|  |
| --- |
| [[Home]](http://faculty.chass.ncsu.edu/garson/PA765/index.htm)  [[Syllabus]](http://faculty.chass.ncsu.edu/garson/PA765/pa765syl.htm)  [[Statnotes]](http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm)  [[Links]](http://faculty.chass.ncsu.edu/garson/PA765/links.htm)  [[Lab]](http://hcl.chass.ncsu.edu/ssl/ssl.htm)  [[Instructor]](http://faculty.chass.ncsu.edu/garson/PA765/garson.htm)  [[Home]](http://faculty.chass.ncsu.edu/garson/PA765/index.htm) |

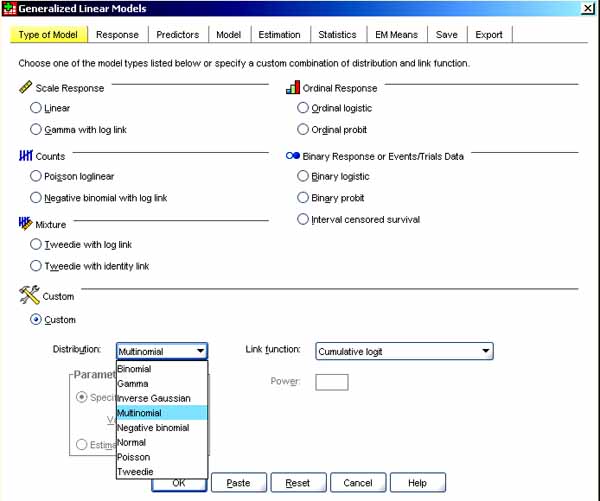
**Generalized Linear Models   
and Generalized Estimating Equations**

|  |  |
| --- | --- |
| **Overview**  The Generalized Linear Model module in SPSS is, as the name suggests, a generalization of the general linear model (GLM) discussed separately with regard to [Anova/Ancova](http://faculty.chass.ncsu.edu/garson/PA765/anova.htm) and [Manova/Mancova](http://faculty.chass.ncsu.edu/garson/PA765/manova.htm) models, as well as regression models. GLM is for variance and regression models which analyze normally distributed dependent variables using an identity link function (that is, prediction is directly of the values of the dependent). In contrast, the generalized linear model (for convenience here abbreviated as GZLM, though note some use GRZ) allows for dependent variables with non-normal distributions and for many link functions other than identity. This means that GZLM supports not only traditional regression models but also logistic models for binary dependents, loglinear analysis of count data, Poisson regression for count data, gamma regression, complementary log-log models for interval-censored survival data, and many others. In SPSS, select Analyze, Generalized Linear Models, Generalized Linear Models...  The Generalized Estimating Equations (GEE) module in SPSS extends GZLM even further by providing support for non-independent data, such as repeated measures and clustered data. GEE thus supports repeated measures logistic regression and many other models for time series or correlated data. In SPSS, select Analyze, Generalized Linear Models, Generalized Estimating Equations. | **Contents**  [Key concepts and terms](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#concepts) [Model type](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#type) [Dependent variable](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#dependent) [Predictor variables](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#predictors) [Models](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#model) [Statistics](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#statistics) [Contrast analysis](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#contrast) [Repeated measures](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#repeated) [Estimation methods](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#estimation) [Assumptions](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#assume) [Frequently asked questions](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#faq) [Bibliography](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#biblio) |

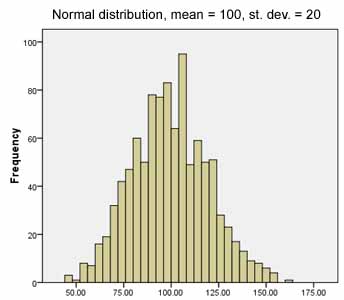
**Key Concepts and Terms**

* + **Type of model**. The "Type of model" tab is identical in both GZLM and GEE models in SPSS. Here one specifies the distribution of the dependent (response) variable, and also specifies the link function by which it is modeled. Where GLM and traditional approaches assume a continuous, normally distributed dependent linearly linked to the predictor variables using an identity link function, GZLM and GEE allow for dozens of model combinations, not just one. The dependent may be non-continuous and may assume any of several different distributions other than normal (ex., Poisson). The dependent also may be nonlinearly linked to the the predictor variables (ex., a log link).

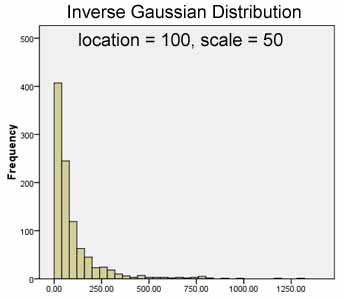
There are two main ways to select the distribution and link function: (a) one can click on the "Custom" radio button and select any of eight distributions for the dependent and any of 15 link functions (depending on the distribution selected); or (b) one can select from a list of "short cut" distribution/link combinations that represent common models, grouped by data level of the dependent variable.



* + - *Distribution*. The dependent may be distributed in any of these forms in the SPSS 16 implementation of GZLM/GEE:
      1. Normal. The distribution of the dependent takes the form of the famous bell-shaped symmetrical curve centered on the mean. This implies the dependent is continuous numeric. A normal distribution is illustrated below. (This and ensuing figures are generated using the SPSS random number generator and distribution functions. Parameters, such as standard deviations, are indicated. Different random runs and different parameters will affect the exact shape of the distribution.)

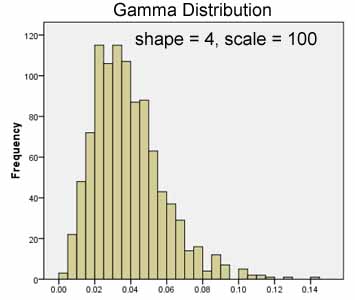


* + - 1. Inverse Gaussian. Also called the inverse normal or Wald distribution, this is used for dependents which are [positively skewed](http://faculty.chass.ncsu.edu/garson/PA765/assumpt.htm#skew) and have values always greater than 0. Values must be greater than 0 or are dropped. The inverse Gaussian distribution has been used to model diffusion processes, insurance claims, and repair times as well as in other reliability variables.

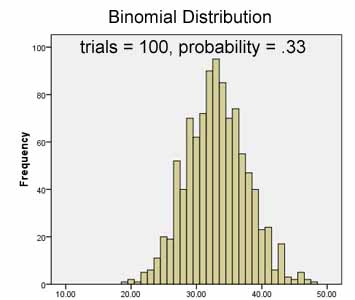


* + - 1. Gamma. This is an alternative for positively skewed dependents. It is highly sensitive to the shape parameter. When the shape parameter is greater than 1, the gamma distribution is mounded but positively skewed as shown in the figure below. When the shape parameter is 1, the gamma distribution is exponentially declining. When the shape parameter is less than 1, the gamma distribution is also exponentially declining and asymptotic to the axes.

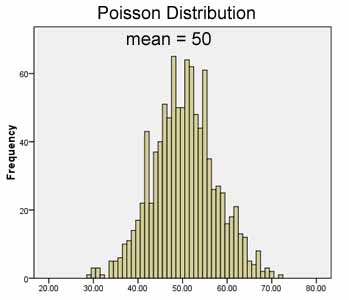
The gamma distribution is used to model variables whose values are positively skewed and always greater than 0. Values not greater than 0 are dropped in gamma models. The selection of gamma over inverse Gaussian is particular to the researcher's dataset. Survival times are typically gamma distributions, though they may also be exponential. The gamma distribution has thus been used in survival analysis (modeling duration-to-event data), but can only be so used if there are no censored cases (no cases still waiting for the event when the research study period ended). It has also been used in varied applications such as estimating insurance claims and rainfall amounts, modeling death rates, and in queuing applications. See also [event history analysis](http://faculty.chass.ncsu.edu/garson/PA765/event.htm) and [Cox regression](http://faculty.chass.ncsu.edu/garson/PA765/cox.htm), which can handle such censored cases.



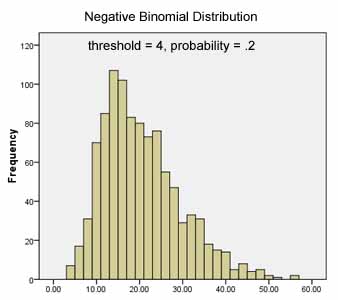
* + - 1. Multinomial. This distribution is used when the dependent has a finite number of categories, has text string values, or is ordinal. The distribution among categories, not shown, is arbitrary.
      2. Binomial. Used when the dependent is binary. The count of events in a fixed number of trials also has a binary distribution. Examples of binomial data are attribute present/not present, innovation adopted/not adopted, or sucess/failure data. It is assumed that the two values have a fixed rather than changing probability of occurrence (as in coin-flipping), even if that probability is not known.



* + - 1. Poisson. The Poisson distribution is also used for count data and is preferred when events are rare, as in modeling accidents, wars, or epidemics. A rule of thumb is to use a Poisson rather than binomial distribution when n is 100 or more and the probability is .05 or less. Where the binomial distribution is used where the variable of interest is count of successes per given number of trials, the Poisson distribution is used for count of successes per given number of time units. The Poisson distribution is also used when "event occurs" can be counted but non-occurrence cannot be counted. In Poisson distributions, the mean equals the variance. All values are non-negative integers (thus count data, which cannot be negative, are better represented by Poisson than normal distributions). Cases with negative values or non-integers must be dropped from the analysis.



* + - 1. Negative binomial. This is similar to the Poisson distribution, also used for count data, but is used when the variance is larger than the mean. Typically this is characterized by "there being too many 0's." It is not assumed all cases have an equal probability of experiencing the rare event, but rather that events may cluster. The negative binomial model is therefore sometimes called the "overdispersed Poisson model." Values must still be non-negative integers. The negative binomial is specified by an ancillary (dispersion) parameter, k. When k=0, the negative binomial is identical to the Poisson distribution. The researcher may specify k or allow it to be estimated by the program.



