* [Using Poisson Regression to Analyze Ship Damage Rates (Generalized Linear Models)](http://127.0.0.1:54857/help/topic/com.ibm.spss.modeler.tutorial/spss/tutorials/genlin_ships_intro.htm)

# Using Poisson Regression to Analyze Ship Damage Rates (Generalized Linear Models)

A generalized linear model can be used to fit a Poisson regression for the analysis of count data. For example, a dataset presented and analyzed elsewhere [1](http://127.0.0.1:54857/help/topic/com.ibm.spss.modeler.tutorial/spss/tutorials/genlin_ships_intro.htm#fntarg_1) concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the values of the predictors, and the resulting model can help you determine which ship types are most prone to damage.

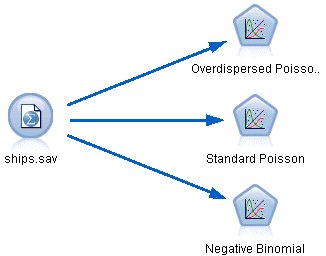
This example uses the stream ships\_genlin.str, which references the data file ships.sav. The data file is in the Demos folder and the stream file is in the streams subfolder.

Modeling the raw cell counts can be misleading in this situation because the Aggregate months of service varies by ship type. Variables like this that measure the amount of "exposure" to risk are handled within the generalized linear model as offset variables. Moreover, a Poisson regression assumes that the log of the dependent variable is linear in the predictors. Thus, to use generalized linear models to fit a Poisson regression to the accident rates, you need to use Logarithm of aggregate months of service.

# Fitting an "Overdispersed" Poisson Regression

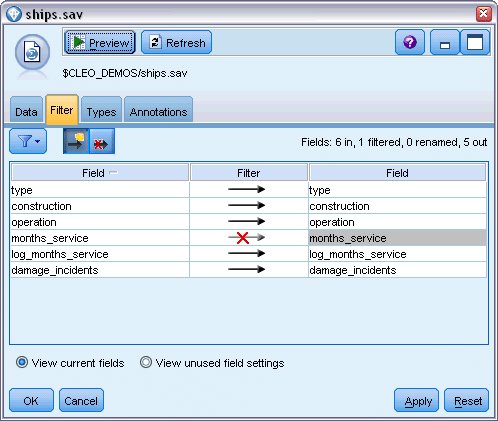
1. Add a Statistics File source node pointing to ships.sav in the Demos folder.

*Figure 1. Sample stream to analyze damage rates*



1. On the Filter tab of the source node, exclude the field months\_service. The log-transformed values of this variable are contained in log\_months\_service, which will be used in the analysis.

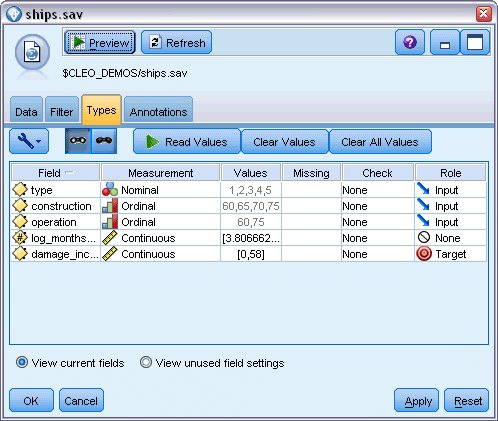
*Figure 2. Filtering an unneeded field*



(Alternatively, you could change the role to **None** for this field on the Types tab rather than exclude it, or select the fields you want to use in the modeling node.)

