Chapter 326

Negative Binomial Regression

Introduction

Negative binomial regression is similar to regular multiple regression except that the dependent (*Y*) variable is an observed count that follows the negative binomial distribution. Thus, the possible values of *Y* are the nonnegative integers: 0, 1, 2, 3, and so on.

Negative binomial regression is a generalization of Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial regression model, commonly known as NB2, is based on the Poisson-gamma mixture distribution. This formulation is popular because it allows the modelling of Poisson heterogeneity using a gamma distribution.

Some books on regression analysis briefly discuss Poisson and/or negative binomial regression. We are aware of only a few books that are completely dedicated to the discussion of count regression (Poisson and negative binomial regression). These are Cameron and Trivedi (2013) and Hilbe (2014). Most of the results presented here were obtained from these books.

This program computes negative binomial regression on both numeric and categorical variables. It reports on the regression equation as well as the goodness of fit, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform a subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values.

The Negative Binomial Distribution

The Poisson distribution may be generalized by including a gamma noise variable which has a mean of 1 and a scale parameter of ν. The Poisson-gamma mixture (negative binomial) distribution that results is

$$\Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}$$

where

$$\mu_i = t_i\mu$$

$$\alpha = \frac{1}{\nu}$$

The parameter μ is the mean incidence rate of y per unit of exposure. Exposure may be time, space, distance, area, volume, or population size. Because exposure is often a period of time, we use the symbol t_i to represent the exposure for a particular observation. When no exposure given, it is assumed to be one.

The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, t.

The results below make use of the following relationship derived from the definition of the gamma function

$$\ln\left(\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right)=\sum_{j=0}^{y_i-1}\ln(j+\alpha^{-1})$$

The Negative Binomial Regression Model

In negative binomial regression, the mean of *y* is determined by the exposure time *t* and a set of *k* regressor variables (the *x*'s). The expression relating these quantities is

$$\mu_i = exp(\ln(t_i)+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki})$$

Often, $x_1 \equiv 1$, in which case β_1 is called the *intercept*. The regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$ are unknown parameters that are estimated from a set of data. Their estimates are symbolized as $b_1, b_2, \ldots, b_k$.

Using this notation, the fundamental negative binomial regression model for an observation *i* is written as

$$\Pr(Y=y_i|\mu_i,\alpha)=\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

Solution by Maximum Likelihood Estimation

The regression coefficients are estimated using the method of maximum likelihood. Cameron (2013, page 81) gives the logarithm of the likelihood function as

$$\mathcal{L}=\sum_{i=1}^{n}\{\ln[\Gamma(y_i+\alpha^{-1})]-\ln[\Gamma(\alpha^{-1})]-\ln[\Gamma(y_i+1)]-\alpha^{-1}\ln(1+\alpha\mu_i)-y_i\ln(1+\alpha\mu_i)+y_i\ln(\alpha)+y_i\ln(\mu_i)\}$$

Rearranging gives

$$\mathcal{L}=\sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1}\ln(j+\alpha^{-1})\right)-\ln\big(\Gamma(y_i+1)\big)-(y_i+\alpha^{-1})\ln(1+\alpha\mu_i)+y_i\ln(\mu_i)+y_i\ln(\alpha)\right\}$$

The first derivatives of $\mathcal{L}$ were given by Cameron (2013) and Lawless (1987) as

$$\frac{\partial\mathcal{L}}{\partial\beta_j}=\sum_{i=1}^{n}\frac{x_{ij}(y_i-\mu_i)}{1+\alpha\mu_i},\qquad j=1,2,\ldots,k$$

$$\frac{\partial\mathcal{L}}{\partial\alpha}=\sum_{i=1}^{n}\left\{\alpha^{-2}\left(\ln(1+\alpha\mu_i)-\sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}}\right)+\frac{y_i-\mu_i}{\alpha(1+\alpha\mu_i)}\right\}$$

$$\frac{-\partial^2\mathcal{L}}{\partial\beta_r\partial\beta_s}=\sum_{i=1}^{n}\frac{\mu_i(1+\alpha y_i)x_{ir}x_{is}}{(1+\alpha\mu_i)^2},\qquad r,s=1,2,\ldots,k$$

$$\frac{-\partial^2\mathcal{L}}{\partial\beta_r\partial\alpha}=\sum_{i=1}^{n}\frac{\mu_i(y_i-\mu_i)x_{ir}}{(1+\alpha\mu_i)^2},\qquad r=1,2,\ldots,k$$

$$\frac{-\partial^2\mathcal{L}}{\partial\alpha^2}=\sum_{i=1}^{n}\left\{\sum_{j=0}^{y_i-1}\left(\frac{j}{1+\alpha j}\right)^2+2\alpha^{-3}\ln(1+\alpha\mu_i)-\frac{2\alpha^{-2}\mu_i}{1+\alpha\mu_i}-\frac{(y_i+\alpha^{-1})\mu_i^2}{(1+\alpha\mu_i)^2}\right\}$$

Equating the gradients to zero gives the following set of likelihood equations

$$\sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1 + \alpha\mu_i} = 0, \qquad j = 1, 2, \ldots, k$$

$$\sum_{i=1}^{n} \left\{ \alpha^{-2} \left(\ln(1 + \alpha\mu_i) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_i - \mu_i}{\alpha(1 + \alpha\mu_i)} \right\} = 0$$

Distribution of the MLE's

Cameron (2013) gives the asymptotic distribution of the maximum likelihood estimates as multivariate normal as follows

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\alpha} \end{bmatrix} \sim \mathrm{N}\left[\begin{bmatrix} \boldsymbol{\beta} \\ \alpha \end{bmatrix} \middle| \begin{matrix} \mathrm{V}(\hat{\boldsymbol{\beta}}) & \mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\alpha}) \\ \mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\alpha}) & \mathrm{V}(\hat{\alpha}) \end{matrix} \right]$$

where

$$\mathrm{V}(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \frac{\mu_i}{1 + \alpha\mu_i} \mathbf{x_i} \mathbf{x_i'} \right]^{-1}$$

$$\mathrm{V}(\hat{\alpha}) = \sum_{i=1}^{n} \left\{ \alpha^{-4} \left(\ln(1 + \alpha\mu_i) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right)^2 + \frac{\mu_i}{\alpha^2(1 + \alpha\mu_i)} \right\}^{-1}$$

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \alpha) = [\mathbf{0}]$$

Deviance

The deviance is twice the difference between the maximum achievable log-likelihood and the log-likelihood of the fitted model. In multiple regression under normality, the deviance is the residual sum of squares. In the case of negative binomial regression, the deviance is a generalization of the sum of squares. The maximum possible log likelihood is computed by replacing μ_i with y_i in the likelihood formula. Thus, we have

$$\begin{aligned} D &= 2[\mathcal{L}(y_i) - \mathcal{L}(\mu_i)] \\ &= 2 \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{y_i}{\mu_i}\right) - (y_i + \alpha^{-1}) \ln\left(\frac{1 + \alpha y_i}{1 + \alpha\mu_i}\right) \right\} \end{aligned}$$

Akaike Information Criterion (AIC)

Hilbe (2014) mentions the Akaike Information Criterion (AIC) as one of the most commonly used fit statistics. It has two formulations:

$$AIC(1) = -2[\mathcal{L} - k]$$

and

$$AIC(n) = -\frac{2}{n}[\mathcal{L} - k]$$

Note that *k* is the number of predictors including the intercept.