* + - 1. Tweedie. The Tweedie distribution was used first to model rainfall in crop science in the 1980s, and since has been used for varied topics including study of survival times, performance degradation, womens' paid working hours, rainfall amounts, and other topics. If selected, the Tweedie distribution requires a parameter, p, which the researcher enters to determine the shape of the distribution:

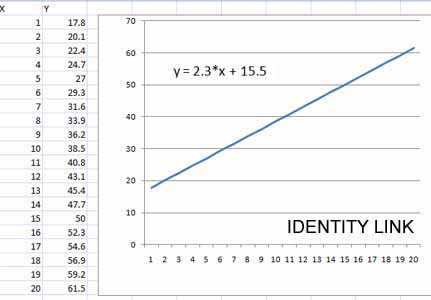
p=0: normal distribution  
p=1: Poisson distribution  
1< p< 2: for continuous data with exact zeros (the default in SPSS is 1.5)  
p=2: gamma distribution  
p>2: for positive continuous data

In a Tweedie distribution, values must be numeric and equal to or greater than 0, or are dropped. When, as in measuring rainfall, many values (days) have a 0 value, and the overall distribution has a [positive skew](http://faculty.chass.ncsu.edu/garson/PA765/assumpt.htm#skew), the Tweedie distribution may work well. See Gilchrist & Drinkwater (2000).

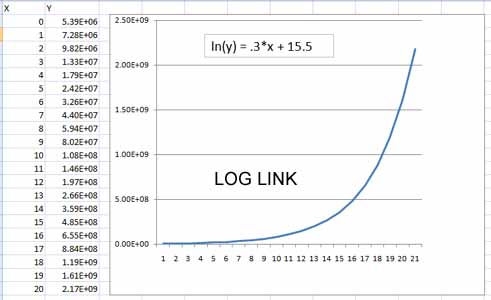
* + - *Link function*. The link function is.the transformation of the dependent variable which is modeled. Transformation may be needed because the predictors have a nonlinear effect on the dependent. OLS regression and GLM in general use an identity function, modeling the actual value of the dependent variable in a linear fashion.

For link functions in general, the model coefficients times the corresponding predictors (plus error) equal the expected value of the link function of the dependent. This implies the link function serves to maintain a linear relationship between the coefficients and predictors on the right-hand side of the model equation and the dependent transformed by the link function on the left-hand side of the equation. The choices for link function depend on the distribution selected.

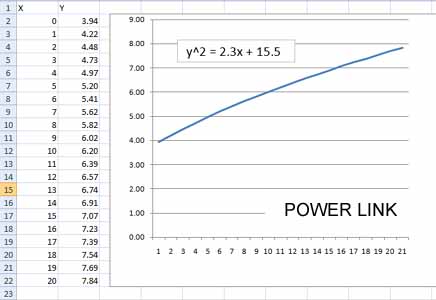
* + - 1. *Normal, gamma, inverse Gaussian, Poisson, and Tweedie distributions* all support the identity, log, and power (exponential) link functions only. In fact these three link functions may work with any type of distribution.
         * Identity. This means the dependent variable is predicted directly and is not transformed. Identity is the link function in OLS regression and in GLM.



* + - * + Log. The natural log of the dependent variable is modeled. Loglinear models assume a Poisson distribution and use a log link function, for instance. In general, log link functions are used in Poisson regression.



* + - * + Power. Some exponent of the dependent variable is modeled.



* + - 1. *Negative binomial distributions* can be used with identity, log, or power link functions, or with the negative binomial link function. Only the negative binomial distribution can use this link function. The SPSS default for negative binomial distributions, however, is the log link function since it allows easier comparison with Poisson models.
         * Negative binomial. What is modeled is log(x /(x+k-1)), where x is the dependent and k is the ancillary parameter of the negative binomial distribution, discussed above.
      2. *Binomial distributions* can be used with identity, log, or power link functions, or with the link functions below. Note that the link functions below can only be used with dependents which have a binomial distribution.
         * Logit. f(x)=log(x /(1-x)). The logit link function transforms the dependent variable, Y, by taking the natural log of the quantity, the probability the dependent equals a given value (usually 1, indicating an event has occurred or a trait is present) divided by the probability it does not. Put another way, logit models model the probability of success (y=1) given the values of various covariates.

Binomial logistic regression models assume a binomial distribution with a logit link function. Likewise, ordinal regression (which is built on a series of binomial logistic regressions) uses a logit link function. High and low values of the dependent are equally probable.

* + - * + Probit. f(x)=F-1(x), where F-1 is the inverse standard normal cumulative distribution function. Probit models are often used for response rate data, such as recovery rates in response to various levels of dosage in a medical study, or other data where the dependent is a proportion. The probit link function transforms the dependent (response) variable by using the normal curve value that corresponds to the raw percent value (ex., 0 on the normal curve corresponds to 50% of the area; 1.64 corresponds to 95% since 95% of the area of the normal distribution is to the left of the normal standard deviate of 1.64). To avoid negative numbers, the probit transform often is made to add 5 to the computed normal standard deviate for any given proportion. Probit regression models assume a binomial distribution with a probit link function.
        + Complementary log-log. f(x)=log(-log(1-x)). This link function may be used when the probability of an event is very small or very large. This link is sometimes used with discrete survival data. The complementary log-log link is used particularly when the estimated probability is asymmetric with respect to the predictor, with higher values of the dependent being more probable. Logit and probit links are symmetric functions and are not appropriate for asymmetric probabilities, hence complementary log-log models are used instead. Sometimes these are referred to as "clog" or "cloglog" models.
        + Negative log-log. f(x)=-log(-log(x)). Used when the estimated probability is asymmetric with respect to the predictor, with lower values of the dependent being more probable.
        + Log complement. f(x)=log(1-x). Sometimes referred to as "logc" models, this link function has been used when computing health risk estimates.
        + Odds power. f(x)=[(x/(1-x))a-1]/a, if a ? 0. f(x)=log(x), if a=0. a is the required number specification and must be a real number.
      1. *Multinomial distributions* can be used with identity, log, or power link functions, or with the link functions below. Note that the link functions below can only be used with dependents which have a multinomial distribution.
         * Cumulative logit. f(x)=ln(x / (1-x)), applied to the cumulative probability of each category of the response. This is the SPSS default and is the usual link function for multinomial (ordinal) logistic regression, paralleling the binary logistic model..
         * Cumulative probit. f(x)=F-1(x), applied to the cumulative probability of each category of the response, where F-1 is the inverse standard normal cumulative distribution function. This is the link function for ordinal probit models.
         * Cumulative Cauchit. f(x) = tan(p (x – 0.5)), applied to the cumulative probability of each category of the response. Used when there are many extreme values.
         * Cumulative complementary log-log. f(x)=ln(-ln(1-x)), applied to the cumulative probability of each category of the response. It is used particularly when the estimated probability is asymmetric with respect to the predictor, with higher values of the dependent being more probable.
         * Cumulative negative log-log. f(x)=-ln(-ln(x)), applied to the cumulative probability of each category of the response. Used when the estimated probability is asymmetric with respect to the predictor, with lower values of the dependent being more probable.
    - *Common model types*. The SPSS "Type of model" tab offers several common models, grouped by data level. Each model implies a certain distribution and link function.
      1. For scale/interval dependents:
         * Linear. Specifies a normal distribution and identity link function. Regression, anova, and GLM models are of this type.
         * Gamma with log link. Specifies a gamma distribution and log link function. *Gamma regression* is often used when the response (dependent) variable is [positively skewed,](http://faculty.chass.ncsu.edu/garson/PA765/assumpt.htm#skew) with a long tail to the right. The degree of skew is estimated by the scale parameter, set under the Estimation tab in SPSS (default is fixed). Gamma regression assumes that the dependent/response variable has no negative values and takes on an asymmetric distribution with a long right tail. That is, gamma regression may be appropriate when modeling dependents with non-negative values which are skewed in distribution. For instance, Chard & Arango (2003) used gamma regression in a World Bank study where the dependent was a Likert scale item where 85% of respondents had chosen 4 or 5 on a 5-point scale. In general, gamma regression may be preferred to linear regression when the normality assumption is violated. Gamma regression is also used in survival analysis (analysis of duration-to-event data), as discussed [above](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#gammadistribution). Gamma regression assumes the distribution of the response variable is continuous. When the response variable has a small to moderate number of values, Poisson or negative binomial regression may be preferred (Hilbe & Hardin, 2001: 69).

In SPSS, select Analyze, General Linear Models, General Linear Models; under the "Type of Model" tab, select "Gamma with log link"; under the "Response" tab, select the dependent variable (ex., the survival/duration variable); fill out other tabs as usual.

*Gamma family variations*. Gamma distributions may be modeled with other link functions described below. The alternative with the lowest deviance (or lowest AIC or BIC) is usually preferred. This is found in the "Goodness of Fit" table in SPSS output, described in a later section of this module.

*Identity-gamma regression* specifies the gamma distribution but uses an identity link function. Identity-gamma regression is often used to model duration data.

*Gamma with inverse log link*. Another variant of gamma regression uses the inverse (power -1) link function (this is the canonical link function for the gamma distribution). Note that when gamma regression is parameterized using an inverse power link, the signs of the coefficients are reversed compared to gamma with log link models.

* + - 1. For ordinal dependents:
         * Ordinal logistic. Specifies a multinomial (ordinal) distribution with a cumulative logit link function. Ordinal logistic regression models are of this type.
         * Ordinal probit. Specifies a multinomial (ordinal) distribution with a cumulative probit link function. Ordinal probit regression models are of this type.
      2. For count data, which requires that the duration of the observation period for each unit of analysis is the same (ex., in counts of walk-in Social Security clients by field office, the observation periods for each office must be regular intervals which are the same, such as months or years) :
         * Poisson loglinear. Specifies a Poisson distribution with a log link function. This is the Poisson regression model, which can be implemented in either GZLM or GEE, depending on whether there are repeated measures or not..

*Poisson regression for count data*. In SPSS, select "Poisson loglinear" in the "Counts" section of the "Type of Model" tab of the GZLM dialog, thus selecting a model for a dependent variable with a Poisson distribution, using a natural log link function. Set the dependent variable under the "Response" tab. Under the "Predictors" tab, enter factors and/or covariates as predictors. Complete "Model," "Statistics," and other tabs as usual.

