Assumptions for General Linear Models

To make sure that the results of linear regression are valid, it is necessary to verify that four assumptions for GLMs are true--the assumptions of linearity in the parameters, normality of the errors, constant variance of the errors, and independence of the errors. Let's look at these assumptions in more detail.

We'll use this graph as a simple example, which shows the linear relationship between a continuous outcome variable (Y) and one predictor variable (X). The four assumptions can be described in three statements. The first statement explains the assumption of linearity in the parameters: the mean of the response variable (Y) is accurately modeled by a function of the predictor variable (X) that is linear in the parameters. In other words, the mean of the response variable is linearly related to the value of the predictor variable. In the graph, a straight line connects the means of the response variable at each value of the predictor variable. Note that linearity in the parameters does not imply linearity in the predictors. Linearity in parameters means that the equation has a linear combination of the predictor variables and the parameters (that is, the betas). However, a model that is linear in parameters might contain predictors that are not linear – in other words, powers or interactions of predictors. You will learn more about these higher-order terms later in this lesson.

The second statement combines two assumptions: normality of the errors and constant variance of the errors. Remember that the errors are the differences between the observed and predicted response values. In this graph, the errors are the differences between the observed values (the dots) and the predicted values (the regression line). The vertical distance between each dot and the corresponding point on the regression line, at each distinct level of the predictor X, is an error. According to this statement, the random error term, ε, has a normal distribution with a mean of zero and a constant variance, σ^2. Another way of saying that the random error term has a mean of zero is that the mean value of ε conditional upon the given X is zero. The "constant variance of the errors" assumption can also be referred to as the assumption of equal spread or, more technically, homoscedasticity. In simpler terms, the variation around the regression line (which is the line of average relationship between Y and X) is the same across the X values; it neither increases nor decreases as X varies.

The third statement states that the error terms are independent at each value of the predictor variable. Remember that these assumptions are presented here in the context of a simple case--a linear relationship between a continuous outcome variable and one predictor variable. The same assumptions are made for more complicated general linear models, such as those with more than one predictor variable, and those where the relationship between the outcome and predictor variables might not be linear (for example, polynomial regression models).

So, why do you need to know about these assumptions? Well, if you want to be sure that the results of linear regression are valid, remember that you must verify that the four assumptions for GLMs are true. And it's important to understand how violations of these assumptions affect your results. Let's look at each assumption individually.

Failing to meet the assumption of linearity in the parameters indicates that the model was misspecified. In this case, the model results are not meaningful. Violation of the normality assumption for the errors does not affect the parameter estimates. However, this violation does affect the tests of significance and the confidence intervals of the parameter estimates. Similarly, violation of the assumptions of constant variance and independence of errors does not affect the parameter estimates. However, violations of these assumptions compromise the standard errors.

Suggested methods for verifying the assumptions include the following: plotting the residuals versus the predicted values and versus the predictor variable, and doing a univariate analysis of the residuals. You'll learn more about these methods later. For additional details about the normality assumption, click the Information button.

Close