AIC(1) is usually output by statistical software applications.

Bayesian Information Criterion (BIC)

Hilbe (2014) also mentions the Bayesian Information Criterion (BIC) as another common fit statistic. It has three formulations:

$$BIC(R) = D - (df)\ln(n)$$
$$BIC(L) = -2\mathcal{L} + k\ln(n)$$
$$BIC(Q) = -\frac{2}{n}(\mathcal{L} - k\ln(k))$$

Note that *df* is the residual degrees of freedom.

Note that *BIC*(*L*) is given as SC in SAS and simply *BIC* in other software.

Residuals

As in any regression analysis, a complete residual analysis should be employed. This involves plotting the residuals against various other quantities such as the regressor variables (to check for outliers and curvature) and the response variable.

Raw Residual

The raw residual is the difference between the actual response and the value estimated by the model. Because in this case, we expect that the variances of the residuals to be unequal, there are difficulties in the interpretation of the raw residuals. However, they are still popular. The formula for the raw residual is

$$r_i = y_i - \hat{\mu}_i$$

Pearson Residual

The Pearson residual corrects for the unequal variance in the residuals by dividing by the standard deviation of *y*. The formula for the Pearson residual is

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \alpha\hat{\mu}_i^2}}$$

Anscombe Residual

The Anscombe residual is another popular residual that is close to a standardized deviance residual. It normalizes the raw residual so that heterogeneity and outliers can be quickly identified. Its formula is

$$a_i = \frac{\frac{3}{\alpha}\left\{(1+\alpha y_i)^{2/3} - (1+\alpha\hat{\mu}_i)^{2/3}\right\} + 3\left(y_i^{2/3} - \hat{\mu}_i^{2/3}\right)}{2\left(\hat{\mu}_i + \alpha\hat{\mu}_i^2\right)^{1/6}}$$

Subset Selection

Subset selection refers to the task of finding a small subset of the available regressor variables that does a good job of predicting the dependent variable. Because negative binomial regression must be solved iteratively, the task of finding the best subset can be time consuming. Hence, techniques which look at all possible combinations of the regressor variables are not feasible. Instead, algorithms that add or remove a variable at each step are used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

Categorical Variables

An issue that often comes up during data analysis is what to do about the binary variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. With two or three categorical variables, a large number of binary variables result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms used here search on model terms rather than on the individual variables. Thus, the whole set of binary variables associated with a given term is considered together for inclusion in, or deletion from, the model. It is all or none. If you want the subset algorithm to deal with them individually, you can generate the set of binary variables manually and include them individually as Numeric Variables.

Hierarchical Models

Another practical modelling issue is how to handle interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to consider only hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the best (closest to zero) log-likelihood. Enter this term into the model.
3. Continue adding terms until a preset limit on the maximum number of terms in the model is reached.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if any have a more attractive log-likelihood. If a switch is found, it is made and the candidate terms are again searched to determine if another switch can be made.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, set the maximum subset size to a large value such as 10. By studying the Subset Selection reports, you can quickly determine an optimum number of terms. Reset the maximum subset size to this number and make a second run. This two-step procedure works better than relying on some F-to-enter and F-to-remove tests whose properties are not well understood to begin with.

Data Structure

At a minimum, datasets to be analyzed by negative binomial regression must contain a dependent variable and one or more independent variables. For each categorical variable, the program generates a set of binary (0 and 1) variables. For example, in the table below, the discrete variable AgeGroup will be replaced by the variables Ag2 through Ag6 (Ag1 is redundant).

Koch et. al. (1986) present the following data taken from the Third National Cancer Survey. This dataset contains the number of new melanoma cases in 1969-1971 among white males in two areas for various age groups. The size of the estimated population at risk is given in the variable Population.

Koch36 dataset

Melanoma	Area	AgeGroup	Population	AG1	AG2	AG3	AG4	AG5	AG6
61	0	<35	2880262	1	0	0	0	0	0
76	0	35-44	564535	0	1	0	0	0	0
98	0	45-54	592983	0	0	1	0	0	0
104	0	54-64	450740	0	0	0	1	0	0
63	0	65-74	270908	0	0	0	0	1	0
80	0	>74	161850	0	0	0	0	0	1
64	1	<35	1074246	1	0	0	0	0	0
75	1	35-44	220407	0	1	0	0	0	0
68	1	45-54	198119	0	0	1	0	0	0
63	1	54-64	134084	0	0	0	1	0	0
45	1	65-74	70708	0	0	0	0	1	0
27	1	>74	34233	0	0	0	0	0	1

Missing Values

If missing values are found in any of the independent variables being used, the row is omitted. If only the value of the dependent variable is missing, that row will not be used during the estimation process, but its predicted value will be generated and reported on.

Procedure Options

This section describes the options available in this procedure.

Variables, Model Tab

This panel specifies the variables and model are used in the analysis.

Variables

Dependent Y

Specify the dependent (response) variable. This is the variable to be predicted by the independent variables. The values in this variable should be non-negative integers (zero is okay).

Exposure T

Specify an optional variable containing exposure values. If this option is left blank, all exposures will be set to 1.0. This variable is specified when the exposures are different for each row.

The exposure is the amount of time, space, distance, volume, or population size from which the dependent variable is counted. For example, exposure may be the time in days, months, or years during which the values on that row were obtained. It may be the number of individuals at risk or the number of man-years from which the dependent variable is measured.

Each exposure value must be a positive (non-zero) number. Otherwise the row is ignored during the estimation phase.

Numeric X's

Specify the numeric (continuous) independent variables. By numeric, we mean that the values are numeric and at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are better analyzed when you specify them as Categorical Independent Variables.

If you want to create powers and cross-products of these variables, specify an appropriate model in the 'Custom Model' field under the Model tab.

If you want to create predicted values of *Y* for values of *X* not in your database, add the *X* values to the bottom of the database. They will not be used during estimation, but predicted values will be generated for them.

Categorical X's

Specify categorical (nominal or group) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories.

Regression analysis is only defined for numeric variables. Since categorical variables are nominal, they cannot be used directly in regression. Instead, an internal set of numeric variables must be substituted for each categorical variable.

Suppose a categorical variable has *G* categories. ***NCSS*** automatically generates the *G*-1 internal, numeric variables for the analysis. The way these internal variables are created is determined by the Recoding Scheme and, if needed, the Reference Value. These options can be entered separately with each categorical variable, or they can specified using a default value (see Default Recoding Scheme and Default Reference Value below).

The syntax for specifying a categorical variable is *VarName*(*CType; RefValue*) where *VarName* is the name of the variable, CType is the recoding scheme, and RefValue is the reference value, if needed.

CType

The recoding scheme is entered as a letter. Possible choices are B, P, R, N, S, L, F, A, 1, 2, 3, 4, 5, or E. The meaning of each of these letters is as follows.

- **B** for **binary** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	B1	B2	B3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

- **P** for **Polynomial** of up to 5th order (you cannot use this option with category variables with more than 6 categories.
  Example: Categorical variable Z with 4 categories.