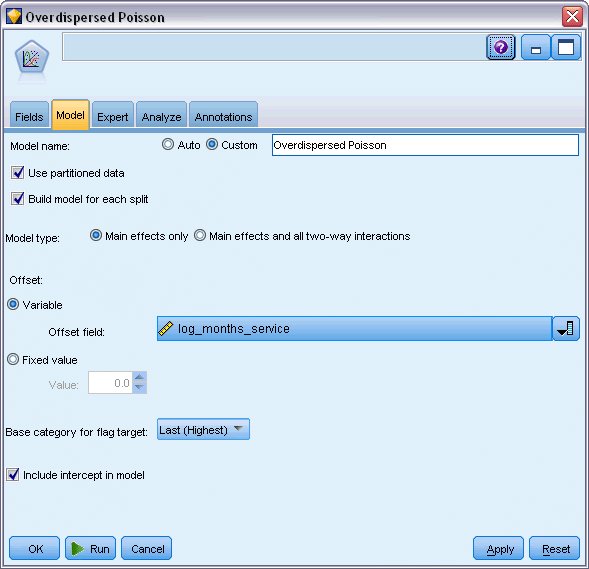
1. On the Types tab of the source node, set the role for the damage\_incidents field to **Target**. All other fields should have their role set to **Input**.
2. Click **Read Values** to instantiate the data.

*Figure 3. Setting field role*



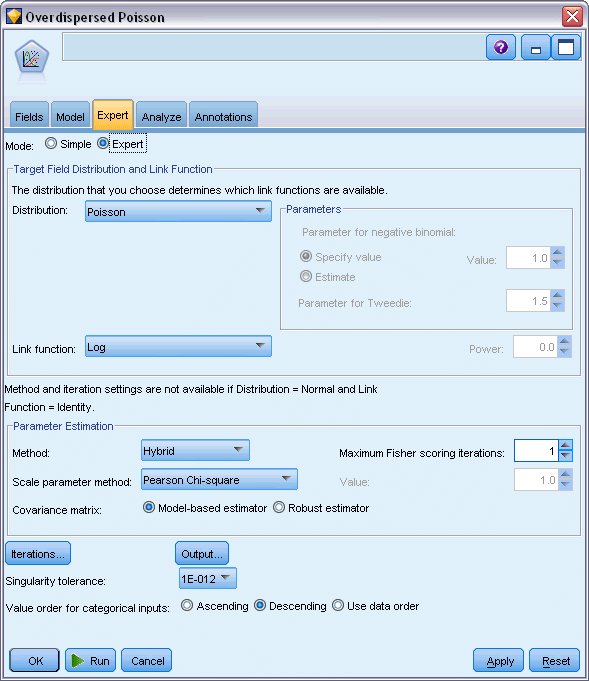
1. Attach a Genlin node to the source node; on the Genlin node, click the **Model** tab.
2. Select log\_months\_service as the offset variable.

*Figure 4. Choosing model options*



1. Click the **Expert** tab and select **Expert** to activate the expert modeling options.

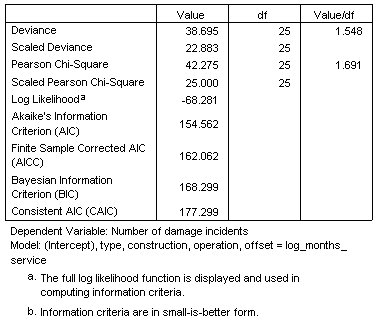
*Figure 5. Choosing expert options*



1. Select **Poisson** as the distribution for the response and **Log** as the link function.
2. Select **Pearson Chi-Square** as the method for estimating the scale parameter. The scale parameter is usually assumed to be 1 in a Poisson regression, but McCullagh and Nelder use the Pearson chi-square estimate to obtain more conservative variance estimates and significance levels.
3. Select **Descending** as the category order for factors. This indicates that the first category of each factor will be its reference category; the effect of this selection on the model is in the interpretation of parameter estimates.
4. Click **Run** to create the model nugget, which is added to the stream canvas, and also to the Models palette in the upper right corner. To view the model details, right-click the nugget and choose **Edit** or **Browse**, then click the **Advanced** tab.

