Using Link Functions

In generalized linear regression, you determine which link function to use based on the type of response variable and its distribution. For each distribution, there is a link function that can be determined algebraically from the distribution equation. This link is called the canonical link. It is not a requirement to use the canonical link, but in most cases it is the one used for that distribution.

When you have a categorical response variable, you know that you use logistic regression to analyze your data. If the categorical response variable is binary – for example, Yes/No, you typically code the values numerically, such as Yes=1 and No=0. You can then model the probability of 1 (let's call that *p*). One obstacle is that the predicted values from our regression analysis can take on any value in the set of real numbers. However, the probabilities are by definition bounded between 0 and 1.

Another problem is that the relationship between the probability of the response and a predictor variable is usually nonlinear rather than linear. In fact, the relationship often resembles an S-shaped curve (a sigmoidal relationship).

To create a linear model and to constrain the predicted probabilities between 0 and 1, you apply a type of link function called the logit transformation to the probability. Another name for this transformation is the logit link function.

The logit transformation is the natural log of odds, where odds is the ratio of the probability of the outcome to the probability of no outcome. Unlike a probability, the logit is unbounded because transforming the probability to odds removes the upper bound. Taking the natural logarithm of the odds removes the lower bound. The model (also called the logistic regression model) is now linear because the logit is linear in its parameters. Furthermore, the model gives estimated probabilities that are between 0 and 1. The logit transformation also satisfies the assumption in logistic regression that the logit has a linear relationship with the predictor variables.

You can also use a generalized linear model with discrete response variables that are counts. Count data, which consists of non-negative integers, is typically modeled using a Poisson distribution. You can apply the log link function to count data that have nonnegative integer values. The log transformation removes the lower bound and creates a linear model. Click the Information button for more information on calculating canonical links for binary response and count data.

Close