Loglinear models. Note that loglinear data is count data (cell counts within tables). Therefore Poisson regression models (Poisson distribution, log link function) are loglinear models when all the predictors are factors (have finite levels, like row and column variables in a table) and one models factors and/or interactions involving factors. Parameter estimates will be the same as in Analyze, Loglinear, General in SPSS menus.

*Checking overdispersion*. In Poisson models, the variance should equal the mean. If the variance exceeds the mean, there is "overdispersion" and the data may be better modeled with the negative binomial distribution. Specifically, when there is overdispersion, Poisson regression underestimates standard errors, causing Type I errors (thinking predictors are significant when they are not). Overdispersion is indicated in the "Goodness of Fit" table in the "Value/df" column for the "Deviance" or "Pearson chi-square" rows. The ratio of deviance (or Pearson chi-square) to degrees of freedom should be close to 1.0 in a well-fitting Poisson model. If it is appreciably greater, the negative binomial model may fit better. One can use the likelihood ratio test to see if a negative binomial model is significantly better, as explained [below](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#lratiotests). In this test, the Poisson model is the "reduced model," nested within the negative binomial model as the "full model," since the negative binomial model has one additional term, k, representing a dispersion parameter . It is also possible to have "underdispersion," indicated by a deviance/df ratio appreciably smaller than 1.0. With either under- or over-dispersion, one should ask for robust estimation of coefficients, selected under the "Estimation" tab.

*Poisson regression for rate data*. While Poisson regression is for count data, it can be adapted for rate (percentage, ratio) data by using an offset variable, as discussed [below](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#offset). In models of this type, rate is count per measure of exposure (ex., per population size, per area, or per length of time). For instance, to model number of lost pieces of luggage per total air miles traveled, the count variable (number of pieces lost) is entered as a response variable with the offset being the log of exposure (log of number of air miles traveled, the rate's denominator). Exposure is not a predictor but an offset (in log form, since the log link function is used in Poisson regression). Rate cannot be modeled directly since Poisson models are for counts, not rates. The offset variable is added to the Poisson regression equation as a variable whose coefficient is believed to be 1. The effect is to make the expected value of the dependent increase in proportion to the size of the exposure variable which in log form serves as the offset. In practical terms, this means the researcher must create an offset variable which is equal to the log of the exposure variable and use that, not the exposure variable itself, as the offset variable.

In SPSS, select "Poisson loglinear" in the "Counts" section of the "Type of Model" tab of the GZLM dialog, thus selecting a model for a dependent variable with a Poisson distribution, using a natural log link function. Set the dependent variable (number of pieces of luggage lost, in our example) under the "Response" tab. Under the "Predictors" tab, enter factors and/or covariates as predictors, then click the "Variable" radio button under "Offset" and enter the offset variable name (the variable containing the natural log of number of air miles traveled, in our example). Complete "Model," "Statistics," and other tabs as usual, but note the Offset Variable cannot be used as a model effect as it is cannot be a predictor.

* + - * + Negative binomial with log link. Specifies a negative binomial distribution (with the ancillary parameter = 1) with a log link function. This is the negative binomial regression model. This model is common for count data when a Poisson model is marked by overdispersion, discussed [above](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#overdispersion). That is, negative binomial regression is used when modeling count data that violates the Poisson assumption of equality of mean and variance. Also, negative binomial regression is thought to be more stable than Poisson regression for small datasets (Rogers, 1993).

*Lagrange multiplier test*. The Lagrange multiplier test may be used to test if a negative binomial model is significantly different from a Poisson model. Since the negative binomial model is the same as the Poisson model when the binomial model's ancillary (dispersion) parameter, k, equals 0, the Lagrange multiplier test is a test of the null hypothesis that k = 0. A non-significant Lagrange test coefficient indicates that k cannot be assumed to be different from 0, and hence a Poisson model would be preferred over a negative binomial model on parsimony grounds (negative binomial models have one more parameter, k).

* + - 1. For binary or events/trials dependents:
         * Binary logistic. Specifies a binomial distribution with a logit link function. This is the binary logistic regression model.

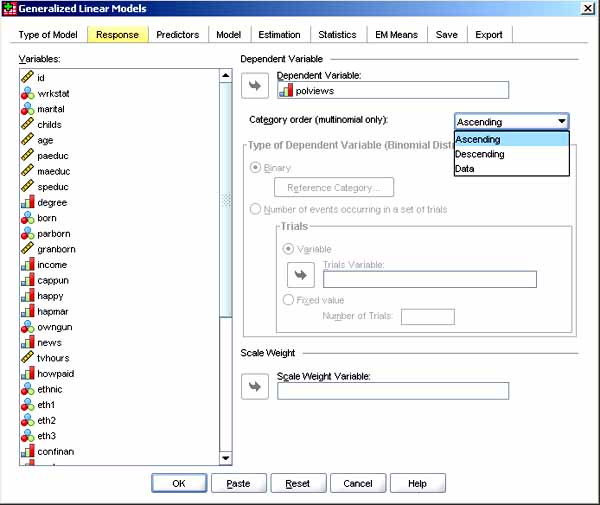
*Logistic regression in SPSS GZLM*. In SPSS, select Analyze, Generalized Linear Models, Generalized Linear Models; under the Type of Model tab, specify Binary Logistic (one of the pre-set models - this selects the binomial distribution with logit link function); under the Response tab, specify the Dependent Variable and click on the Reference Category button to set the reference category (this is under the Binary radio button, which will already be checked); under the Predictors tab enter the factors and covariates; under the Model tab specify the main effects and interactions you wish to model; under the Estimation and Statistics tabs you may accept defaults or set options; if you have factors as predictors, you can call for contrast analysis under the EM Means tab; optionally you may ask to save variables under the Save tab or save the model under the Export tab; click OK at the bottom to run the model.

* + - * + Binary probit. Specifies a binomial distribution and with a probit link function. This is the binary probit regression model.

*Probit regression in SPSS GZLM*. In SPSS, select Analyze, Generalized Linear Models, Generalized Linear Models; under the Type of Model tab, specify Binary Probit (one of the pre-set models - this selects the binomial distribution with probit link function); under the Response tab, specify the Dependent Variable and if applicable check the radio button for "Variable represents number of events occurring in a set of trials", and enter a fixed number of trials or specify the variable containing the Trials Variable; under the Predictors tab enter the factors and covariates; under the Model tab specify the main effects and interactions you wish to model and uncheck the "Include intercept in model" checkbox (this is because if you have an intercept, the coefficient for the highest factor level will be 0 - unless you add the estimated intercept value to it and to the coefficients for all the other levels); under the Estimation and Statistics tabs you may accept defaults or set options; if you have factors as predictors, you can call for contrast analysis under the EM Means tab; optionally you may ask to save variables under the Save tab or save the model under the Export tab; click OK at the bottom to run the model.

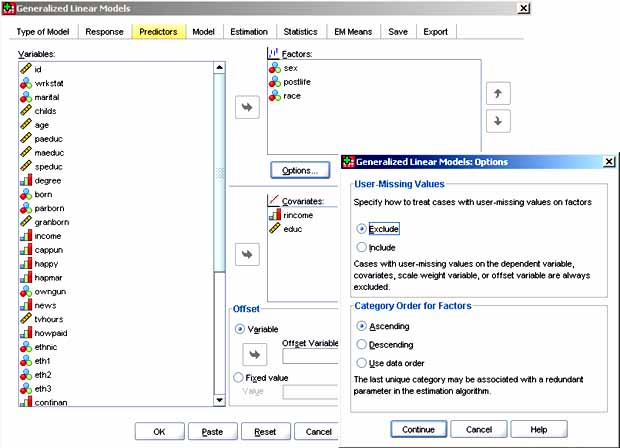
* + - * + Interval censored survival. Specifies a binomial distribution with a complementary log-log link function.
      1. Mixture.
         * Tweedie with log link. Specifies a Tweedie distribution with a log link function.
         * Tweedie with identity link. Specifies a Tweedie distribution with an identity link function.

* + **The dependent (response) variable** is specified in SPSS under the "Response" tab, identical in either GZML or GEE. There are relatively few parameters to be filled out on the "Response" dialog box:



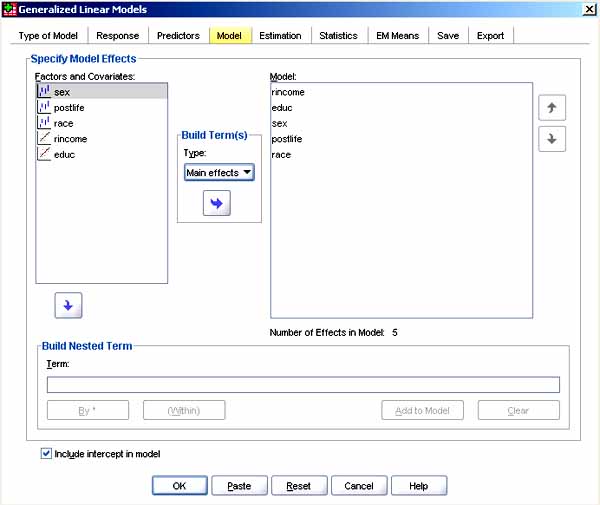
* + - *Dependent variable*. The dependent variable is also called the response variable. Move the desired dependent variable into the "Dependent Variable" textbox from the "Variables" listbox. As noted above, if one has repeated measures and thus is in GEE, the Subject variable(s) and the Within Subject variable(s) cannot be used as the dependent (response) variable.
    - *Category order*. This applies only if the dependent variable is multinomial in distribution, in which case one selects Ascending, Descending, or Data (which means the first value encountered in the data defines the first category, the last value encountered defines the last category). .
    - *Binary*. If the dependent variable has been specified as binary in distribution, then the Binary radio button is clicked on and one gets to click on "Reference Category" to set the reference category as the highest, lowest, or a custom value. Highest is default, meaning that what is modeled is the probability of the dependent being 1 rather than 0.
    - *Number of events occuring in a set of trials*. If the dependent variable has been specified as binary in distribution and one has events-in-trials data, then this radio button may be clicked on and one may enter the name of the variable (the trials variable) recording the number of trials if this varies by subject, or if a fixed value for all subjects, the "Fixed value" radio button may be clicked and the number of trials simply entered as a positive integer. Of course, the number of trials must be greater than or equal to the number of events for a subject. If in GEE, the Subject variable(s) and the Within-Subjects (repeated measures) variables cannot be used as the trials variable.
    - *Scale Weight* allows the researcher to move from the "Variables" listbox the name of the scale weight variable, if any. Values of this variable must be greater than 0 or the case will be dropped.

