Practice 5.1 (Level 1): Fitting a Poisson Regression Model Using PROC GENMOD

Task

A survey was undertaken to examine which factors are related to ear infections among swimmers. The response variable (**Infections**) is the number of self-diagnosed ear infections reported by the participant. The data is stored in the **mydata.earinfection** data set. In this practice, you use PROC GENMOD to fit a Poisson regression model.

Reminder: Make sure you've defined the **mydata** library.

1. Copy and paste the following code in the editor, and then submit the code. Examine the histogram of the dependent variable **Infections**, and the superimposed normal distribution and nonparametric curve. Is the normal distribution a good fit?

```
title;
proc sgplot data=mydata.earinfection;
   histogram Infections;
   density Infections;
   density Infections / type=kernel;
run;
```

   As shown in the histogram with the normal and kernel densities superimposed, the normal distribution is **not** a good fit for the data.

2. Write a PROC UNIVARIATE step to further explore the data by looking at summary statistics. Use an ODS SELECT statement to look at the Moments table and the Goodness of Fit table. How would you describe the distribution of this variable?

```
ods select moments basicmeasures goodnessoffit;
proc univariate data=mydata.earinfection normal;
   var Infections;
   histogram / normal;
run;
```

   In the results, the Basic Statistical Measures table indicates that the mean number of ear infections is about 1.39, which is low enough that the Poisson distribution would be appropriate for the data. However, the variance is about 5.47, much higher than the mean, so overdispersion might be a problem. The Goodness-of-Fit Tests for Normal Distribution table shows that the tests for normality have low p-values. This indicates that there is enough evidence to reject the assumption of normality for this data set.

3. Write a PROC GENMOD step to fit a Poisson regression model to the data. Which factors seem to be significant? Is overdispersion a problem for this model?

```
proc genmod data=mydata.earinfection;
   class Swimmer Location Gender;
   model Infections=Swimmer Location Age Gender /
        dist=poisson link=log type3;
run;
```

   Examine the results. In the Criteria for Assessing Goodness of Fit table, the scaled deviance and scaled Pearson chi-square for the Value/DF column are 2.6951 and 3.4170, respectively. These values are not very close to 1, which indicates possible overdispersion.

   As indicated by the Analysis of Maximum Likelihood Parameter Estimates table and the LR Statistics For Type 3 Analysis table, the significant factors that affect the average value of the number of infections include **Swimmer**, **Location**, and **Age**. The variable **Gender** is not significant. However, your data might have

overdispersion, which means that the standard errors might be underestimated. You should adjust for the overdispersion before deciding which factors are significant in this model.

4. Use PROC GENMOD to fit a negative binomial regression model to the data. Does this model account for the overdispersion? What factors are now significant?

```
proc genmod data=mydata.earinfection;
   class Swimmer Location Gender;
   model Infections=Swimmer Location Age Gender /
         dist=negbin link=log type3;
run;
```

Examine the results. The goodness-of-fit statistics indicate that the negative binomial model is a good fit to the data. The chi-square value divided by the degrees of freedom is close to 1. Now, only **Swimmer** and **Location** are significant. **Age** is no longer significant in the negative binomial model.

Hide Solution

Statistics 2: ANOVA and Regression

Close