## Using PROC GENMOD to Fit a Poisson Regression Model

In this demonstration, we'll fit a Poisson regression model for count data in the **crab** data set.

Let's begin with some exploratory data analysis. We want to explore the distribution of the outcome variable, **Satellites**. We'll use the SGPLOT procedure to create a histogram of the data. We use DENSITY statements to superimpose a normal distribution and a nonparametric curve fit to the data.

Let's run the program.

```
title;
proc sgplot data=mydata.crab;
   histogram satellites;
   density satellites;
   density satellites / type=kernel;
run;
```

The histogram shows that the data are skewed to the right. The kernel and normal density superimposed on the histogram indicate that the variable **Satellites** does not follow a normal distribution. Let's further explore the data by looking at summary statistics from PROC UNIVARIATE. Again, the variable is **Satellites**. The ODS statement specifies the moments and basic measures tables.

```
ods select moments basicmeasures;
proc univariate data=mydata.crab;
   var satellites;
run;
```

We run the code. Let's look at the first table, with the heading Moments. For a normal distribution, we'd see a skewness statistic of 0. In this case, the skewness statistic is 1.145 is greater than 0, indicating positively skewed data. For a normal distribution, the kurtosis is also 0. Recall that kurtosis measures peakedness. The kurtosis of this data is considerably higher than 0, indicating that this is not a normal distribution.

The mean value (2.92) is greater than the median (2.0). These indicate a skewed-to-the-right distribution. The average for **Satellites** is relatively small, and the variance (9.91) is much bigger than the mean (2.92). Remember that the Poisson distribution assumes that the mean and variance are equal. Because this is not the case for this data, the model will not fit well.

Now, let's use the GENMOD procedure to fit a Poisson regression model. In the MODEL statement, we specify the outcome variable **Satellites**, as well as the continuous variables **Width** and **Weight**; and the categorical predictors **Color** and **Spine**. Because they are categorical variables, we also list **Color** and **Spine** in the CLASS statement.

In the MODEL statement, we'll use the DIST= option to specify the Poisson distribution and the LINK= option to specify the log link function. We'll also specify the TYPE3 option, which requests that SAS compute statistics for type 3 contrasts for each effect specified in the MODEL statement.

```
proc genmod data=mydata.crab;
   class color spine;
   model satellites=width weight color spine
        / dist=poi link=log type3;
   title 'Poisson Model';
run;
```

Let's run the program and review the results.

The Criteria For Assessing Goodness Of Fit table provides statistics for testing the goodness of fit of the model. The measures are deviance and the Pearson chi-squared statistic. The values of these statistics divided by the squared scale parameter (that is, the dispersion parameter) are called scaled deviance and scaled Pearson chi-squared. Because the scale parameter by definition is 1 for Poisson regression, the statistics (original and scaled) are equal.

The Value/DF values are computed by dividing the goodness-of-fit statistics by the degrees of freedom. The degrees of freedom for the Deviance and Pearson Chi-Square are equal to the number of observations minus the number of

regression parameters estimated. These values for the scaled deviance or the scaled Pearson chi-square are useful for assessing the goodness of model fit. Values close to 1 indicate good model fit. The Value/DF column in the table has 3.3308 for scaled deviance and 3.2353 for scaled Pearson chi-square. These values are not close to 1. This might indicate overdispersed data, which can occur frequently in Poisson regression and occasionally in logistic regression. Overdispersion does not affect the parameter estimates, but it does cause the estimates of the standard error of the parameter estimates to be underestimated. You'll learn more about overdispersion later in this lesson.

Other fit statistics include the Akaike information criterion (AIC), the corrected Akaike information criterion (AICC), and the Bayesian information criterion (BIC). Each is a measure of goodness of model fit that balances model fit against model simplicity. These criteria are useful in selecting among models, with smaller values representing better model fit.

The scale parameter can be estimated from your data. For the Poisson distribution, you divide the Pearson chi-square statistic (or the deviance statistic) by the degrees of freedom (which is indicated by the Value/DF column), and then take the square root.

Let's look at the last couple of tables in these results. The Analysis Of Maximum Likelihood Parameter Estimates table provides the parameter estimates and the *p*-values for testing whether the estimates are different from zero.

The LR Statistics for Type 3 Analysis table gives the tests of significance for each of the parameters. However, because the data exhibits overdispersion, these results might not be reliable. Note that the likelihood ratio test for the predictor variables indicates a significant **Weight** effect and a significant **Color** effect. You can compare these tables to those obtained after correcting for the overdispersion.

---



Close