Z	P1	P2	P3
1	-3	1	-1
3	-1	-1	3
5	1	-1	-3
7	3	1	1

- **R** to compare each with the **reference value** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **N** to compare each with the **next** category.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	0	0
3	-1	1	0
5	0	-1	1
7	0	0	-1

- **S** to compare each with the **average of all subsequent** values.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-3	0	0
3	1	-2	0
5	1	1	-1
7	1	1	1

- **L** to compare each with the **prior** category.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-1	0	0
3	1	-1	0
5	0	1	-1
7	0	0	1

- **F** to compare each with the **average of all prior** categories.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	1	1
3	1	1	-1
5	1	-2	0
7	-3	0	0

- **A** to compare each with the **average of all** categories (the Reference Value is skipped).
  Example: Categorical variable Z with 4 categories. Suppose the reference value is 3.

Z	S1	S2	S3
1	-3	1	1
3	1	1	1
5	1	-3	1
7	1	1	-3

- **1** to compare each with the **first** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	-1	-1	-1
B	1	0	0
C	0	1	0
D	0	0	1

- **2** to compare each with the **second** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	-1	-1	-1
C	0	1	0
D	0	0	1

- **3** to compare each with the **third** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	-1	-1	-1
D	0	0	1

- **4** to compare each with the **fourth** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **5** to compare each with the **fifth** category after sorting.
  Example: Categorical variable Z with 5 categories.

Z	C1	C2	C3	C4
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1
E	-1	-1	-1	-1

- **E** to compare each with the **last** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

RefValue
A second, optional argument is the reference value. The reference value is one of the categories. The other categories are compared to it, so it is usually a baseline or control value. If neither a baseline or control value is evident, the reference value is the most frequent value.

For example, suppose you want to include a categorical independent variable, State, which has four values: Texas, California, Florida, and NewYork. Suppose the recoding scheme is specified as *Compare Each with Reference Value* with the reference value of *California*. You would enter

State(R;California)

Default Recoding Scheme

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. The options are outlined in the sections below.

The contrast type may also be designated within parentheses after the name of each categorical independent variable, in which case the default contrast type is ignored.

If your model includes interactions of categorical variables, this option should be set to 'Contrast with Reference'or Compare with All Subsequent' in order to match GLM results for factor effects.

- **Binary** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	B1	B2	B3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

- **Polynomial** of up to 5th order (you cannot use this option with category variables with more than 6 categories.
  Example: Categorical variable Z with 4 categories.

Z	P1	P2	P3
1	-3	1	-1
3	-1	-1	3
5	1	-1	-3
7	3	1	1

- **Compare Each with Reference Value** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **Compare Each with Next**.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	0	0
3	-1	1	0
5	0	-1	1
7	0	0	-1

- **Compare Each with All Subsequent**.
  Example: Categorical variable Z with 4 categories.
  Z S1 S2 S3
  1 -3 0 0
  3 1 -2 0
  5 1 1 -1
  7 1 1 1

- **Compare Each with Prior**
  Example: Categorical variable Z with 4 categories.
  Z S1 S2 S3
  1 -1 0 0
  3 1 -1 0
  5 0 1 -1
  7 0 0 1

- **Compare Each with All Prior**
  Example: Categorical variable Z with 4 categories.
  Z S1 S2 S3
  1 1 1 1
  3 1 1 -1
  5 1 -2 0
  7 -3 0 0

- **Compare Each with Average**
  Example: Categorical variable Z with 4 categories. Suppose the reference value is 3.
  Z S1 S2 S3
  1 -3 1 1
  3 1 1 1
  5 1 -3 1
  7 1 1 -3

- **Compare Each with First**
  Example: Categorical variable Z with 4 categories.
  Z C1 C2 C3
  A -1 -1 -1
  B 1 0 0
  C 0 1 0
  D 0 0 1

- **Compare Each with Second**
  Example: Categorical variable Z with 4 categories.
  Z C1 C2 C3
  A 1 0 0
  B -1 -1 -1
  C 0 1 0
  D 0 0 1

- **Compare Each with Third**
  Example: Categorical variable Z with 4 categories.
  Z C1 C2 C3
  A 1 0 0
  B 0 1 0
  C -1 -1 -1
  D 0 0 1

- **Compare Each with Fourth**
  Example: Categorical variable Z with 4 categories.
  Z C1 C2 C3
  A 1 0 0
  B 0 1 0
  C 0 0 1
  D -1 -1 -1

- **Compare Each with Fifth**
  Example: Categorical variable Z with 5 categories.
  Z C1 C2 C3 C4
  A 1 0 0 0
  B 0 1 0 0
  C 0 0 1 0
  D 0 0 0 1
  E -1 -1 -1 -1

- **Compare Each with Last**
  Example: Categorical variable Z with 4 categories.
  Z C1 C2 C3
  A 1 0 0
  B 0 1 0
  C 0 0 1
  D -1 -1 -1

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. The choices are

- **First Value after Sorting – Fifth Value after Sorting**
  Use the first (through fifth) value in alpha-numeric sorted order as the reference value.

- **Last Value after Sorting**
  Use the last value in alpha-numeric sorted order as the reference value.

Frequencies

This is an optional variable containing the frequency (observation count) for each row. Usually, you would leave this option blank and let each row receive the default frequency of one.

If your data have already been summarized, this option lets you specify how many actual rows each physical row represents.

Dispersion Parameter (α)

α Input Type

Select the type of input you would like to use to specify the dispersion parameter α.

The choices are:

Estimate α from the data

α is estimated from the data using maximum likelihood. This will also provide an estimate of its standard error.

Enter α directly

Enter a fixed value for α. You might do this for one of two reasons:

1. To speed up the convergence once a reasonable value of α has been determined.

2. To force the fit of a geometric regression model which is the name of the special case in which α is equal to one.

α (Dispersion)

The dispersion parameter α specifies the amount of overdispersion in a Poisson-gamma mixture model (commonly called the negative binomial regression model). It is added to enhance a Poisson regression model with a more flexible specification of the variance (since the Poission distribution forces the unrealistic assumption that the variance is equal to the mean).

Range

α must be greater than zero and is usually less than 4.

Why specify α

1. Maximum likelihood estimation may not converge if α is near zero.

2. You might use this option to shorten the runtime for large datasets for which a previous run gives an estimate of α.

3. Geometric Regression is a special case of negative binomial regression in which α is set to one.

Regression Model

Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a straight-forward regression model, select *1-Way*.

The options are

- **Up to 1-Way**

  This option generates a model in which each variable is represented by a single model term. No cross-products, interactions, or powers are added. Use this option when you want to use the variables you have specified, but you do not want to generate other terms.

  This is the option to select when you want to analyze the independent variables specified without adding any other terms.

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C

- **Up to 2-Way**
  This option specifies that all individual variables, two-way interactions, and squares of numeric variables are included in the model. For example, if you have three numeric variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*A + B*B + C*C

  On the other hand, if you have three categorical variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C

- **Up to 3-Way**
  All individual variables, two-way interactions, three-way interactions, squares of numeric variables, and cubes of numeric variables are included in the model. For example, if you have three numeric, independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C + A*A + B*B + C*C + A*A*B + A*A*C + B*B*C +A*C*C + B*C*C

  On the other hand, if you have three categorical variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

- **Up to 4-Way**
  All individual variables, two-way interactions, three-way interactions, and four-way interactions are included in the model. Also included would be squares, cubes, and quartics of numeric variables and their cross-products.