* + **The independent (predictor) variables** are specified in SPSS under the "Predictors" tab, identical in either GZML or GEE. The SPSS manual warns, "Note: When the response is binomial with binary format, the procedure computes deviance and chi-square goodness-of-fit statistics by subpopulations that are based on the cross-classification of observed values of the selected factors and covariates. You should keep the same set of predictors across multiple runs of the procedure to ensure a consistent number of subpopulations." There are only a few sorts of information to enter in this dialog box:



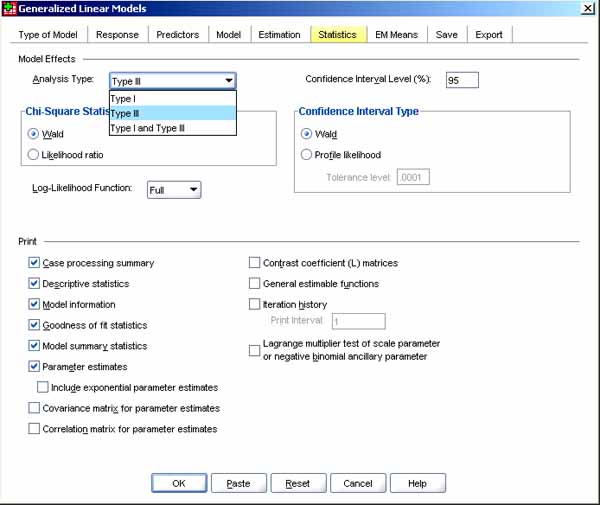
* + - *Factors* Categorical independents can be dragged from the "Variables" listbox to the "Factors" textbox.
    - *Covariates*. Scale (interval numerically coded) predictors can be dragged from the "Variables" listbox to the "Covariates" textbox.
    - *Missing values and factor order options*. The Options button allows the researcher to set two things: (a) whether to include/exclude cases which have missing values on the dependent, on factors, on the scale weight variable, or on the offset variable (exclude is the default); and (b) if factors are in the model, if their order is to be ascending, descending, or in order encountered in the data. While the order of factor levels will not affect predicted values of the dependent or goodness of fit measures, changing the order may change which value is the reference category and thus change the meaning of the coefficients. Also, changing order when there is a factor-by-covariate interaction may well change the value of the coefficient for that covariate. Order will also affect the intercept and its significance.
    - *Offset Variable*. If desired, the values of some variable can be added to the linear predictor of the dependent variable. (The linear predictor is the right-hand side of the model equation). Alternatively, a fixed value may be added. Offsets are commonly used in Poisson regression models for rate data, as discussed [above](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#poissonrate).

* + **Specifying the Effects to Be Modeled** is done in SPSS under the "Model" tab, identical in GZLM and GEE. The interface is similar to that for specifying effects in GLM. There are three main sections to enter in the "Model" dialog:



* + - *Specify Model Effects*. The "Factors and Covariates" box will list variables entered under the "Predictors" tab. The researcher enters these variables into the "Model" box using the "Build Terms" operator in the middle, which has such choices as Main Effects, Interaction, Factorial, All 2-way, All 3-way, etc., up to All 5-way, which will apply to whatever variables the researcher has highlighted in the "Factors and Covariates" list. The "Factorial" choice includes in the model all main effects and all possible interactions, for the selected variables. If nothing is entered in the "Model" box, the default model is intercept only.
    - *Build Nested Term*. Any factor or covariate from the "Factors and Covariates" box may also be entered into the "Term" textbox of the "Build Nested Term" portion of the "Model" dialog. Normally this is a grouping variable which does not interact with other main effects in the model. For instance, "Agency" might be a nested term in a study of employee attitudes: since any given employee is only employed by one agency, one would assume no interaction of agency with other factors and covariates. One would enter Attitude into the "Build Nested Term" textbox, click the "Within" button, and then enter Agency, creating the term Attitude(Agency); then click Add to the Model. One can also create interactions this way: enter Attitude into the "Build Nested Term" textbox, click the "By" button, and then enter Agency, creating the term Attitude\*Agency; then click Add to the Model. There can be multiple levels of nesting. A given factor or covariate cannot be a nested term if it is also in the "Model" box. Factors cannot be nested within covariates.
    - *Include Intercept in Model*. Including the intercept is the default. This box would be unchecked only if the researcher was certain that the slope passed through the origin. Note models with a multinomial ordinal distribution do not have a single intercept but rather threshold parameters that define transition points between adjacent categories. The thresholds are always included in the model and are not optional.

* + **Statistics**. The "Statistics" tab gives output options, almost identical in GZLM and GEE. GZLM offers a choice of confidence interval type and a Lagrange multiplier test not available in GEE. GEE will print a working correlation matrix and employs a "log quasi-likelihood function" rather than the usual log-likelihood function in GZLM. This is because the log-likelihood function assumes independent data, whereas GEE deals with repeated measures.



* + - Options
      1. *Analysis type*: Analysis type refers to the type of sum of squares used in the algorithms used for computing likelihood ratio and Wald statistics in GZLM (see chi-square statistics, below) or in the generalized estimating equations in GEE. From the pull-down menu, one may select Type I, Type III, or Type I and Type III. If "Type I and Type III" is selected, tests of model effects are reported both ways. Type III is the default, used for most research (by far).

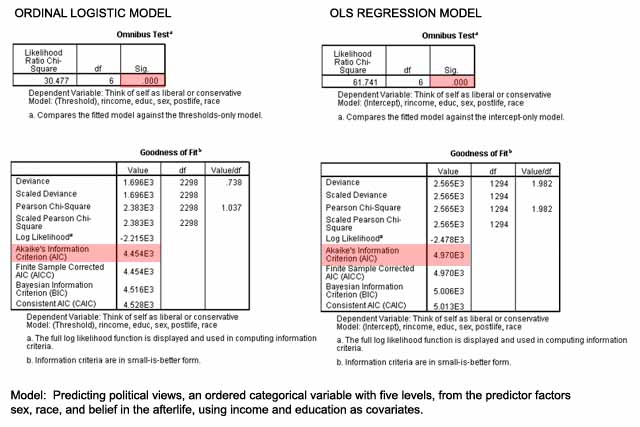
Type I analysis is recommended when the researcher is able to stipulate beforehand the order of model predictors, in a hierarchical design where main effects are specified before first-order interaction effects, and first-order interaction effects are specified before second-order interaction effects, etc. Type I is also used for purely nested models where a first effect is nested within a second effect, the second within a third, etc.; and is used in polynomial regression models where simple terms are specified before higher-order terms (ex., squared terms). Type I tests may be thought of as a test of sequential effects: if an effect is non-significant by Type I, this means it cannot be assumed to be different from 0 when that effect is entered into a model containing only effects that precede it in order of entry.

* + - 1. *Chi-square statistics*: Where F tests are used to test linear models, in GZLM either Wald chi-square or likelihood ratio chi-square (GZLM) or generalized score chi-square (GEE) statistics are used to test the model. Thus in GZLM one may select between the Wald (the default) and likelihood ratio chi square. In GEE one may select between Wald (the default) and generalized score chi-square statistics. Wald is the default in both cases. Likelihood ratio tests are preferred in multinomial logit models, however. Also, Wald tests may be inaccurate with small samples.
      2. *Confidence intervals*: In either GZLM or GEE one may specify any confidence limits from 50 to 100, but 95 is the default. In GEE, Wald confidence intervals are always computed regardless of the "Chi-square statistics" selection. In GZLM one may select between Wald intervals or profile likelihood intervals. Profile confidence intervals add a small degree of accuracy while being computationally intense.
      3. *Likelihood functions*: GZLM uses the log likelihood function while GEE uses the log quasi-likelihood function. In either case, the researcher may select between "Full" or "Kernel" format for the likelihood function. Full is the default. Full adds a constant term to the format which, since it has no effect on parameter estimation, is left out of the format of some other software. Other software may be emulated by selecting Kernel. In sum, this choice makes little difference.
    - Output
      1. *Case processing summary/Descriptive statistics/Model information*: These output options, checked by default in SPSS, generate (a) the number and percentage of cases included and excluded from the analysis and the Correlated Data Summary table; (b) descriptive statistics and summary information about the dependent variable, covariates, and factors; and (c) the dataset name, dependent variable or events and trials variables, offset variable, scale weight variable, probability distribution, and link function, respectively.
      2. *Goodness of fit statistics*: This output option is a default also, causing GZLM to display deviance, scaled deviance, Pearson chi-square, scaled Pearson chi-square, log-likelihood, Akaike's information criterion (AIC), finite sample corrected AIC (AICC), Bayesian information criterion (BIC), and consistent AIC (CAIC) goodness of fit measures in the "Goodness of Fit" table in SPSS. GEE displays quasi-likelihood under the independence model criterion (QIC) for choosing the best correlation structure and another QIC measure for choosing the best subset of predictors. QIC is an extension of AIC.
         * *Likelihood ratio tests*. The likelihood value reflects how likely it is that the model would result in the observed patterns in the actual data. The larger the likelihood value, the better the fit of the model to the data. The likelihood ratio in the "OmnibusTest" table is a function of the difference in likelihood values between the researcher's model and the model with the intercept only (the null model). If the likelihood ratio is significant, the researcher concludes that the coefficients in the model are different from 0 and the model is accepted (technically, the researcher rejects the null hypothesis that the coefficients are not different from 0).

