Remediating Violations of the Assumptions

Let's start by examining what alternatives are available if your model does not meet the assumptions for linear regression. You know that when the normality assumption is violated, it affects the test results, for example, tests of significance and confidence intervals of the parameter estimates. One way to remediate this problem is to transform the dependent variable to normalize the distribution. Transformations are often suggested as a means to "normalize" univariate data which can be skewed left or right. For more information about variable transformations, including sample programs, see the Help and Resources section of this course.

You can also use SAS procedures with the appropriate options to accommodate the distribution exhibited by your data. You can fit a generalized linear model using the GENMOD or GLIMMIX procedures with the distribution and link options. If your data violates the constant variance assumption, you can request tests using both the usual covariance matrix and the heteroscedasticity-consistent covariance matrix. In the MODEL statement of PROC REG, you specify the ACOV, HCC, or WHITE option.

You can also transform the dependent variable to stabilize the variance, or use different procedures to model the nonconstant variance. These procedures include:

- PROC GENMOD or PROC GLIMMIX with the appropriate distribution function, using the DIST= option
- PROC MIXED with the GROUP= option, which allows you to define an effect specifying heterogeneity in the covariance structure, the TYPE= option, or the power-of-mean models, and
- PROC SURVEYREG for survey data.

The SURVEYREG procedure fits linear models for survey data and computes regression coefficients and their variance-covariance matrix. For time series data, you can use the procedures provided in SAS/ETS software.

You can also use a weighted least squares regression model when variances are not constant. The ordinary least squares method does not provide minimum variance parameter estimates when variances are not constant. However, if the errors are assumed to be independent, then you can use weighted least squares to reduce the influence of highly variable observations. Information about weighted least squares, including a sample program, is included in the Help and Resources section of this course.

You know that when the independence assumption is violated, it can affect the standard errors of the parameter estimates, confidence intervals, and significance tests for the parameters. Recall that correlated observations, and therefore error terms, often arise from time series data, repeated measures on a given subject, data gathered from a nested design, or from a complex survey design. You can use several modeling tools to account for correlated observations. If the data to be analyzed is time series data, you use SAS/ETS procedures like PROC AUTOREG or PROC ARIMA. For more information on analyzing correlated data, including sample programs, see the Help and Resources section of this course.

To model correlations arising with data that include repeated measures from each subject, you can use PROC MIXED, PROC GENMOD, or PROC GLIMMIX. To model the data gathered from a complex survey design, you should use PROC SURVEYREG.

When a plot of the residuals versus the predicted values exhibits a discernable pattern, you know that this indicates misspecification of the model. This means that the relationship between response variable and predictors does not follow a linear relationship. This could also mean that there is a missing component in the model: an interaction, a polynomial, or another variable entirely. To remediate this problem, you might consider fitting the regression model with polynomial effects or splines.

When the relationship between the dependent variable and one or more predictor variables does not follow a linear relationship, you might consider transforming the predictor variables to obtain the linearity. (Neter, Wasserman, and Kutner 1990) Another option is to transform the predictor variables to obtain the linearity. For applications in physical fields such as pharmacokinetics, chemistry, and biology, the relationship might not be linear in terms of parameters. In that case, you might want to fit a nonlinear regression model using PROC NLIN. For more information on fitting a nonlinear regression model, including sample programs, see the Help and Resources section of this course.

When the parametric form of the relationship is difficult or impossible to define, you might want to fit a local regression model using PROC LOESS. The idea of local regression is that near any chosen value of X, the regression function can be well approximated by low-degree polynomials. For more information on fitting a local regression model, including sample programs, see the Help and Resources section of this course.

Close