# Goodness-of-Fit Statistics

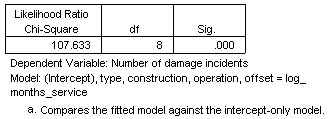
*Figure 1. Goodness-of-fit statistics*



The goodness-of-fit statistics table provides measures that are useful for comparing competing models. Additionally, the Value/df for the Deviance and Pearson Chi-Square statistics gives corresponding estimates for the scale parameter. These values should be near 1.0 for a Poisson regression; the fact that they are greater than 1.0 indicates that fitting the overdispersed model may be reasonable

# Omnibus Test

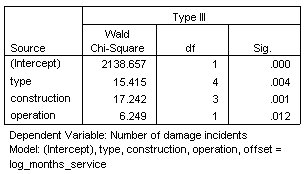
*Figure 1. Omnibus test*



The omnibus test is a likelihood-ratio chi-square test of the current model versus the null (in this case, intercept) model. The significance value of less than 0.05 indicates that the current model outperforms the null model.

# Tests of Model Effects

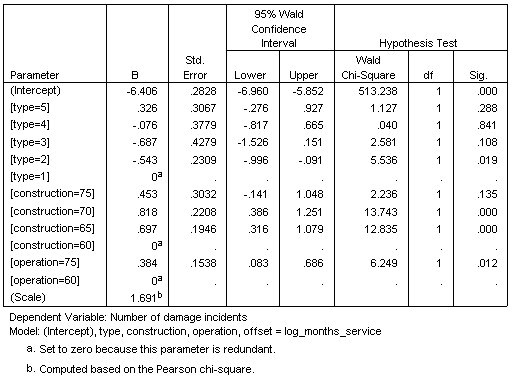
*Figure 1. Tests of model effects*



Each term in the model is tested for whether it has any effect. Terms with significance values less than 0.05 have some discernible effect. Each of the main-effects terms contributes to the model

**Parameter Estimates**

*Figure 1. Parameter estimates*



The parameter estimates table summarizes the effect of each predictor. While interpretation of the coefficients in this model is difficult because of the nature of the link function, the signs of the coefficients for covariates and relative values of the coefficients for factor levels can give important insights into the effects of the predictors in the model.

* For covariates, positive (negative) coefficients indicate positive (inverse) relationships between predictors and outcome. An increasing value of a covariate with a positive coefficient corresponds to an increasing rate of damage incidents.
* For factors, a factor level with a greater coefficient indicates greater incidence of damage. The sign of a coefficient for a factor level is dependent upon that factor level's effect relative to the reference category.

You can make the following interpretations based on the parameter estimates:

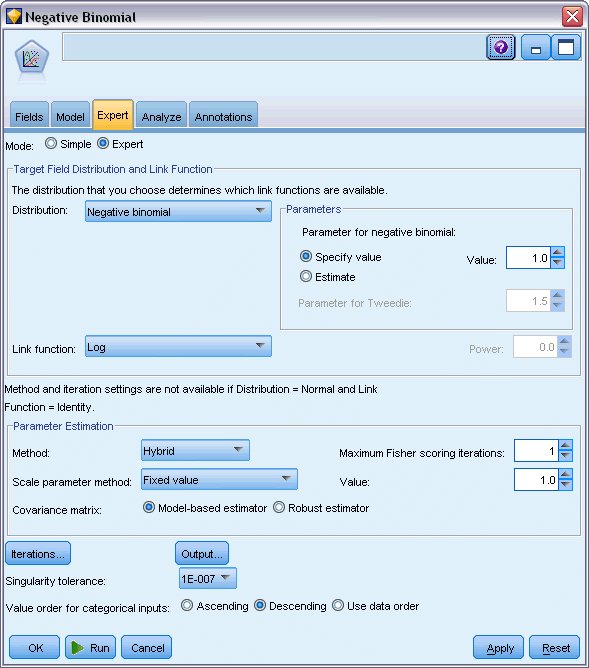
* Ship type *B [type=2]* has a statistically significantly (*p* value of 0.019) lower damage rate (estimated coefficient of –0.543) than type *A [type=1]*, the reference category. Type *C [type=3]* actually has an estimated parameter lower than *B*, but the variability in *C*'s estimate clouds the effect. See the estimated marginal means for all relations between factor levels.
* Ships constructed between 1965–69 *[construction=65]* and 1970–74 *[construction=70]* have statistically significantly (*p* values <0.001) higher damage rates (estimated coefficients of 0.697 and 0.818, respectively) than those built between 1960–64 *[construction=60]*, the reference category. See the estimated marginal means for all relations between factor levels.
* Ships in operation between 1975–79 *[operation=75]* have statistically significantly (*p* value of 0.012) higher damage rates (estimated coefficient of 0.384) than those in operation between 1960–1974 *[operation=60]*.