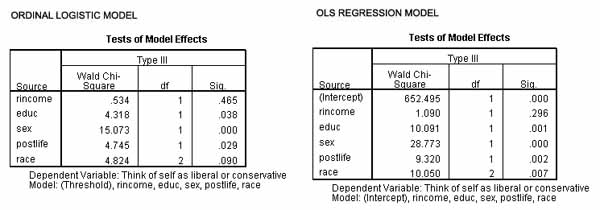
More generally, the likelihood ratio test may be used to compare the researcher's full model not only against the intercept-only model but against any nested model. The likelihood ratio statistic equals -2\*(log likelihood of the intercept-only or other reduced model minus the log-likelihood of the researcher's full model), where in SPSS the log likelihood is printed in the "Goodness of Fit" table. One could, for instance, run an intercept-only model, get its log-likelihood and note its degrees of freedom, then run one's full model; compute the difference in log-likelhood values times -2; look up this likelihood ratio value in a chi-square table using as degrees of freedom the difference in df between the two models; and get the same value as in the likelihood ratio test in the SPSS "Omnibus Test" table. However, the comparison nested model need not be the intercept-only model but can be any two models where one is nested within the other.

* + - * + *Deviance ratios*. In a well-fitting model, the ratio of the deviance value to its degrees of freedom should be close to 1. This ratio is shown in SPSS in the Goodness of Fit table, which also shows the Pearson chi-square/df ratio, an alternative statistic which should also be close to 1. By the same token, if deviance or chi-square ratios are high, it is usually a sign the model is misspecified and/or includes outlier data points with high residuals. If ratios are greater than 1 even after fixing problems of misspecification and outliers, such large ratios may also mean in the case of binomial and Poisson distributions of the dependent, that there is overdispersion (variance is greater than the mean) and a negative binomial model may fit better.
        + *Information criteria tests*. For AIC, AICC, BIC, and CAIC, a smaller value is a better fit when comparing models, which do not need to be nested. AICC is a correction to AIC for small sample sizes (it will be the same for large samples). BIC penalizes for model complexity as well as small sample size (AIC does not penalize for model complexity nor reward parsimony). CAIC (Consistent AIC) also penalizes for both sample size and model complexity (lack of parsimony). CAIC's penalty is greater than AIC but less than BIC. These measures are discussed at greater length in the section on [structural equation modeling](http://faculty.chass.ncsu.edu/garson/PA765/structur.htm#infotheory).

In the tables below, while both the ordinal logistic regression model and the OLS linear regression model are highly significant by the likelihood-ratio test, by the AIC criterion, the ordinal logistic model is somewhat better since the smaller value is the better fit.

* + - 1. 
      2. *Model summary statistics*: Also a default, for GZLM this option displays likelihood-ratio statistics for the model fit omnibus test (the table is labeled "Omnibus Test") and a test for each effect (the table is labeled "Tests of Model Effects"). The "Omnibus Test" table gives the likelihood ratio test for the researcher's model vs. the intercept-only model, as discussed above. The "Tests of Model Effects" table gives a test (by default, a Wald chi-square test using Type III sums of squares) for each model effect, with a corresponding significance level. If the significance > .05, the effect is not significant.

In the "Tests of Model Effects" tables below, one comes to generally the same substantive conclusions for both the ordinal logistic and linear regression models, with the exception that race is a significant predictor of political attitudes in the regression model but not in the ordinal logistic model, once other variables in the model are taken into account.

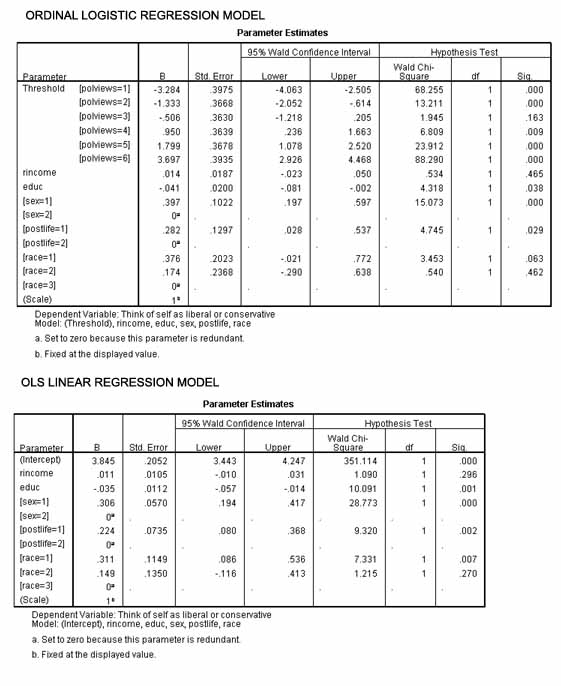


QIC and QICC. For GEE, the "Model summary statistics" option displays a "Goodness of Fit" table with QIC and QICC coefficients instead of GZLM's "Omnibus Test" table (the "Tests of Model Effects" table appears for both GZLM and GEE). QIC is the "quasi-likelihood under independence criterion" and QICC is a corrected version which penalizes for model complexity (that is, it rewards parsimony). Both are adaptations of AIC for repeated measures, where quasi log likelihood is used instead of log likelihood. As with AIC, smaller is better model fit for both QIC and QICC. QIC is used for choosing the best working correlation structure assumption in the "Repeated" tab (simply run the model for different working correlation structure assumptions, choose the assumption with the lowest QIC value). QICC is used for choosing the best subset of predictors (simply run a model and a nested model dropping one of the predictors, then compare QICC coefficients, with lower being better fit).

Warnings: Note that if Type I sums of squares are requested in the "Analysis Type" options of the Statistics tab, effect tests will be sensitive to the entry order of the predictors (unlike Type III, which is not sensitive to entry order). Regardless of Type I or III, the Wald test gives overly large standard errors for large coefficients, leading to Type 2 errors (accepting a false null hypothesis, concluding there is no relationship when there is one). For large coefficients, a likelihood ratio test is preferred (this is done as described [above](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#lratiotests) using a full model and a reduced model with one variable dropped). Finally, model coefficients are partial coefficients: they will change depending on which variables are predictors. Many recommend that non-significant variables should be dropped from the model one at a time until all remaining predictors are significant.

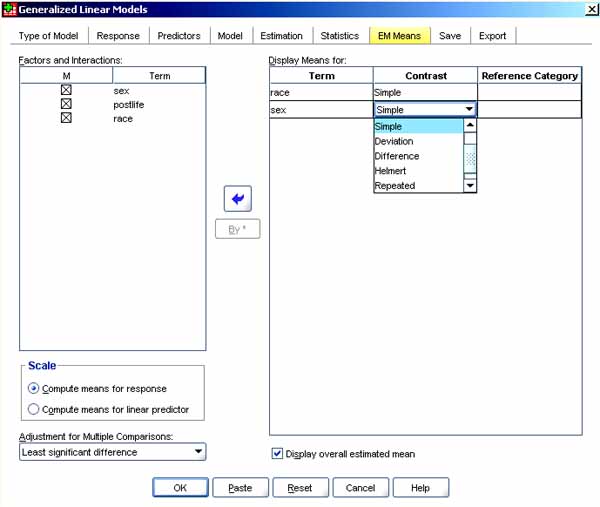
* + - 1. *Parameter estimates*: Also a default in SPSS, this option displays parameter estimates and corresponding test statistics and confidence intervals (the table is labeled "Parameter Estimates"). The "Parameter Estimates" table gives the actual model parameter for each effect (the b coefficient), the standard error of b, and the 95% Wald confidence intervals for b. The researcher may optionally display exponentiated parameter estimates in addition to the raw parameter estimates. In logistic models, for instance, exp(b) is the odds ratio, which is easier to interpret than raw b.

As the tables below illustrate, the significance levels for the covariates (here income and education) and the binary variables (sex, belief in the afterlife) are the same as reported above in the "Tests of Model Effects" table. However, the "Parameter Estimates" table also enables the researcher to assess the significance of particular values of the categorical predictors like race, which has three levels, in relation to the reference categories.