  For example, if you have four categorical variables A, B, C, and D, this would generate the model:

  A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + A*C*D + B*C*D + A*B*C*D

- **Interaction**
  Mainly used for categorical variables. A saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-variable combinations (you can have unequal numbers of observations for each combination of the categorical variables). No squares, cubes, etc. are generated.

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

  Note that the discussion of the Custom Model option discusses the interpretation of this model.

- **Custom Model**
  The model specified in the *Custom Model* box is used.

Remove Intercept

Unchecked indicates that the intercept term, β_0, is to be included in the regression. Checked indicates that the intercept should be omitted from the regression model. Note that deleting the intercept distorts most of the diagnostic statistics. In most situations, you should include the intercept in the model.

Replace Custom Model with Preview Model (button)

When this button is pressed, the Custom Model is cleared and a copy of the Preview model is stored in the Custom Model. You can then edit this Custom Model as desired.

Maximum Order of Custom Terms

This option specifies that maximum number of variables that can occur in an interaction (or cross-product) term in a custom model. For example, A*B*C is a third order interaction term and if this option were set to 2, the A*B*C term would not be included in the model.

This option is particularly useful when used with the bar notation of a custom model to allow a simple way to remove unwanted high-order interactions.

Custom Model

This options specifies a custom model. It is only used when the *Terms* option is set to *Custom*. A custom model specifies the terms (single variables and interactions) that are to be kept in the model.

Interactions

An interaction expresses the combined relationship between two or more variables and the dependent variable by creating a new variable that is the product of the variables. The interaction between two numeric variables is generated by multiplying them. The interaction between to categorical variables is generated by multiplying each pair of indicator variables. The interaction between a numeric variable and a categorical variable is created by generating all products between the numeric variable and the indicator variables generated from the categorical variable.

Syntax

A model is written by listing one or more terms. The terms are separated by a blank or plus sign. Terms include variables and interactions. Specify regular variables (main effects) by entering the variable names. Specify interactions by listing each variable in the interaction separated by an asterisk (*), such as Fruit*Nuts or A*B*C.

You can use the bar (|) symbol as a shorthand technique for specifying many interactions quickly. When several variables are separated by bars, all of their interactions are generated. For example, A|B|C is interpreted as A + B + C + A*B + A*C + B*C + A*B*C.

You can use parentheses. For example, A*(B+C) is interpreted as A*B + A*C.

Some examples will help to indicate how the model syntax works:

A|B = A + B + A*B

A|B A*A B*B = A + B + A*B + A*A + B*B

Note that you should only repeat numeric variable. That is, A*A is valid for a numeric variable, but not for a categorical variable.

A|A|B|B (Max Term Order=2) = A + B + A*A + A*B + B*B

A|B|C = A + B + C + A*B + A*C + B*C + A*B*C

(A + B)*(C + D) = A*C + A*D + B*C + B*D

(A + B)|C = (A + B) + C + (A + B)*C = A + B + C + A*C + B*C

Subset Selection

Search Method

This option specifies the subset selection algorithm used to reduce the number of independent variables that used in the regression model. Note that since the solution algorithm is iterative, the selection process can be very time consuming. The Forward algorithm is much quicker than the Forward with Switching algorithm, but the Forward algorithm does not usually find as good of a model.

Also note that in the case of categorical independent variables, the algorithm searches among the original categorical variables, not among the generated individual binary variables. That is, either all binary variables associated with a particular categorical variable are included or not—they are not considered individually.

Hierarchical models are such that if an interaction is in the model, so are the terms that can be derived from it. For example, if A*B*C is in the model, so are A, B, C, A*B, A*C, and B*C. Statisticians usually adopt hierarchical models rather than non-hierarchical models. The subset selection procedure can be made to consider only hierarchical models during its search.

The subset selection options are:

- **None – No Search is Conducted**

  No subset selection is attempted. All specified independent variables are used in the regression equation.

- **(Hierarchical) Forward**

  With this algorithm, the term with the largest log likelihood is entered into the model. Next, the term that increases the log likelihood the most is added. This selection is continued until all the terms have been entered or until the maximum subset size has been reach.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for selection. For example, the interaction term A*B will not be considered unless both A and B are already in the model.

  When using this algorithm, you must make one run that allows a large number of terms to find the appropriate number of terms. Next, a second run is made in which you decrease the maximum terms in the subset to the number after which the log likelihood does not change significantly.

- **(Hierarchical) Forward with Switching**

  This algorithm is similar to the Forward algorithm described above. The term with the largest log likelihood is entered into the regression model. The term which increases the log likelihood the most when combined with the first term is entered next. Now, each term in the current model is removed and the rest of the terms are checked to determine if, when they are used instead, the likelihood function is increased. If a term can be found by this switching process, the switch is made and the whole switching operation is begun again. The algorithm continues until no term can be found that improves the likelihood. This model then becomes the best two-term model.

  Next, the subset size is increased by one, the best third term is entered into the model, and the switching process is repeated. This process is repeated until the maximum subset size is reached. Hence, this model finds the optimum subset for each subset size. You must make one run to find an appropriate subset size by looking at the change in the log likelihood. You then reset the maximum subset size to this value and rerun the analysis.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for addition or deletion. For example, the interaction term A*B will not be considered unless both A and B are already in the model. Likewise, the term A cannot be removed from a model that contains A*B.

Stop search when number of terms reaches

Once this number of terms has been entered into the model, the subset selection algorithm is terminated. Often you will have to run the procedure twice to find an appropriate value. You would set this value high for the first run and then reset it appropriately for the second run, depending upon the values of the log likelihood.

Note that the intercept is counted in this number.

Iterations Tab

Estimation Options

The following options are used during the likelihood maximization process.

Maximum Iterations

Specifies the maximum number of iterations allowed during the iteration procedure. If this number is reached, the procedure is terminated prematurely. Typically, the maximum likelihood procedure converges in 20 to 30 iterations, so a value of twenty here should be ample.

Convergence Zero

This option specifies the convergence target for the maximum likelihood estimation procedure. When all of the maximum likelihood equations are less than this amount, the algorithm is assumed to have converged. In theory, all of the equations should be zero. However, about the best that can be achieved is 1E-13, so you should set this value to a number a little larger than this such as the default of 1E-9.

The actual value can be found by looking at the Maximum Convergence value on the Run Summary report.

Reports Tab

The following options control which reports are displayed.

Alpha

Alpha Level

Alpha is the significance level used in the hypothesis tests. One minus alpha is the confidence level of the confidence intervals. A value of 0.05 is most commonly used. This corresponds to a chance of error of 1 in 20. You should not be afraid to use other values since 0.05 became popular in pre-computer days when it was the only value available.

Typical values range from 0.001 to 0.20.

Select Reports – Summaries

Run Summary ... Means

Each of these options specifies whether the indicated report is calculated and displayed.

Select Reports – Subset Selection

Subset Selection - Summary and Subset Selection - Detail

Indicate whether to display these subset selection reports.

Select Reports – Estimation

Regression Coefficients ... Rate Coefficients

Indicate whether to display these estimation reports.

Select Reports – Goodness-of-Fit

Lack-of-Fit Statistics ... Log-Likelihood and R-Squared

Indicate whether to display these model goodness-of-fit reports.

Select Reports – Row-by-Row Lists

Residuals ... Incidence

Indicate whether to display these list reports. Note that since these reports provide results for each row, they may be too long for normal use when requested on large databases.

