Print

# Summary: Lesson 5: Generalized Linear Models

This summary contains topic summaries, syntax, and sample programs.

## Topic Summaries

*To go to the movie where you learned a task or concept, select a link.*

### Introduction to Generalized Linear Models

The general linear model is mathematically represented as $Y_i= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ where: $Y_i$ is the $i^{th}$ observed value for the response variable. $x_{1i}$ through $x_{ki}$ are the $i^{th}$ values for the predictor variables $x_1$ through $x_k$. $\beta_1$ through $\beta_k$ are the regression coefficients for the corresponding predictor variables $x_1$ through $x_k$, respectively, and $\varepsilon_i$ is the $i^{th}$ value of random errors.

A generalized linear model extends the general linear model in three ways: First, no assumption of normality is required. Second, the variance of the response variable can be expressed as a function of its mean. Third, a link function, *g*, is used to fit the linear model. Note that we are not modeling the individual values of the response variable, Y, but the expected value or mean of Y.

The link function is any monotonic differentiable function g that relates the mean of y to a linear combination of predictor variables Xβ. In general, if we have a distribution for our outcome data and this distribution has certain restrictions on the parameter values, we create the link function to ensure that these restrictions are upheld.

In generalized linear regression, you determine which link function to use based on the type of response variable and its distribution. For each distribution, there is a link function that can be determined algebraically from the distribution equation. This link is called the canonical link.

When you have a categorical response variable, you know that you use logistic regression to analyze your data. One obstacle is that the predicted values from our regression analysis can take on any value in the set of real numbers. Another problem is that the relationship between the probability of the response and a predictor variable is usually nonlinear rather than linear.

To create a linear model and to constrain the predicted probabilities between 0 and 1, you apply a type of link function called the logit transformation to the probability. The logit transformation is the natural log of odds, where odds is the ratio of the probability of the outcome to the probability of no outcome.

You can also use a generalized linear model with discrete response variables that are counts. You can apply the log link function to count data that have nonnegative integer values. The log transformation removes the lower bound and creates a linear model.

You construct a generalized linear model by selecting the appropriate link function and response probability distribution based on the type of response variable(s) in your data.

A number of link functions and probability distributions are available in SAS procedures such as GENMOD and GLIMMIX. The default link function for linear regression is the identity function. When we use the identity link function, we assume that the expected value of the response can be modeled by a linear combination of the predictors without any transformation (link function) needed.

You use logistic regression when you have a dichotomous, or binary, response variable that uses the binomial distribution. When your response variable is discrete with a Poisson distribution, you apply the log link function.

For the gamma distribution, you use the inverse function, which is the default link function in PROC GENMOD. However, when practical or theoretical limitations indicate, analysts often use the log link function to constrain the predicted values to be positive.

### Poisson Regression and Negative Binomial Regression

Poisson regression is a type of generalized linear model often used to analyze count data. Poisson regression assumes that the response variable follows a Poisson distribution that is conditional on the values of the predictor variables.

Although ordinary least squares regression can be used to analyze count data, Poisson regression has the advantage of being precisely customized to the discrete, often skewed distribution of count data. In addition to being skewed, the sample distribution should have a fairly small mean if Poisson regression is the method of choice.

Count data can include a zero value. When count data have an incidence of zeros greater than expected for the underlying probability distribution, you must use a zero-inflated Poisson model.

The Poisson distribution is used when the variable represents a nonnegative count of some relatively rare event and is skewed to the right. It is fully defined by one parameter, the mean (**λ**), which must be positive. An unusual property of the Poisson distribution is that the mean and variance are equal.

As the mean (**λ**), increases, the skewness decreases, and the distribution becomes more bell-shaped. It starts to approximate the normal distribution. As a discrete distribution, the Poisson distribution might be more conventionally represented by a bar chart.

To ensure that the mean remains positive for all linear predictors, as well as for all parameter and covariate combinations, Poisson regression models use a log link function that relates the expected value of the response variable to the linear combination of predictors.

We estimate the parameters of Poisson regression models by using the method of maximum likelihood. This method finds the parameter estimates that are most likely to occur given the data.

The mean satisfies the exponential relationship, and the fitted values are the exponentiation of the linear predictor, that is, $\mu = e(\beta_1X_1+ \beta_2X_2+...+\beta_kX_k)$. You can see that the effect of the predictors on the mean is multiplicative, not additive, and that you need to consider how the mean will change based on the changing value of a predictor.

Changes in the mean are measured by percent of change. This percent of change can be calculated by subtracting 1 from the exponentiation of the parameter estimate and then multiplying by 100.

You can use PROC GENMOD to fit generalized linear models with a number of built-in link functions and probability distributions. The PLOTS= option in the PROC GENMOD statement controls the plots produced through ODS Graphics. The CLASS statement names the classification variables to be used as explanatory variables in the analysis.

The MODEL statement specifies the response, or dependent variable, and the predictors, also known as the effects or explanatory variables. By default, the model includes an intercept term. You can use the NOINT option to remove the intercept. You can specify the response as a single variable where each observation is a row in the data. You can also specify the response using an event/trials syntax.

You can use the ESTIMATE statement to obtain a test for a specified hypothesis concerning the model parameters. The ‘label’ option identifies the contrast on the output.

```
PROC GENMOD DATA=SAS-data-set <options>;
    CLASS variables;
    MODEL response-effect </ options>;
    ESTIMATE 'label' effect-values </ options>;
RUN;
```

```
MODEL events/trials = <effects> <options>;
```

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. 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.

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.

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.