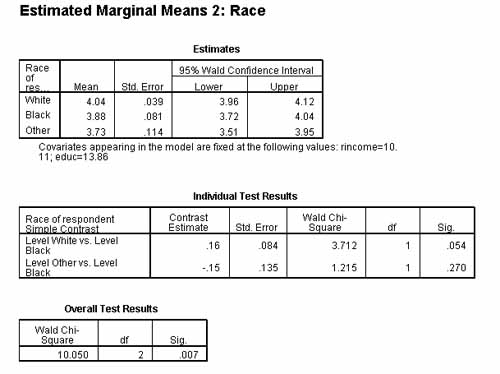
* + - * + *Parameter significance vs. effect significance*. Significance levels reported in the "Parameter Estimates" table also usually repeat significance levels reported in the "Test of Model Effects" table, but note the two tests do test different things and a variable effect may be significant while a corresponding parameter coefficient may be non-significant. If there is a difference, hypothesis-testing whether the effect of a variable is significantly different from 0 should use the significance levels reported in the "Test of Model Effects" table. The "Tests of Model Effects" table reports Type III (and Type I if requested) Wald chi-square tests for the null hypothesis that none of the parameter estimates (b coefficients) for a predictor are different from 0 (a finding of significance means that at least one of the parameter estimates is significant). In contrast, the Wald chi-square test in the "Parameter Estimates" table, if significant, means that that parameter is significantly different from 0. For categorical variables, the parameter for the reference category is not shown, nor is its significance, but the reference categories can be switched by changing the factor order using the Options button under the Predictors tab. For the variable gender, for instance, the parameter significance level for gender=1 with gender=0 as reference category might be non-significant while the parameter significance level for gender=0 with gender=1 as reference category might be significant. This can happen when the model includes interaction terms (ex., gender\*vote). The b coefficient parameters are partial coefficients, controlling for other terms in the model, so that in this example, the parameter for gender=1 may be non-significant after controlling for the interaction terms in the model. If the model only involves main effects, significance levels in the "Model Effects" table will correspond to significance levels in the "Parameter Estimates" table for binary predictors and covariates, but are not directly comparable for categorical predictors..
        + *Odds ratios*. When requested (to do so, check "Display exponentiated parameter estimates" under the "Statistics" tab), exp(B) - the odds ratio - is the last column in the "Parameter Estimates" table. Odds ratios are the model b coefficients raised to the power of the natural logarithm, It is this value which is commonly used to interpret the importance of the effects of covariates and of levels of factors used as predictors. Note that for levels of factors, the odds ratio must be interpreted in relation to the reference category (usually the highest-coded factor level). For example, if the odds ratio for the lowest level of a factor is .5, then we may say that a subject in that factor category has half the odds of those in the highest category. The odds referred to vary by model: for binary logistic regression, it is the odds of the event (odds the dependent = 1); for Poisson rate models, it is the percent of the predicted rate in the reference group. The closer the odds ratio is to 1.0, the more the independent variable's categories (ex., male and female for gender) don't matter and are independent of the dependent variable, with 1.0 representing full statistical independence. In the case of factor levels, the closer the odds ratio is to 1.0, the more similar that group is to the reference group for the given factor. The more the odds ratio is greater than 1.0 for covariates, the more that covariate increases the dependent, or if less than 1.0, decreases it. The more the odds ratio is greater than 1.0 for a level of a factor, the more membership in that factor group increases the odds compared to the reference group, or if less than 1.0, decreases it compared to the reference group. For further discussion of the interpretation of odds ratios, see the section on [logistic regression](http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm#odds).
      1. *Covariance/correlation matrices*: Not a default but if checked, the covariance and/or correlation matrices for the parameter estimates may be displayed.
      2. *Contrast coefficient matrices*: Not a default but if checked, this option displays contrast coefficients for the default effects and for the estimated marginal means, if requested on the EM Means tab and if available for the particular type of model the researcher has requested.
      3. *General estimable functions*: Not a default but if selected displays the matrices for generating the contrast coefficient (L) matrices.
      4. *Iteration history*: Not a default but if selected displays an iteration history table with parameter estimates for every nth iterations beginning with the 0th iteration (the initial estimates), where n is the value of the print interval set by the researcher in the dialog box in SPSS. The last iteration is always displayed regardless of n.
      5. *Lagrange multiplier test*: Not the default but if selected displays Lagrange multiplier test statistics for assessing the validity of a scale parameter that is computed using the deviance or Pearson chi-square, or set at a fixed number, for the normal, gamma, inverse Gaussian, and Tweedie distributions. For the negative binomial distribution, this tests the fixed ancillary parameter. Not available in GEE.
      6. *Working correlation matrix*: For GEE models only, this non-default option displays the values of the matrix representing within-subject dependencies. Its structure depends upon the specifications in the Repeated tab, as discussed above.

* + - **Contrast analysis**. The "EM Means" tab in both GZLM and GEE allows the research to select which factor levels and factor interactions are used for the display of estimated marginal means on the dependent variable. Covariates are not available for use. The overall estimated mean is also displayed. This allows comparision among groups formed by levels of the factors. EM Means is not available for ordinal multinomial models. Note some additional options are available in syntax (command) mode in SPSS, discussed further below.



* + - 1. *Factors and Interactions.*. This listbox on the left of the dialog screen contains the variables and interactions specified in the Predictors and Model tabs respectively, as discussed above. There is also a By \* button which can be used to add interactions even if not specifed under the Models tab (shift-select multiple variables to activate the By\* button, then click it).
      2. *Display Means For.*. This listbox on the right of the dialog screen lists the researcher-selected factors and factor interactions for which to display estimated means. That is, it is used for contrast analysis. One may click on the "Contrast" column header to select among the following types of contrast. The "Reference" column allows the reference category to be specified. .

The output below is for the OLS linear regression model, but was not available for the ordinal logistic regression model used for comparison in above. The OLS regression model treats the dependent variable, political views, as if its category codings 1 to 5 are interval numbers (not mulitnomial categories as in ordinal logistic regression). Thus the "Estimates" table gives the estimated means on political views for each race, since estimated marginal means were requested for race. The "Individual Test Results" table shows contrasts between each of the first two racial groups with the third reference category (by default the highest category, here black = 3). Neither contrast is significant for these data. The "Overall Test Results" table shows that race is nonetheless a significant predictor of political views in the linear regression model (the .007 significance level is the same as reported in the "Tests of Model Effects" table).



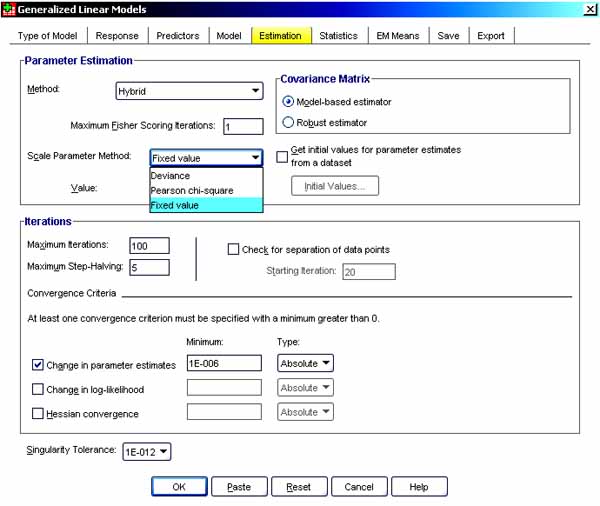
* + - * + *None*. No contrasts. No comparisons are computed for the given factor or interaction.
        + *Simple contrasts*. The mean of each factor level is compared to the mean of the reference level. In experimental research, the reference level represents the control group. This was the option used in the figure above.
        + *Pairwise contrasts*. Comparisons are computed for every pair of levels of the factors or implied factor for interactions. There is no reference category. "None" and "Pairwise" are the only available options for interactions.
        + *Deviation contrasts*. The mean of each factor level is compared to the grand mean. No reference category is specified.
        + *Difference contrasts*. The mean of each factor level except the first is compared to the mean of previous levels. No reference category is specified. Difference contrasts are also called reverse Helmert contrasts.
        + *Helmert contrasts*. Compares the mean of each level of the factor except the last to the mean of all subsequent levels. No reference category is specified.
        + *Repeated contrasts*. Compares the mean of each level of the factor except the last to the mean of the next subsequent level. No reference category is specified.
        + *Polynomial contrasts*. This type of contrast is used to estimate whether polynomial trends exist for the mean of the dependent across levels of the factor. In the output, the first degree of freedom represents a linear effect across all levels; the second degree of freedom represents a quadratic effect; etc. No reference category is specified.
      1. *Scale*. If the radio button for "Compute means for response" is selected (the default), estimated marginal means are for the dependent (response) variable are computed, based on the original scale of the dependent variable. If the radio button for "Compute means for linear predictor" is selected, estimated marginal means are computed for the linear predictor, based on the dependent variable as transformed by the link function.
      2. *Adjustment for Multiple Comparisons.*. Multiple comparisons involved in hypothesis-testing using multiple contrasts distort the actual significance level. In 20 contrasts, for instance, one would expect 1 to be a false positive at the usual .05 significance level. The following adjustments ("corrections") to the overall significance level are available: c
         * *Least significant difference*. No adjustment. This was the option used in the figure above.
         * *Bonferroni.*. The more comparisons, the higher the operational signficance level applied to maintain the nominal level. For instance, a .001 operational level might be needed to maintain the nominal .05 significance level. This method is commonly used but is very conservative (it is difficult to establish significant differences between levels because a stringent operational significance level may be required).
         * *Sequential Bonferroni*. This method modifies the Bonferroni algorithm to be much less conservative.
         * *Sidak*. An alternative to the Bonferroni adjustment, providing even tighter significance bounds than the Bonferroni.
         * *Sequential Sidak*. This method modifies the Sidak algorithm to be much less conservative.
      3. *Display overall estimated mean*. If checked (not the default), this checkbox causes the overall estimated mean to be printed along with the contrasts.

* + - **Subject variables and within-subjects variables**. If one has repeated measures, one must use the GEE module and specify the Subject variable(s). For instance, there may be an ID variable for each subject, or it may be that both a School ID and a Pupil ID variable are needed to specify subjects. Also, one must specify which variables are repeated measures over time or across clusters. These are the "Within Subjects" variables for which multiple measures are recorded for each subject, such as tests of memory over time for the same individuals. Note the Subject variable(s) and the Within Subject variable(s) cannot be used as the dependent (response) variable or as the trials variable if one has events-within-trials data. They can, however, be factors (predictors) in the model.
      1. *Data structure*. Repeated measures data in SPSS requires each subject have as many rows of data as there are time periods measured. The same subject ID is used for each row of data, with the time variable column having different values (ex., week = 1, 2, 3, etc.). In SPSS, the Data, Restructure menu choice may be useful in transforming data into this format if necessary. The time variable (ex., week) is a within-subjects variable since it applies to the same individual over repeated measurements.
      2. The "Repeated" tab of the GEE procedure also supports a variety of options for repeated measures:
         * *Sort cases by subject and within-subject variables*. This checkbox, which is checked by default, sorts cases properly. The only time one would uncheck this box is in the rare situation when (a) one wants to save a little computer processing time on a very large dataset by not declaring a within-subjects variable, and (b) cases are already sorted in the dataset so each subject's repeated measurements are properly ordered in a contiguous block of cases without missing rows (ex., not leaving out row 4 for subject 12 because subject 12 was not available for measurement at time 4). Not declaring a within-subject variable and relying just on data order will cause errors if there are missing observations/values. Declaring the within-subject variable(s) and leaving this checkbox checked is recommended.
         * *Covariance Matrix*. In SPSS one may choose "Robust Estimator" (the default) or "Model Based Estimator." The former is a corrected version of the latter which gives more consistent estimates of covariance even when the working correlation matrix is misspecified. See further discussion [below](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#cov_estimators).
         * *Working Correlation Matrix*. In this section of the "Repeated" tab dialog, the researcher specifies what assumptions should be made about the nature of the correlation matrix formed by the repeated measures (ex., the matrix formed by measurements of memory at Year1, Year2, etc., as both the rows and columns of the matrix). As estimation is iterative, for large samples parameter estimates will be consistent even if the structure of the working correlation matrix is misspecified. For small samples, parameter estimates are more likely to be affected by specification. One may select the best assumption for working correlation matrix by using the [QIC goodness of fit coefficient](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#qic), discussed below in the section on statistical output. Any of five possible assumptions may be specified:

