Modeling Overdispersion by Using the Negative Binomial Distribution

In this demonstration, we'll use PROC GENMOD to fit a negative binomial model to the crab data set in order to account for overdispersion.

The syntax for PROC GENMOD is very similar to what we used to fit the Poisson regression model. In the model statement we use the DIST= option to specify the negative binomial distribution and the LINK= option to specify the log link function. Again, we specify the TYPE3 option, which requests that SAS compute contrasts for each effect specified in the MODEL statement.

```
proc genmod data=mydata.crab;
   class color spine;
   model satellites=width weight color spine
        / dist=negbin link=log type3 ;
   title 'Negative Binomial to Account for Overdispersion';
run;
```

Let's run the program and review the results, beginning with the Criteria for Assessing Goodness of Fit table.

Because the extra dispersion parameter in the negative binomial distribution models the overdispersion, the Pearson Chi-Square / DF value is relatively close to 1, as expected.

Let's compare the fit statistics shown here for the negative binomial model to the fit statistics for the Poisson model. We can see that in all cases the statistic from the negative binomial model is smaller, indicating that the negative binomial model fits the data better.

Now let's look at the Analysis of Maximum Likelihood Parameter Estimates table. The dispersion parameter is estimated at 1.0363 and is significantly different from zero (as indicated by the confidence interval). Recall that for the negative binomial distribution, the limiting value of 0 for the dispersion parameter corresponds to a Poisson regression model. This implies that overdispersion is evident if a Poisson model were used. Thus, the standard errors from the negative binomial model are more appropriate than those from the Poisson model.

We'll scroll down and look at the LR Statistics for Type 3 Analysis table. With the negative binomial model, none of the effects is significant at an alpha level of 0.05. If you want to reduce your model, you can remove the nonsignificant factors one at a time, starting with the least significant one (**Width**). Reducing your model in this way, the final model has only one significant term, **Weight**.

Let's refit the model with **Weight** as the only predictor variable. We'll turn on the influence diagnostics with the DIAGNOSTICS option and use the PLOTS=ALL option to request all available plots. So that we can identify potentially influential or outlying observations, we use an ODS GRAPHICS statement with the option IMAGEMAP=ON. This provides the HTML format for the output with tooltips turned on.

```
ods graphics/ reset=all imagemap=on;
proc genmod data=mydata.crab plots(unpack)=all;
   model satellites= weight  / dist=negbin link=log type3 diagnostics;
title2 'Reduced Model';
run;
```

Let's run the program and again review the Criteria for Assessing Goodness of Fit table. The AICC (754.79) and BIC (764.10) fit statistics for the reduced model are slightly improved when compared to the full model (AICC = 764.43, BIC = 791.70).

Now we'll look at the Analysis of Maximum Likelihood Parameter Estimates table.

When you use the negative binomial distribution to account for overdispersion, the only factor that is significant in predicting the number of satellites is **Weight**. When you compare this to the previous Poisson model that exhibited overdispersion, **Color** was incorrectly identified as a significant factor. Perhaps this indicates that the Poisson model suffered from a Type I error.

The plots created by ODS Graphics with the DIAGNOSTICS option in the MODEL statement include plots of the standardized deviance residuals, Cook’s D statistic, and standardized DFBETA plots for the intercept and **Weight**.

The plot of the standardized deviance residuals shows random scatter about the reference line, indicating a good model fit. However, standardized residuals higher than 2 or less than -2 should be examined. The residual for observation #15 falls outside this range.

The deviance residuals can take unusual patterns, but unlike standard residuals from a general linear model, they are not used to validate model assumptions. Instead, you use deviance residual plots to look for points with large deviance values that are separated from other points. These points might be influential observations or outliers (Allison, 2012).

Because **Weight** is the only variable in this model, look there for unusual values. This female weighs 2.3 kilograms, close to the average of 2.4 kilograms. What is unusual about this crab is that she has 14 satellites outside her nest. Recall from the data exploration that the mean value of **Satellites** was 2.9. The model does not fit this unusual observation well.

The other data point with a standardized deviance residual more than 2 is observation 149. She weighs 1.9 kilograms, which is below average. She had 10 satellites outside her nest. That is more than the model predicts, given her weight.

This same observation exhibits a positive influence on the parameter estimate for the intercept and a negative influence on the parameter estimate for **Weight**. This data point is also flagged as influential by the plot of the Cook's D statistic. (Notice that the values shown below the observation numbers in the graphs are the values of the statistic plotted on the vertical axis.)

Observation 141 appears to be exhibiting negative influence on the parameter estimate for **Weight** and seems to be potentially influential in the leverage plot. This data point represents a very large crab weighing 5.2 kilograms with 7 satellites.

The three observations flagged in these plots, plus any others that appear to be influential, should be investigated to make sure that they are not data entry errors.

Close