Incidence Counts

Up to five incidence counts may be entered. The probabilities of these counts under the Poisson regression model will be displayed on the Incidence Report.

These values must be non-negative integers.

Exposure Value

Specify the exposure (time, space, distance, volume, etc.) value to be used as a multiplier on the Incidence Report. All items on that report are scaled to this amount. For example, if your data was scaled in terms of events per month but you want the Incidence report scaled to events per year, you would enter '12' here.

Report Options Tab

These options control format of the reports.

Variable Labels

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Stagger label and output if label length is ≥

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable's results printed on two lines.

Enter *100* when you want each variable's results printed on a single line.

Decimal Places

Precision

Specifies whether unformatted numbers (designated as decimal places = 'All') are displayed as single (7-digit) or double (13-digit) precision numbers in the output. All calculations are performed in double precision regardless of the Precision selected here.

Single

Unformatted numbers are displayed with 7-digits. This is the default setting. All reports have been formatted for single precision.

Double

Unformatted numbers are displayed with 13-digits. This option is most often used when the extremely accurate results are needed for further calculation. For example, double precision might be used when you are going to use the Multiple Regression model in a transformation.

Double Precision Format Misalignment

Double precision numbers require more space than is available in the output columns, causing column alignment problems. The double precision option is for those instances when accuracy is more important than format alignment.

Comments

1. This option does not affect formatted numbers such as probability levels.

2. This option only influences the format of the numbers as they presented in the output. All calculations are performed in double precision regardless of the Precision selected here.

Y ... Incidence Rate Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter **All** to display all digits available. The number of digits displayed by this option is controlled by whether the **Precision** option is *Single* or *Double*.

Plots Tab

These options control the attributes of the various plots.

Select Plots

Incidence (Y/T) vs X Plot ... Resid vs Row Plot

Indicate whether to display these plots. Click the plot format button to change the plot settings.

Edit During Run

This is the small check-box in the upper right-hand corner of the format button. If checked, the graphics format window for this plot will be displayed while the procedure is running so that you can format it with the actual data.

Plot Options

Residual Plotted

This option specifies which of the three types of residuals are shown on the residual plots.

Storage Tab

These options let you specify if, and where on the dataset, various statistics are stored.

Warning: Any data already in these columns are replaced by the new data. Be careful not to specify columns that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the dataset when the procedure is run.

- **Do not store data**
  No data are stored even if they are checked.

- **Store in empty columns only**
  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**
  Beginning at the *Store First Item In* column, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Item In

The first item is stored in this column. Each additional item that is checked is stored in the columns immediately to the right of this column.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these columns is automatically replaced, so be careful..

Data Storage Options – Select Items to Store

Expanded X Values ... Covariance Matrix

Indicated whether to store these row-by-row values, beginning at the column indicated by the *Store First Item In* option. Note that several of these values include a different value for each group and so they require several columns when they are stored.

Expanded X Values

This option refers to the experimental design matrix. They include all binary and interaction variables generated.

Example 1 – Negative Binomial Regression using a Dataset with Indicator Variables

This section presents several examples. In the first example, the data shown earlier in the Data Structure section and found in the Koch36 dataset will be analyzed. Koch et. al. (1986) presented this dataset. It contains the number of new melanoma cases in 1969-1971 among white males in two areas for various age groups. The size of the estimated population at risk is given in the variable Population.

This dataset is instructive because it shows how easily categorical variables are dealt with. In this example, two categorical variables, Area and AgeGroup, will be included in the regression model. The dataset can also be used to validate the program since the results are given in Koch (1986).

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Negative Binomial Regression window.

1 Open the Koch36 dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Koch36.NCSS**.
- Click **Open**.

2 Open the Negative Binomial Regression window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Negative Binomial Regression** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Negative Binomial Regression window, select the **Variables, Model tab**.
- Double-click in the **Dependent Y** box. This will bring up the variable selection window.
- Select **Melanoma** from the list of variables and click **Ok**. *Melanoma* will appear in the **Dependent Y** box.
- Double-click in the **Exposure T** variable box.
- Select **Population** from the list of variables and click **Ok**.
- Double-click in the **Categorical X's** box.
- Enter **Area(B;0) AgeGroup(B;<35)** in the **Categorical X's** box. The values in parentheses give the reference value for each variable.
- The rest of this panel can be left at the default values.

4 Specify the model.

- Set the **Terms** option to **1-Way**.
- Set the **Subset Selection** option to **None**.

5 Specify the reports.

- Select the **Reports tab**.
- Check all of the reports and plots. Normally, you would not want all of them, but we specify them now for documentation purposes.
- Set the **Incidence Counts** to **5 10 15 20 25**.
- Set the **Exposure Value** to **100000**.

6 Specify the decimals.

- Select the **Report Options tab**.
- Set the number of **decimal places for Probability** to **6**.

7 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary

Item	Value	Item	Value
Dependent Variable	Melanoma	Rows Used	12
Exposure Variable	Population	Sum of Frequencies	12
Frequency Variable	None	Iterations	21
Ind. Var's Available	2	Convergence Setting	1E-09
No. of X's in Model	6	Rel LogLike Change	0.00796603
LogLike: Max Possible	-54.0583	Subset Method	None
LogLike: Model	-54.2572	Alpha	0.27586

This report provides several details about the data and the MLE algorithm.

Dependent, Exposure, and Frequency Variables

These variables are listed to provide a record of the variables that were analyzed.

Ind. Var's Available

This is the number of independent variables that you have selected.

No. of X's in Model

This is the number of actual *X*-variables generated from the terms in the model that was used in the analysis.

LogLike: Max Possible

This is the maximum possible value of the log likelihood.

LogLike: Model

This is the value of the log likelihood that was achieved for this run.

Rows Used

This is the number of rows used by the estimation algorithm. Rows with missing values and filtered rows are not included. Always check this value to make sure that you are analyzing all of the data you intended to.

Sum of Frequencies

This is the number of observations used by the estimation algorithm. If you specified a Frequency Variable, this will be greater than the number of rows. If not, they will be equal.

Iterations

This is number of iterations used by the estimation algorithm.

Convergence Setting

When the relative change in the log-likelihood is less than this amount, the maximum likelihood algorithm stops. The algorithm also stops when the maximum number of iterations is reached.

Rel LogLike Change

This is the relative change of the log-likelihoods from the last two iterations.

Subset Method

This is the type of subset selection that was run.

Alpha

This line gives the estimated value of alpha, the dispersion parameter.

Model Summary

Model	Model DF	Error DF	Log Likelihood	Deviance	AIC(1)	Pseudo R^2
Intercept	1	11	-65.8419			
Model	7	5	-54.2572	0.3976	122.5143	0.9831
Maximum	12		-54.0583			

This report is analogous to the analysis of variance table. It summarizes the goodness of fit of the model.

Model

This is the term(s) that are reported about on this row of the report. Note that the model line includes the intercept.

Model DF

This is the number of variables in the model.

Error DF

This is the number of observations minus the number of variables.

Log Likelihood

This is the value of the log-likelihood function for the intercept only model, the chosen model, and the saturated model that fits the data perfectly. By comparing these values, you obtain an understanding of how well you model fits the data.

Deviance

The deviance is the generalization of the sum of squares in regular multiple regression. It measures the discrepancy between the fitted values and the data.