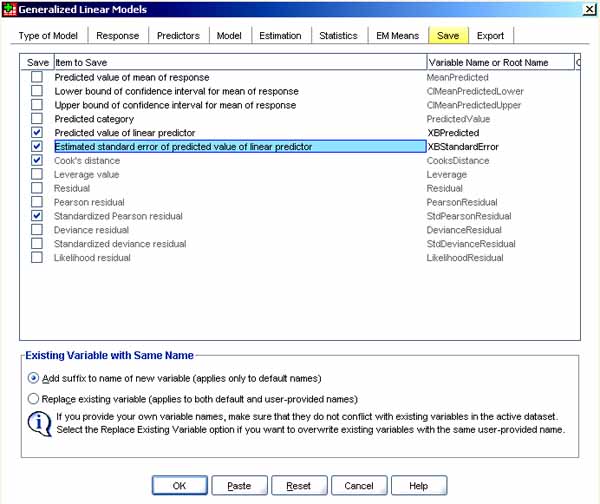
a. Independent: This assumes measurements for the repeated measure are uncorrelated. Independent is the default. The initial structure of the correlation matrix of within-subjects variables is assumed by the estimation algorithm to start with 1's on the diagonal and 0's off the diagonal, signifying variables are correlated with themselves at any given time but are not correlated with measurements at other times. Often this is a wrong assumption for repeated measures.   
b. AR(1): Repeated measures are assumed to have a first-order autoregressive relationship, which means current values are functions of prior values plus error. This assumes time series data with equal time intervals. The estimation algorithm assumes 1's on the diagonal, while off-diagonal elements are a function of time: measures one time period apart have a correlation of alpha; two periods apart have correlations of alpha-squared; three apart, alpha-cubed. In AR(1) models, correlation diminishes exponentially with time.   
c. Exchangeable: Correlations for within-subjects variables are homogenous, displaying compound symmetry. Thus the estimation algorithm assumes 1's on the diagonal and equal correlation for all off-diagonal elements. Correlation is assumed, therefore, not to change with time.   
d. M-dependent: Pairs of data elements separated by consecutive repeated measurements have a common correlation coefficient, pairs separated by a third data element have a common correlation coefficient, etc., up to m-1 intervening measurements. Measurements with m or greater intervening measurement are assumed uncorrelated. The researcher specifies m. Thus the algorithm assumes 1's on the diagonal, 0's for elements separated by m or more time units, and all others an equal correlation. That is, correlation does not change with time until time m, when it drops to zero.  
e. Unstructured: A completely general correlation matrix is assumed. That is, correlation estimates are estimated without constraints. While this is the obvious choice when the researcher cannot make any of the foregoing assumptions, estimates for this method may be unreliable for small samples.

* + - * + *Correlation estimator adjustment*. There is a checkbox for "Adjust estimator for number of nonredundant parameters. The default is checked. If unchecked, correlation estimates will be unaffected by subject-level replications in the data.
        + *Maximum iterations*. The researcher may specify the maximum number of iterations for the estimation algorithms for the linear part of the model. The default is 100.
        + *Update matrix*. This checkbox, which is checked by default, causes the estimation algorithms to use an updated working correlation matrix at each k iterations. By default k is 1 but the researcher may set the number. If unchecked, the initial working correlation matrix is used throughout.
        + *Convergence criteria*.There is a "Change in Parameter Estimates" checkbox, checked by default. When checked, estimation iterations stop when less than a specified value. The researcher may set if the value is absolute or relative change, and what the cutoff level is. Absolute change < .006 is the default.
        + *Hessian convergence.* This checkbox, unchecked by default, when checked, causes iterations to stop and convergence to be assumed if a statistic based on the Hessian is less than the value specified by the researcher, which is an absolute positive number. This may be used when convergence by default criteria fails.

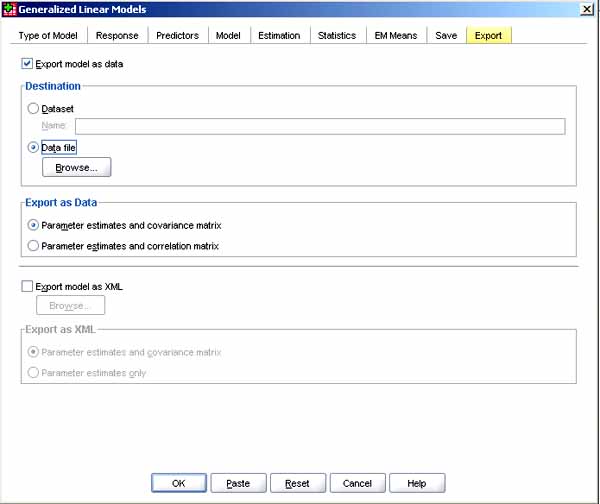
* + - **Estimation** methods and parameters may be set in SPSS under the "Estimation" tab, identical in GZLM and GEE except with respect to covariance matrices, discussed below..Often researchers accept the default settings for this tab.



* + - 1. *Parameter Estimation*. Model coefficients are estimated by maximum likelihood (ML) algorithms. However, iterations in ML can be accomplished by various methods. The "Parameter" tab's "Method" box provides for different method options, allowing selection among the Newton-Raphson, Fisher scoring, or the hybrid method. The hybrid method is to use Fisher scoring iterations which, if convergence is achieved before maximum iterations is reached, then switches to the Newton-Raphson method. If the dependent distribution is normal, "Method" is not settable.
      2. *Scale Parameter.* The scale parameter is an estimated model parameter related to the variance of the dependent variable. The "Scale parameter method" box allows one to estimate the scale parameter based on a selection among maximum-likelihood methods (the default, except not available if the dependent variable has a negative binomial, Poisson, binomial, or multinomial distribution). Alternative methods are deviance, Pearson chi-square, or a researcher-set fixed value the scale parameter. If there is a scale weight variable to weight cases, the scale parameter is divided by the scale weights for each observation. Cases with scale weight values that are less than or equal to 0 or are missing are not used in the analysis.
      3. *Initial values* if checked invokes a procedure to automatically compute initial values for parameters. Alternatively, the researcher can specify initial values for the parameter estimates.
      4. *Covariance Matrix*. This section of the "Estimation" tab dialog lets the researcher choose between model-based estimators and robust estimators of covariance/correlation when estimating the variance in GZLM models. Robust estimation is generally preferred as it provides consistent covariarnce estimates even when link functions are incorrectly modeled and/or variance is incorrectly specified. However, the robust method requires no missing data and is not recommended when the number of clusters is < 20. synonyms for the robust method are Huber/White, sandwich, and empirical estimation.
      5. *Iterations*. The researcher may also set the maximum number of iterations for the estimation algorithm while attempting to reach convergence (100 is default); the maximum step-halving (by default at each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached; or one can enter a positive integer other than 5).
      6. *Check for separation of data points*. When this checkbox is selected, the estimation algorithm tests that estimates of the dependent for each subject have unique values. That is, the algorithm checks for "separation," which is unique estimates for each data point. This option pertains to and is available for multinomial and binomial distributions of the dependent.
      7. *Convergence Criteria.*. Here the researcher can set the criterion for when the algorithm stops the cycle of iterations. The default is an absolute value of less than .006 change in parameter estimates. The researcher can set a different absolute value. Alternatively, one may specify a given minimum change in log-likelihood or in Hessian convergence.
      8. *Singularity tolerance*. When variables (columns) are linearly dependent, convergence may be impossible or unreliable. Matrices with this condition are "singular," evidenced by a very samll determinant. The default criterion for when matrices are singular is a determinant < 1e-12. the researcher may set a different positive value.
    - *Saving variables to one's dataset*. The "Save" tab in the SPSS GZLM/GEE dialog saves up to 14 variables for GZLM and 8 for GEE back to one's dataset. Note that the "Save" dialog uses specific variable names which, if present in the dataset, get overwritten unless the researcher uses suffixes to create unique names. The suffix-addition option is at the bottom of the "Save" dialog box.



* + - 1. *Residual analysis.* . Many of the variables which can be saved are useful for analysis of residuals and outliers. For instance, one can save predicted values (use "Predicted value of linear predictor" to maintain the link transform) and Pearson residuals, then plot residuals on the Y axis against predicted values on the X axis, perhaps getting one such plot for each time period if repeated measures apply or each factor group of interest (in SPSS menus, after saving, select Graphs, Legacy Dialogs, Scatter/Dot, Simple Scatter, Define, Panel by).
      2. *Variables available in GEE*
         * *Predicted value of mean of response*, saved as "MeanPredicted," saves model-predicted values of the dependent for each case or, in the case of binary dependent variables, saves predicted probabilities. For multinomial dependents, the item label becomes "Cumulative predicted probability," and the the cumulative predicted probability for each category of the dependent is saved, except the last.
         * *Lower bound of confidence interval for mean of response*, saved as "CMeanPredictedLower," is the lower bound of the confidence interval for the mean of the dependent variable for the given case.For multinomial dependents, the item label becomes "Lower bound of confidence interval for cumulative predicted probability," and the lower bound for each category of the response except the last is saved.
         * *Upper bound of confidence interval for mean of response*, saved as "CMeanPredicted," is the same thing for upper bounds.
         * *Predicted category*. saved as "PredictedValue," saves the dependent/response variable's predicted coding. This is available only for binary response/dependent variables or for multinomial dependents, in which case what is saved is the predicted response category for each case.
         * *Predicted value of linear predictor*, saved as "XBPredicted," saves the model-predicted values for each case in the as transformed by the link function specified under the Type of Model tab).
         * *Estimated standard error of predicted value of linear predictor*, sved as "XBStandardError," saves the standard error of the linear predictor (the dependent as transformed by the link function). If the dependent is multinomial, the estimated standard error for each category of the response except the last is saved.
         * *Residual*, saved as "Residual," is the raw residual: the observed value minus the value predicted by the model. Not available for multinomial dependents.
         * *Pearson residuals*, saved as "PearsonResidual," is the square root of the contribution of the given case to the Pearson chi-square statistic, with the sign of the raw residual. Not available for multinomial dependents.
      3. *Variables available in GZLM but not GEE*. However, even in GZLM, these statistics are not available if the dependent/response variables is multinomial:
         * *Cook's distance*, saved as "CooksDistance," is used for spotting outliers,Cook's distance measures how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. A large Cook's D indicates that the excluded case is particularly influential on residuals, which represent error.
         * *Leverage values*, saved as "Leverage," is an alternative measure for spotting outliers. It also measures the influence of a point on the fit of the regression. The centered leverage ranges from 0 (no influence) to (N-1)/N.
         * *Standardized Pearson residual*, saved as "StPearsonResidual," is the Pearson residual multiplied by the square root of the inverse of the product of the scale parameter and 1-leverage for the case.
         * *Deviance residual,* saved as "DevianceResidual," is the square root of the contribution of a case to the deviance statistic, with the sign of the raw residual.
         * *Standardized deviance residual,* saved as "StDevianceResidual,". is the deviance residual multiplied by the square root of the inverse of the product of the scale parameter and 1-leverage for the case.
         * *Likelihood residual*, saved as "LikelihoodResidual," is the square root of a weighted average (based on the leverage of the case) of the squares of the standardized Pearson and standardized Deviance residuals, with the sign of the raw residual.
    - *Exporting data*. The "Export" tab exports the user's model, consiisting of an SPSS dataset containing the columns listed below, in order. This matrix file can be used as the initial values for further model estimation. Not all procedures accept this data format. Even if the format is accepted, the researchre should be careful that all parameters have the same meaning for the procedure reading the file.



