (or observed) values are to the predicted values. Ideally, for a good model fit, observed and predicted values should go hand in hand.

A partial regression leverage plot is the plot of the residuals for the response variable against the residuals for a selected predictor. This example plots residuals for price against horsepower. The residuals for the response variable are calculated by regressing the response variable on all predictor variables, except for the selected predictor. The residuals for the selected predictor are calculated from a model where the selected predictor is regressed on the remaining predictor variables.

The slope of the linear regression line in the partial regression leverage plot is the regression coefficient for that predictor variable in the full model. (Sall 1990) This plot helps you evaluate whether you have specified the relationship between the response and the predictor variables correctly. A pattern in this plot (for example, curvature) indicates that the predictor-response relationship is not correctly accounted for by the full model. These plots are also helpful in detecting outliers and influential points. (Rawlings, Pantula, and Dickey)

In addition to assessing diagnostic plots in order to evaluate the fit of your model, you can examine model-fitting statistics. These statistics include R-square, Adjusted R-square (the higher the value, the better), Akaike's Information Criterion (AIC) or Corrected Akaike's Information criterion (AICC), Schwarz’s Bayesian criterion (the smaller the value, the better), and Mallows' C_p. The model with a C_p value less than or equal to p, the number of parameters including the intercept, will better fit the data.

If your data includes multiple observations (or replicates) for each value of the combination of the independent variables, then you can use the LACKFIT option in the MODEL statement in PROC REG to perform a lack-of-fit test for the regression model. The test procedure involves partitioning the residual sum of squares into two components, the sum of squares due to pure error, and the sum of squares due to lack of fit (also referred to as bias error). If the model is adequate, then both components estimate the nominal level of error, but if the bias component of the error is much larger than the pure error, then this suggests evidences of significant lack of fit.

Close