Evaluating Model Assumptions

You know that for a linear regression analysis to be valid, four assumptions must be met: linearity in the parameters, normality of the errors, constant variance of the errors, and independence of the errors. The mathematical notation shown here represents the assumptions on the error terms. The error term, ε, is assumed to be independent and identically distributed and to have a normal distribution with a mean of zero and constant variance, represented by σ^2. If your model meets these assumptions, plots of the residuals versus the predicted values are a random scatter about a zero reference line, as expected, and you can trust subsequent results. So let's discuss how to evaluate each of these assumptions.

The independence assumption is that the error terms (ε) are independent of each other. That is, there is no correlation between any one observation in the data set and another. Several methods are available to test this assumption. First, you investigate the source of your data. Knowing how your data is gathered helps you evaluate the assumption of independence. If data is gathered over time, or includes repeated measures on a given subject, consists of any type of clustered data, or results from an overly complex survey design, the error terms might be dependent.

Data collected over time might be serially correlated. Serial correlation in the errors (also known as autocorrelation) means correlation between consecutive errors or errors separated by some other number of periods. To evaluate if the error terms for time series data are correlated, you can plot residuals versus time or another ordering component to examine whether there seems to be any positive or negative autocorrelations. A pattern that is not random suggests a lack of independence. You can also evaluate whether the error terms are correlated for time series data by using the REG procedure to calculate the first-order autocorrelation statistic, also known as the Durbin-Watson *d*-statistic.

You can check the assumption that the error terms are normally distributed in several ways. You can use the SGPLOT procedure, or the ODS graphics output from the REG and UNIVARIATE procedures, to examine a histogram of the residuals with a normal curve overlaid, a normal probability plot of the residuals, or a normal quantile plot of the residuals. You can also use PROC UNIVARIATE to produce the formal tests of normality. Remember that the accuracy of all these tests depends on the sample size.

To evaluate the assumption of constant variance of the error terms, also known as homoscedasticity, you can use the REG procedure and ODS Graphics to produce plots of the residuals versus predicted values and plots of the residuals versus independent variables. Examine the plots for evidence of residuals that grow larger as a function of the predicted value. You should see a random scatter of the residual values above and below the reference line at zero.

You can also perform a test for heteroscedasticity by using the SPEC= option in the MODEL statement in PROC REG. However, the SPEC test is a multiple null hypothesis test, so rejection of the null hypothesis in this case may not be due solely to heteroscedasticity. Thus, this test is usually performed in combination with other tests.

In addition, you can compute the Spearman rank correlation coefficient, which measures the correlation between the size of the ordered predicted values and the absolute value of their associated residuals. (Carroll and Ruppert 1988) The Spearman rank correlation coefficient is available as an option in PROC CORR. If the value of the coefficient is close to zero, it means that there is no correlation between the size of the predicted value and the magnitude of the residual. This indicates that the variances are equal. Positive values mean that the magnitude of the residuals increases as the predicted values increase. This indicates that the variance increases as the mean increases. Negative values indicate that the variance decreases as the mean increases.