* + - 1. *Split variables*. Variables defining splits, if any, come first.
      2. *RowType*. This column has the codes COV (covariances), CORR (correlations), EST (parameter estimates), SE (standard errors), SIG (significance levels), and DF (sampling design degrees of freedom). That is, for each model parameter there is a separate case with row type COV (or CORR), plus a separate case for each of the other row types.
      3. *VarName\_*. This column takes values P1, P2, etc., corresponding to an ordered list of estimated model parameters (except the scale or negative binomial parameters), for row types COV or CORR, with value labels corresponding to the parameter strings shown in the Parameter estimates table. The cells are blank for other row types.
      4. *P1, P2, etc.* . These variables correspond to an ordered list of all model parameters (including the scale and negative binomial parameters, as appropriate), with variable labels corresponding to the parameter strings shown in the Parameter estimates table, and take values according to the row type.
      5. *Note:* For redundant parameters, all covariances are set to zero, correlations are set to the system-missing value; all parameter estimates are set at zero; and all standard errors, significance levels, and residual degrees of freedom are set to the system-missing value. For the scale parameter, covariances, correlations, significance level and degrees of freedom are set to the system-missing value. If the scale parameter is estimated via maximum likelihood, the standard error is given; otherwise it is set to the system-missing value. For the negative binomial parameter, covariances, correlations, significance level and degrees of freedom are set to the system-missing value. If the negative binomial parameter is estimated via maximum likelihood, the standard error is given; otherwise it is set to the system-missing value. If there are splits, then the list of parameters must be accumulated across all splits. In a given split, some parameters may be irrelevant; this is not the same as redundant. For irrelevant parameters, all covariances or correlations, parameter estimates, standard errors, significance levels, and degrees of freedom are set to the system-missing value.
      6. *Export model as XML*. Checking this checkbox saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. SmartScore and SPSS Server (a separate product) can use this model file to apply the model information to other data files for scoring purposes.
    - *Syntax (Command) mode options*. Some options in SPSS's GZLM/GEE procedures are available only in the syntax (command) window. In syntax, one may specify initial values for parameter estimates; fix covariates at values other than their means when computing estimated marginal means (using the EMMEANS subcommand); specify custom polynomial contrasts for estimated marginal means (using the EMMEANS subcommand); and specify a subset of the factors for which estimated marginal means are displayed to be compared using the specified contrast type (using the TABLES and COMPARE keywords of the EMMEANS subcommand).See also the GENLIN command syntax in SPSS.

**Assumptions**

* + - **Not assumed**. GZLM/GEE, compared to GLM, do not assume a normally distributed dependent variable (or normally distributed independents), nor linearity between the predictors and the dependent, nor homogeneity of variance for the range of the dependent variable.
    - **Linearity of the link function**. The researcher still must select a link function such that there is a linear relation between the linear predictor (the right-hand side of the model equation) and the link function of the dependent. For example, in logistic regression models, one must check to see if there is "linearity in the logit," meaning the linear predictor is in fact linearly related to the logit of the dependent.
    - **Absence of high multicollinearity**. As in other linear models, presence of high multicollinearity among the independents will inflate standard errors and confound interpretation of the relative contributions of the independents. See the discussion in the section on OLS regression, [here](http://www2.chass.ncsu.edu/garson/PA765/regress.htm#multicol), and at various points in the pages on [regression](http://faculty.chass.ncsu.edu/garson/PA765/regress.htm) and [logistic regression](http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm).
    - **Centered data**. As in OLS regression, centering may be necessary either to reduce multicollinearity or to make interpretation of coefficients meaningful. Centering is almost alway recommended for independent variables which are components of interaction terms in a logistic model. See the full discussion in the section on OLS regression, [here](http://www2.chass.ncsu.edu/garson/PA765/regress.htm#centering).
    - **Data distribution**. The independents may be of any distribution so long as linearity of the link function is maintained. The dependent may assume any of a wide variety of distributions, including normal, inverse normal (inverse Gaussian), binomial, multinomial, and Poisson. See further discussion [below](http://faculty.chass.ncsu.edu/garson/PA765/gzlm_gee.htm#exponential).
    - **Independent vs. correlated data** . In GZLM, observations are assumed to be independent. In GEE, observations are assumed to be independent between subjects within any given time period, cluster, or repeated measures factor, but are assumed to be dependent within the same subject across repeated measures factors. That is, GEE assumes between-subjects independence and within-subjects dependence. In practical terms, if there are repeated measures, GEE, not GZLM, should be chosen and the repeated measures properly specified.
    - **Data levels**. In both GZLM and GEE, the dependent (response) variable may be binary, counts, scale, or events-in-trials. "Scale" means interval or ordinal if it can be assumed that ordinal values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. For ordinal response variables, the model type should be multinomial (ordinal) logistic or ordinal probit. Factors are categorical. Covariates are scale variables. Weight and offset variables, if any, are also assumed to be scale. In GEE, the variables specified as repeated measures within-subjects effects, or the variables used to define subjects, cannot be used as dependent variables.
    - **Missing data**. As in other forms of statistical analysis, missing data can lead to biased coefficients unless data are missing completely at random. [Missing values analysis](http://faculty.chass.ncsu.edu/garson/PA765/missing.htm) is discussed in a separate section.

**Frequently Asked Questions**

* + - *Should I use GRZ or GZLM as the abbreviation for generalized linear models?*

As of late 2007, Google searching suggested that GZLM was preferred 7:1 over GRZ, though this is a matter of preference.

* + - *Can the dependent/response variable in GZLM or GEE be distributed in any manner? Are these distribution-free models?*

No on both counts. The distribution must be specified, and it must be a member of the exponential family of distributions, which include normal, binomial, inverse Gaussian (Wald, inverse normal), gamma, Poisson, Tweedie, and multinomial. In addition, the negative binomial distribution can be estimated and may be used in GZLM/GEE models. There are other members of the exponential family not supported by GSLM/GEE in SPSS as of SPSS 16, such as the Dirichlet, Bernoulli, and Weibull distributions. Also not supported and not compatible with the generalized linear model are such non-exponential distributions as the Cauchy, Laplace, and uniform families of distributions.

* + - *What is a "canonical link"?*

Each distribution in the exponential families has a link function which is sufficient to properly map the linear predictor (the right-hand part of the model equation) onto the domain to the dependent/response variable. This does not mean it is the only possible link function for the distribution with which it is associated, only that it is known to be a sufficient one. Here are common distributions and their corresponding canonical link functions:

* + - * + Normal distribution, identity link function
        + Binomial distribution, logit link function
        + Negative binomial distribution, negative binomial link function
        + Poisson distribution, log link function
        + Inverse normal (inverse Gaussian) distribution, power link function where power = -2
        + Gamma distribution, power link function where power = -1 - that is, an inverse (reciprocal) link function.

**Bibliography**

* + - Agresti, A. (1996). *An introduction to categorical data analysis.* John Wiley & Sons: New York. Perhaps the most-cited introductory text.
    - Atkins, D. C., & Gallop, R. J. (2007). Re-thinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology*, 21, 726-735. On poisson regression models.
    - Chard, Cristina Ling & Arango, Diana J. (2003). *Annual review of world bank staff learning FY03.* Example of using gamma regression, with comparisons to logistic and OLS regression. WBI Evaluation Studies No. EG04-74. Washington, DC: The World Bank Institute. December, 2003Retrieved 3/20/08 from http://info.worldbank.org/etools/library/latestversion.asp?80321.
    - Gilchrist, R. & Drinkwater, D. (2000). The use of the Tweedie distribution in statistical modeling. Pp. 313-318 in Bethlehem, Jelke G. & van der Heijden, Peter G. M., eds. *Proceedings in computational statistics 2000*. Heidelberg, Germany: Physica- Verlag.
    - Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions (2nd ed.)*. College Station, TX: StataCorp LP.
    - Hilbe, Joseph M. (2007). *Negative binomial regression*. New York: Cambridge University Press.
    - Hoffman, J. P. (2004). *Generalized linear models: An applied approach*. Boston: Allyn & Bacon. An accessible extended introduction.
    - McCullagh, P. & J.A. Nelder (1989). *Generalized linear models.* Chapman and Hall: London. A mathematical treatment.
    - Nelder, J. A. & Wedderburn, R. W. N. (1972). Generalized linear models. *Journal of the Royal Statistical Society* 135: 370-384. The seminal article for GZLM.
    - Rogers, W. H. (1993). Comparison of nbreg and glm for negative binomial. *Stata Technical Bulletin* 16: 7.

@c 2006, 2007, 2009 G. David Garson  
Last updated 6/26/2009.

Top of Form

Bottom of Form