You can model the overdispersion by using a related distribution for count data that allows the variance to exceed the mean. To fit a negative binomial model, you specify the DIST= NEGBIN option in the MODEL statement in PROC GENMOD.

```
MODEL response-effects <DIST=NEGBIN>;
```

Another way to account for overdispersion is to apply a multiplicative adjustment factor to adjust the standard errors. To apply a multiplicative adjustment factor, specify the PSCALE or DSCALE option in the MODEL statement in PROC GENMOD.

```
MODEL response-effects <PSCALE DSCALE>;
```

The negative binomial distribution is a generalization of the Poisson distribution for count data that permits the variance to exceed the mean. If the distribution of the response variable is Poisson, given the mean at a fixed setting of the predictors, and the mean itself follows a gamma distribution, then it follows that the marginal distribution for the response variable is a negative binomial.

Unlike the Poisson distribution, the negative binomial distribution provides a way to model subject heterogeneity and account for overdispersion. The negative binomial distribution is appropriate for aggregated events.

For a negative binomial distribution, the relationship between the variance and the mean has a dispersion parameter that must be estimated or set to a fixed value. The dispersion parameter, *k*, enables the variance to exceed the mean and enables the negative binomial distribution to account for overdispersion. If the dispersion parameter is much greater than 0, overdispersion is evident and the standard errors increase.

When outcomes occur over time, space, or some other index of size, it might be more useful to model the rate of occurrence rather than the counts. Rates are simply counts divided by the measure of exposure.

To model Poisson outcomes as rates rather than counts, you assume that the numerator variable (that is, the counts) follows a Poisson distribution and that the denominator standardizes the counts.

To model rates using the Poisson regression model, you use the OFFSET= option in the MODEL statement in PROC GENMOD.

```
MODEL response-effects <OFFSET=>;
```

For Poisson regression models for rates, the log of the incidence (where T is a measure of exposure) is modeled as a linear function of the explanatory variables. If you exponentiate both sides of the model expression, you obtain the expected number of events. Notice that the expected number of events is proportional to the index of exposure times the marginal effects of the explanatory variables.

**Introduction to Gamma Regression**

You can use the gamma distribution when the response variable has continuous positive values, is highly skewed to the right, and has variances that are proportional to the squared mean. The gamma distribution also has lighter tails than a lognormal distribution. Zero and negative values are not allowed in the gamma distribution.

Recall that the Poisson regression is useful for modeling the count or rate of rare events, whose variance increases in proportion to the mean. For skewed distributions with relatively large means, the gamma or lognormal distribution might be better choices than the Poisson distribution. An additional consideration in selecting the correct distribution is the tail behavior. Although all of the distributions in this plot have the same expected value and variance, they have increasingly heavy tails.

You use PROC GLIMMIX to fit generalized linear models for nonnormal responses, including: logistic regression models for binary outcomes, Poisson or negative binomial regression models for counts of rare events, and a gamma regression model for continuous skewed, positive values.

The MODEL statement is required and specifies a single response variable and fixed effects. You can specify the response variable by using either the response syntax or the events/trials syntax. The DIST= option in the MODEL statement identifies the distribution that you want to model. You use the LINK= option in the MODEL statement to specify the link function. For gamma regression, you specify DIST=GAMMA and LINK=log.

```
PROC GLIMMIX DATA=SAS-data-set <options>;
    MODEL response<(response-options)>=<fixed-effects>
        < DIST=keyword LINK=keywordoptions>;
    OUTPUT OUT= SAS-data-set keyword=name(s);
RUN;
```

```
MODEL events/trials = <effects> <options>;
```

## Sample Programs

### Fitting a Poisson Regression Model for Count Data

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

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

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

### Modeling Overdispersion by Using the Negative Binomial Distribution

```
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;

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;
```

### Fitting a Poisson Regression Model for Rate Data

```
proc genmod data=mydata.skin;
   class city age;
   model cases= city age / offset=log_pop dist=poi link=log type3;
   title 'Poisson Regression Model for Skin Cancer Rates';
run;
```

### Fitting a Gamma Regression Model

```
ods select Histogram ParameterEstimates GoodnessofFit;
proc univariate data=mydata.cars2;
   var price;
   histogram /gamma(alpha=est sigma=est theta=est color=blue w=2)
                    vaxis=0 to 14 by 2 midpoints=8 to 50 by 2;
   title 'Testing Gamma Distributions';
run;

proc glimmix data=mydata.cars2 plots=studentpanel(unpack);
   model price=hwympg hwympg2 horsepower / dist=gamma link=log solution;
   id model hwympg hwympg2 horsepower price;
   output out=check1 student
        pred(ilink)=   Pred stderr(ilink)=Stderr lcl(ilink)=LCL
                       ucl(ilink)=UCL
        pred(noilink)= XB stderr(noilink)=StderrXB lcl(noilink)=LCLXB
                       ucl(noilink)=UCLXB;
   title 'Cars Data Set - Gamma Distribution with Log Link';
run;

proc print data=check1 (obs=5);
   var Model Hwympg Hwympg2 Horsepower Price pred stderr lcl ucl xb
       stderrxb lclxb uclxb student;
title2 'Predicted Values';
run;

proc print data=check1;
   where student ge 2 | student le -2;
   var model student price pred hwympg horsepower;
title2 'Outlying Student Residuals';
run;

proc glimmix data=mydata.cars2 plots=studentpanel (unpack);
   model price = hwympg hwympg2 horsepower / dist=gamma link=id solution;
   id model hwympg hwympg2 horsepower price;
   output out=check2 student=Student pred(ilink)=Pred;
   title 'Cars Data Set - Gamma Distribution with Identity Link';
run;

proc print data=check2;
   where student ge 2 | student le -2;
   var model student price pred hwympg horsepower;
title2 'Outlying Student Residuals';
run;
title1;   title2;
```

*Statistics 2: ANOVA and Regression*

Close