## Overdispersion

Poisson regression models assume that the variance is equal to the mean. However, when you model count data, the variances are usually much higher than the mean. This phenomenon is called overdispersion.

When the model for the mean is correct but the variance does not follow what is defined by the Poisson distribution, the maximum likelihood estimates of the model parameters are still consistent, but the standard errors are incorrect. Overdispersion leads to underestimates of the standard errors of the parameter estimates, overestimates of the test statistics, which increase the Type I error rate, and liberal *p*-values. Overdispersion leads to underestimates of the standard errors of the parameter estimates and overestimates of the test statistics, which increase the Type I error rate.

Underdispersion can also occur. In this case, the standard errors are overestimated and the test statistics are underestimated, which increase the Type II error rate.

Overdispersion is not a problem in ordinary least squares (OLS) regression because the normal distribution has a separate parameter, the variance, to describe variability. OLS regression is generally not appropriate for count data because it assumes constant variances, which might lead to incorrect inferences for count data.

So what causes overdispersion?

Poisson regression models assume that the response variable has a Poisson distribution conditional on the values of the predictor variables. If some of the relevant predictor variables are not in the model, or, in other words, if the model is under-specified, then the unexplained variability among the subjects causes greater variation in the response than the Poisson predicts.

If the variance equals the mean when all the relevant predictor variables are controlled for, it exceeds the mean when relevant predictor variables are not controlled for. Also, because there is no random error term in a Poisson regression model, there is no way to account for the extra variability caused by the omitted important predictor variables. The presence of outliers may also result in increased variability and hence cause overdispersion.

Positive correlation between responses in clustered data can also cause overdispersion. Examples of naturally occurring clusters include families, households, litters, and colonies. If your sample includes a large number of observations within a cluster, the positive correlation between the responses might cause overdispersion.

Now that you understand what overdispersion is and what causes it, let's learn how to correct for overdispersion. First, do some quick checks to ensure that your data does not contain errors. Second, recheck that your model includes all the important variables. Re-specify your model if needed.

After you've completed these checks, you can perform one of the following solutions: You can model the overdispersion by using a related distribution for count data that allows the variance to exceed the mean. This is the negative binomial distribution The binomial distribution counts the number of successes in a fixed number of Bernoulli trials. The negative binomial distribution counts the number of Bernoulli trials that are required to obtain a set number of successes. To fit a negative binomial model, you specify the DIST= NEGBIN option in the MODEL statement in PROC GENMOD.

Another way to account for overdispersion is to apply a multiplicative adjustment factor to adjust the standard errors. This fixes the scale parameter at the value of 1 in the estimation procedure. All statistics such as standard errors and likelihood ratio statistics are adjusted appropriately. For binomial and Poisson distributions that have no free scale parameter, this can be used to specify an overdispersed model. To apply a multiplicative adjustment factor, specify the PSCALE or DSCALE option in the MODEL statement in PROC GENMOD.