# Fitting Alternative Models

One problem with the "overdispersed" Poisson regression is that there is no formal way to test it versus the "standard" Poisson regression. However, one suggested formal test to determine whether there is overdispersion is to perform a likelihood ratio test between a "standard" Poisson regression and a negative binomial regression with all other settings equal. If there is no overdispersion in the Poisson regression, then the statistic −2×(log-likelihood for Poisson model − log-likelihood for negative binomial model) should have a mixture distribution with half its probability mass at 0 and the rest in a chi-square distribution with 1 degree of freedom.

1. Select **Fixed value** as the method for estimating the scale parameter. By default, this value is 1.

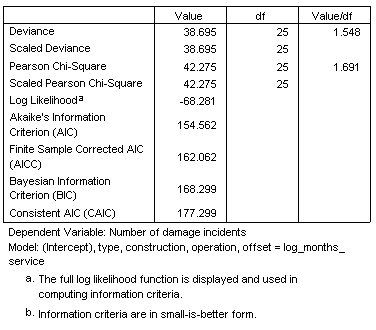
*Figure 1. Expert tab*



1. To fit the negative binomial regression, copy and paste the Genlin node, attach it to the source node, open the new node and click the **Expert** tab.
2. Select **Negative binomial** as the distribution. Leave the default value of 1 for the ancillary parameter.
3. Run the stream and browse the Advanced tab on the newly-created model nuggets.

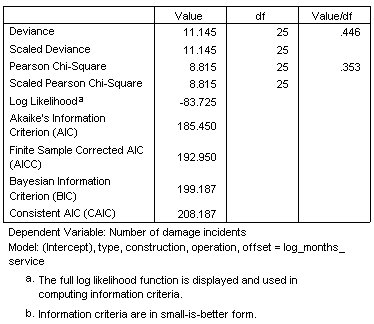
# Goodness-of-Fit Statistics

*Figure 1. Goodness-of-fit statistics for standard Poisson regression*



The log-likelihood reported for the standard Poisson regression is –68.281. Compare this to the negative binomial model.

*Figure 2. Goodness-of-fit statistics for negative binomial regression*



The log-likelihood reported for the negative binomial regression is –83.725. This is actually smaller than the log-likelihood for the Poisson regression, which indicates (without the need for a likelihood ratio test) that this negative binomial regression does not offer an improvement over the Poisson regression.

However, the chosen value of 1 for the ancillary parameter of the negative binomial distribution may not be optimal for this dataset. Another way you could test for overdispersion is to fit a negative binomial model with ancillary parameter equal to 0 and request the Lagrange multiplier test on the Output dialog of the Expert tab. If the test is not significant, overdispersion should not be a problem for this dataset.

# Summary

Using Generalized Linear Models, you have fit three different models for count data. The negative binomial regression was shown not to offer any improvement over the Poisson regression. The overdispersed Poisson regression seems to offer a reasonable alternative to the standard Poisson model, but there is not a formal test for choosing between them.

Explanations of the mathematical foundations of the modeling methods used in IBM® SPSS® Modeler are listed in the *IBM SPSS Modeler* Algorithms Guide.