AIC(1)

This is Akaike's information criterion (AIC) as given earlier. It has been shown that using AIC to compare competing models with different numbers of parameters amounts to selecting the model with the minimum estimate of the mean squared error of prediction.

Pseudo R^2

This is the generalization of regular R^2 in multiple regression. Its formula is

$$R^2 = \frac{LL_{fit} - LL_0}{LL_{\max} - LL_0}$$

Means Report

Variable	Mean	Minimum	Maximum
Melanoma	68.667	27.000	104.000
Population	554422.917	34233.000	2880262.000

This report is analogous to the analysis of variance table. It summarizes the goodness of fit of the model.

Variable

The name of the variable.

Mean

The mean of the variable.

Minimum

The smallest value in the variable.

Maximum

The largest value in the variable.

Regression Coefficients Report

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Z Value H0: β=0	Prob Level	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
Alpha	0.27586	0.02968	9.30	0.000000	0.21770	0.33402
Intercept	-10.64692	0.41232	-25.82	0.000000	-11.45505	-9.83880
(Area=1)	0.81421	0.31189	2.61	0.009040	0.20291	1.42550
(AgeGroup="35-44")	1.79188	0.53895	3.32	0.000885	0.73556	2.84821
(AgeGroup="45-54")	1.89838	0.53853	3.53	0.000423	0.84288	2.95389
(AgeGroup="54-64")	2.22288	0.53863	4.13	0.000037	1.16719	3.27858
(AgeGroup="65-74")	2.37990	0.54175	4.39	0.000011	1.31809	3.44172
(AgeGroup=">74")	2.88030	0.54317	5.30	0.000000	1.81572	3.94489

Estimated Negative Binomial Regression Model

Melanoma =
Exp(-10.6469211619537 + 0.814208330468012*(Area=1) + 1.79188051909463*(AgeGroup="35-44") +
1.89838266017341*(AgeGroup="45-54") + 2.22288137703531*(AgeGroup="54-64") + 2.37990175349948*(AgeGroup="65-74") +
2.88030453282956*(AgeGroup=">74"))

Transformation Note:
Regular transformations must be less the 255 characters. If this expression is longer the 255 characters,
copy this expression and paste it into a text file, then use the transformation FILE(filename.txt)
to access the text file.

This report provides the estimated regression model and associated statistics. It provides the main results of the analysis.

Independent Variable

This item provides the name of the independent variable shown on this line of the report. The *Intercept* refers to the optional constant term. The *Alpha* value is the estimated value of the dispersion coefficient.

Note that whether a line is skipped after the name of the independent variable is displayed is controlled by the *Stagger label and output if label length is* ≥ option in the Format tab.

Regression Coefficient

These are the maximum-likelihood estimates of the regression coefficients, $b_1, b_2, \cdots, b_k$. Their direct interpretation is difficult because the formula for the predicted value involves the exponential function.

Standard Error

These are the asymptotic standard errors of the regression coefficients, the s_{b_i}. They are an estimate the precision of the regression coefficient. The standard errors are the square roots of the diagonal elements of this covariance matrix.

Wald's Chi² H0: β=0

This is the one degree of freedom chi-square statistic for testing the null hypothesis that $\beta_i = 0$ against the two-sided alternative that $\beta_i \neq 0$. The chi-square value is called a *Wald statistic*. This test has been found to follow the chi-square distribution only in large samples.

The test statistic is calculated using

$$\chi_1^2 = \left(\frac{b_i}{s'_{b_i}} \right)^2$$

P-Value

The probability of obtaining a chi-square value greater than the above. This is the significance level of the test. If this value is less than some predefined alpha level, say 0.05, the variable is said to be statistically significant.

Lower and Upper Confidence Limits

These provide a large-sample confidence interval for the values of the coefficients. The width of the confidence interval provides you with a sense of how precise the regression coefficients are. Also, if the confidence interval includes zero, the variable is not *statistically significant.* The formula for the calculation of the confidence interval is

$$b_i \pm z_{1-\alpha/2} s'_{b_i}$$

where $1-\alpha$ is the confidence coefficient of the confidence interval and *z* is the appropriate value from the standard normal distribution.

Rate Report

Independent Variable	Regression Coefficient b(i)	Rate Ratio Exp(b(i))	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
(Area=1)	0.81421	2.257	1.225	4.160
(AgeGroup="35-44")	1.79188	6.001	2.087	17.257
(AgeGroup="45-54")	1.89838	6.675	2.323	19.180
(AgeGroup="54-64")	2.22288	9.234	3.213	26.538
(AgeGroup="65-74")	2.37990	10.804	3.736	31.241
(AgeGroup=">74")	2.88030	17.820	6.145	51.671

This report provides the rate ratio for each independent variable.

Independent Variable

This item provides the name of the independent variable shown on this line of the report. The *Intercept* refers to the optional constant term.

Regression Coefficient

These are the maximum-likelihood estimates of the regression coefficients, $b_1, b_2, \cdots, b_k$. Their direct interpretation is difficult because the formula for the predicted value involves the exponential function.

Rate Ratio

These are the exponentiated values of the regression coefficients. The formula used to calculate these is

$$RR_i = e^{b_i}$$

The rate ratio is mainly useful for interpretation of the regression coefficients of indicator variables. In this case, they estimate the incidence of the response variable (melanoma in this example) in the given category relative to the category whose indicator variable was omitted (usually called the *control* group).

Lower and Upper Confidence Limits

These provide a large-sample confidence interval for the rate ratios. The formula for the calculation of the confidence interval is

$$\exp\left(b_i \pm z_{1-\alpha/2} s'_{b_i}\right)$$

where $1-\alpha$ is the confidence coefficient of the confidence interval and z is the appropriate value from the standard normal distribution.

Lack-of-Fit Statistics

Statistic	Value
Log Likelihood: Max Possible	-54.0583
Log Likelihood: Model	-54.2572
Log Likelihood: Intercept Only	-65.8419
Deviance	0.3976
AIC(1)	122.5143
AIC(n)	10.2095
BIC(R)	-12.0269
BIC(L)	125.9086
BIC(Q)	11.3131

This report provides several goodness-of-fit statistics that were described earlier in this chapter.

Analysis of Deviance

Model Term	DF	Deviance	Increase From Model Deviance (Chi^2)	P-Value
Intercept Only	1	23.5671		
Area	1	8.8615	8.46	0.003623
AgeGroup	5	21.1798	20.78	0.000891
(Full Model)	7	0.3976		

This report is the negative binomial regression analog of the analysis of variance table. It displays the results of a chi-square test of the significance of the individual terms in the regression.

This report is not produced during a subset selection run.

Note that this report requires that a separate regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Model Term

This is the model term that is being tested. The test is formed by comparing the deviance statistic when the term is removed with the deviance of the complete model. Thus, the deviance when the term is left out of the model is shown.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the *Stagger label and output* option in the Report Options tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the chi^2 test displayed on this line.

Deviance

The deviance is equal to minus two times the log likelihood achieved by the model being described on this line of the report. See the discussion given earlier in this chapter for a technical discussion of the deviance. A useful way to interpret the deviance is as the analog of the residual sum of squares in multiple regression. This value is used to create the difference in deviance that is used in the chi-square test.

