StatNews #43

Regression Models for Count Data
December 2000

One of the main assumptions of linear models such as linear regression and analysis of variance is that the residual errors follow a normal distribution. To meet this assumption when a continuous response variable is skew, a transformation of the response variable can produce errors that are approximately normal. Often, however, the response variable of interest is categorical or discrete, not continuous. In this case, a simple transformation cannot produce normally distributed errors.

A common example is when the response variable is the counted number of occurrences of an event. The distribution of counts is discrete, not continuous, and is limited to non-negative values. There are two problems with applying an ordinary linear regression model to these data. First, many distributions of count data are positively skew with many observations in the data set having a value of 0. The high number of 0's in the data set prevents the transformation of a skew distribution into a normal one. Second, it is quite likely that the regression model will produce negative predicted values, which are theoretically impossible.

An example of a regression model with a count response variable is the prediction of the number of times a person perpetrated domestic violence against his or her partner in the last year based on whether he or she had witnessed domestic violence as a child and who the perpetrator of that violence was. Because many individuals in the sample had not perpetrated violence at all, many observations had a value of 0, and any attempts to transform the data to a normal distribution failed.

An alternative is to use a Poisson regression model or one of its variants. These models have a number of advantages over an ordinary linear regression model, including a skew, discrete distribution and the restriction of predicted values to non-negative numbers. A Poisson model is similar to an ordinary linear regression, with two exceptions. First, it assumes that the errors follow a Poisson, not normal, distribution. Second, rather than modeling Y as a linear function of the regression coefficients, it models the natural log of the response variable, ln(Y), as a linear function of the coefficients.

The Poisson model assumes that the mean and variance of the errors are equal. But, usually in practice the variance of the errors is larger than the mean (although it can also be smaller). When the variance is larger than the mean, there are two extensions of the Poisson model that work well. In the over-dispersed Poisson model, an extra parameter is included which estimates how much larger the variance is than the mean. This parameter estimate is then used to correct for the effects of the larger variance on the p-values. An alternative is a negative binomial model. The negative binomial distribution is a form of the Poisson distribution in which the distribution's parameter is itself considered a random variable. The variation of this parameter can account for a variance of the data that is higher than the mean.

A negative binomial model proved to fit well for the domestic violence data described above. Because the majority of individuals in the data set perpetrated 0 times, but a few individuals perpetrated many times, the variance was over 6 times larger than the mean. Therefore, the negative binomial model was clearly more appropriate than the Poisson.

All three variations of the Poisson regression model are available in many general statistical packages, including SAS, Stata, and S-Plus.

If you would like help implementing or interpreting a model for count data, please contact any of the consultants in the CSCU office.

References:

- Gardner, W., Mulvey, E.P., and Shaw, E.C (1995). "Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models", Psychological Bulletin, 118, 392-404.
- Long, J.S. (1997). Regression Models for Categorical and Limited Dependent Variables, Chapter 8. Thousand Oaks, CA: Sage Publications.

Author: Karen Grace-Martin