Increase From Model Deviance (Chi^2)

This is the difference between the deviance for the model described on this line and the deviance of the complete model. This value follows the chi-square distribution in medium to large samples. This value can the thought of as the analog of the residual sum of squares in multiple regression. Thus, you can think of this value as the increase in the residual sum of squares that occurs when this term is removed from the model.

Another way to interpret this test is as a redundancy test because it tests whether this term is redundant after considering all of the other terms in the model.

P-Value

This is the significance level of the chi-square test. This is the probability that a chi^2 value with degrees of freedom DF is equal to this value or greater. If this value is less than 0.05 (or other appropriate value), the term is said to be statistically significant.

Log Likelihood & R^2 Report

Model Term	DF	Log Likelihood if this Term is Omitted	Total R^2 if this Term is Omitted	Increase in R^2 if this Term is Included
Intercept Only	1	-65.8419		
Area	1	-58.4891	0.6240	0.3591
AgeGroup	5	-64.6482	0.1013	0.8818
(Full Model)	7	-54.2572		0.9831
(Perfect Fit)	12	-54.0583		1.0000

This report provides the log likelihoods and R^2 values of various models. This report is not produced during a subset selection run.

Note that this report requires that a separate regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Model Term

This is the term being analyzed on this line. The “(Perfect Fit)” line gives the results for the saturated (complete) model.

DF

This is the degrees of freedom of the term displayed on this line.

Log Likelihood if this Term is Omitted

This is the log likelihood of the regression without the term listed.

Total R^2 if this Term is Omitted

This is the *pseudo-*R^2 of the model without the term listed at the beginning of the line.

Increase in R^2 if this Term is Included

This is amount that R^2 is increased when this term added to the regression model.

Residuals Report

Row	Melanoma (Y)	Predicted Value	Raw Residual	Pearson Residual	Anscombe Residual	Population (T)
1	61	68.4751	-7.4751	-0.2026	-0.9206	2880262
2	76	80.5370	-4.5370	-0.1049	-0.5104	564535
3	98	94.1023	3.8977	0.0774	0.3991	592983
4	104	98.9492	5.0508	0.0955	0.5035	450740
5	63	69.5826	-6.5826	-0.1756	-0.8021	270908
6	80	68.5668	11.4332	0.3094	1.3448	161850
7	64	57.6515	6.3485	0.2034	0.8214	1074246
8	75	70.9800	4.0200	0.1052	0.4727	220407
9	68	70.9725	-2.9725	-0.0778	-0.3554	198119
10	63	66.4460	-3.4460	-0.0962	-0.4265	134084
11	45	40.9972	4.0028	0.1782	0.6154	70708
12	27	32.7380	-5.7380	-0.3166	-1.0345	34233

This report provides the predicted values and various types of residuals. Large residuals indicate data points that were not fit well by the regression model.

Predicted Means Report

Row	Melanoma (Y)	Predicted Value	Standard Error	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit	Population (T)
1	61	68.4751	28.2334	13.1387	123.8116	2880262
2	76	80.5370	33.0823	15.6969	145.3772	564535
3	98	94.1023	38.5855	18.4761	169.7285	592983
4	104	98.9492	40.5706	19.4323	178.4661	450740
5	63	69.5826	28.7805	13.1739	125.9914	270908
6	80	68.5668	28.4483	12.8092	124.3244	161850
7	64	57.6515	23.8026	10.9993	104.3037	1074246
8	75	70.9800	29.1809	13.7866	128.1735	220407
9	68	70.9725	29.1520	13.8357	128.1094	198119
10	63	66.4460	27.3113	12.9169	119.9752	134084
11	45	40.9972	17.0387	7.6019	74.3924	70708
12	27	32.7380	13.6852	5.9155	59.5605	34233

This report provides the predicted values along with their standard errors and confidence limits.

If you want to generate predicted values and confidence limits for X values not on your database, you should add them to the bottom of the database, leaving Y blank (if you are using an exposure variable, set the value of T to a desired value). These rows will not be included in the estimation algorithm, but they will appear on this report with estimated Y's.

Incidence Report when Exposure = 100000

Row	Average Incidence Rate	Prob that Count is 5	Prob that Count is 10	Prob that Count is 15	Prob that Count is 20	Prob that Count is 25
1	2.3774	0.062192	0.002595	0.000064	0.000001	0.000000
2	14.2661	0.039247	0.054161	0.044201	0.028248	0.015661
3	15.8693	0.031887	0.048797	0.044163	0.031299	0.019242
4	21.9526	0.015521	0.030944	0.036485	0.033688	0.026983
5	25.6850	0.010511	0.023255	0.030426	0.031173	0.027707
6	42.3644	0.002635	0.007483	0.012567	0.016526	0.018852
7	5.3667	0.111669	0.036207	0.006943	0.001042	0.000136
8	32.2041	0.005762	0.014470	0.021490	0.024994	0.025217
9	35.8232	0.004279	0.011313	0.017688	0.021656	0.023001
10	49.5555	0.001649	0.004958	0.008820	0.012285	0.014844
11	57.9809	0.001017	0.003214	0.006009	0.008796	0.011171
12	95.6329	0.000203	0.000721	0.001516	0.002495	0.003562

This report gives the predicted incidence rate and Poisson probabilities for various counts.

Row

The row number of the item. If you have excluded some rows by using a filter or if some of the rows had missing values, the row number identifies the original row on the database.

Average Incidence Rate

This is the predicted incidence rate calculated using the formula

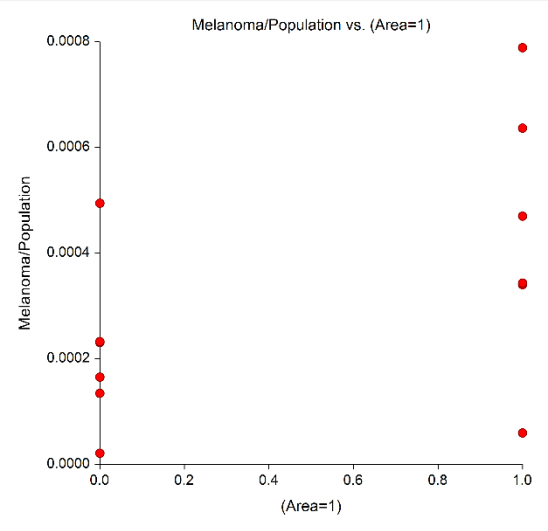
$$\hat{\mu}_i = T\hat{\mu}(\mathbf{x}_i'\mathbf{b})$$

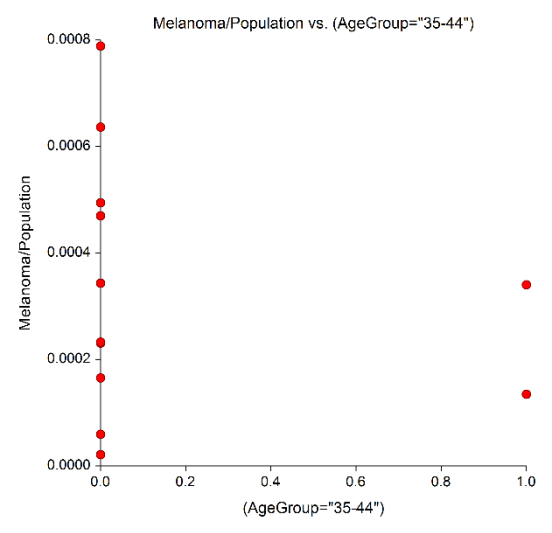
Note that the calculation is made for a specific exposure value, not the value of T on the database. This allows you to make valid comparisons of the incidence rates.

Prob that Count is Y

Using the negative binomial probability distribution, the probability of obtaining exactly Y events during the exposure amount given in the Exposure Value box is calculated for the values of Y specified in the Incidence Counts box.

Plots of Y/T (Incidence) vs X

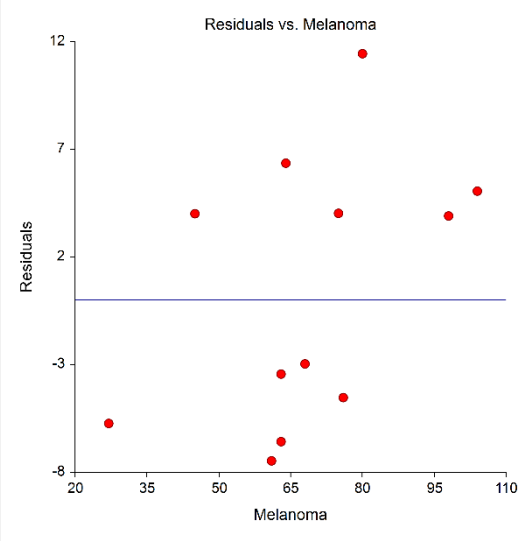


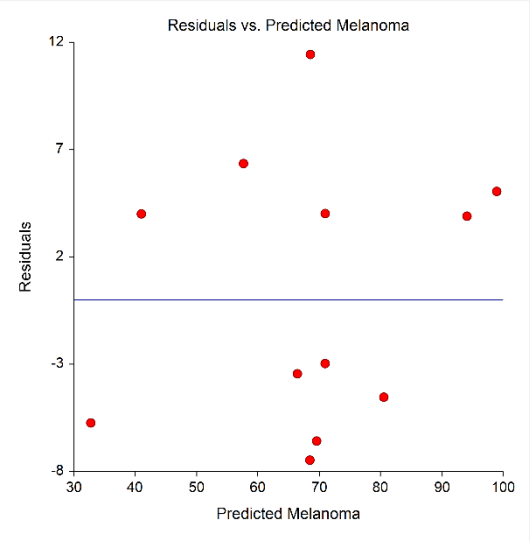


(four more similar plots are shown here)

These plots show each of the independent variables plotted against the incidence as measured by Y/T. They should be scanned for outliers and curvilinear patterns.

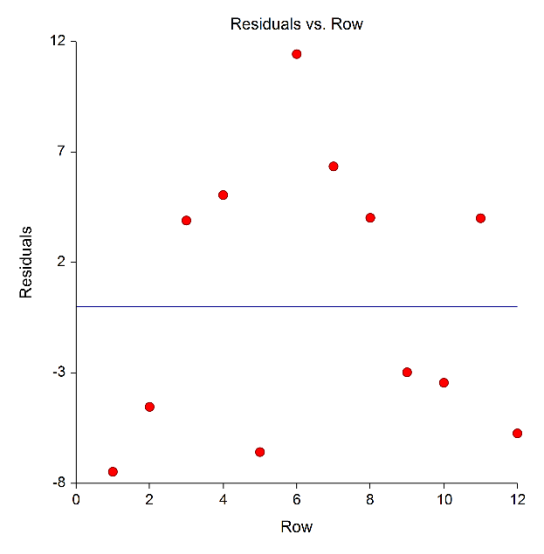
Plots of Residuals vs. Y and Predicted Y





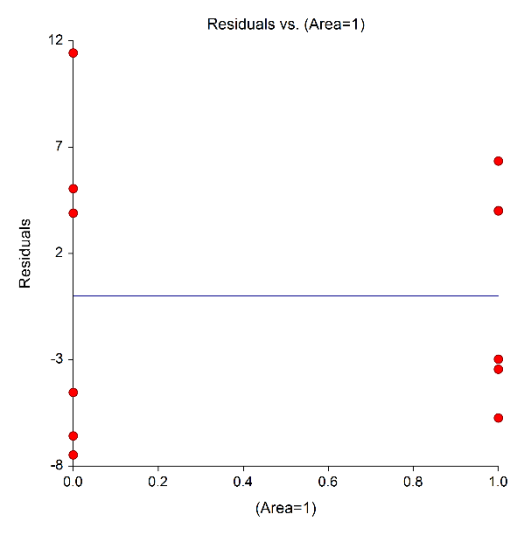
These plots show the residuals versus the dependent variable and the predicted value of the dependent variable. They are used to spot outliers.

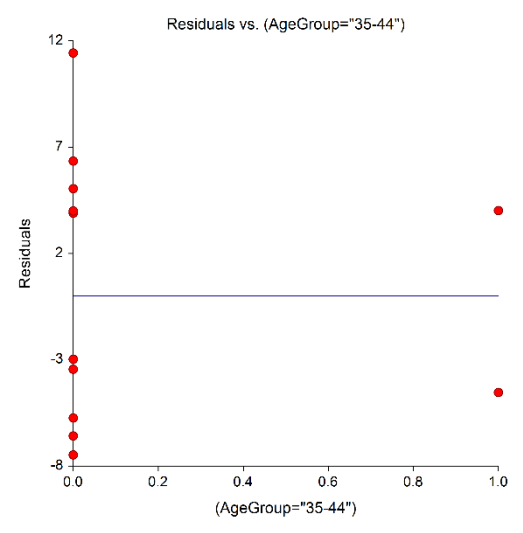
Plots of Residuals and Row



This plot shows the residuals versus the row numbers. It is used to quickly spot rows that have large residuals.

Plots of Residuals and X's





(four more similar plots are shown here)

These plots show the residuals plotted against the independent variables. They are used to spot outliers. They are also used to find curvilinear patterns that are not represented in the regression model.

Example 2a – Subset Selection

This example will demonstrate how to select an appropriate subset of the independent variables that are available. The dataset to be analyzed consists of ten independent variables, a dependent variable, a frequency variable, and an exposure variable. The dependent variable was generated using independent variables X1, X2, and X3 using the formula

$$Count = \text{Int}\{Time\ \text{Exp}(0.6 + 0.1X_1 + 0.2X_2 + 0.3X_3)\}$$

Variables X4, X5, and X6 were copies of X1 plus a small random component. Similarly, X7 and X8 were near copies of X2 and X9 and X10 were near copies of X3. These near copies of the original variables were added to cause confusion to the selection algorithm. The forty rows of data are stored in the PoisReg dataset.

To test the variable search, we assume that we do not know how the data were generated. Our task is to find a subset of the ten independent variables that does a good job of fitting the data. We plan to make two runs. The goal of the first run will be to find an appropriate subset size. Then, in the second run, we will identify the variables in this subset and estimate the various regression statistics.

You may follow along here by making the appropriate entries or load the completed template **Example 2a** by clicking on Open Example Template from the File menu of the Negative Binomial Regression window.

1 Open the PoisReg dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **PoisReg.NCSS**.
- Click **Open**.

2 Open the Negative Binomial Regression window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Negative Binomial Regression** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Negative Binomial Regression window, select the **Variables, Model tab**.
- Set the **Dependent Y** variable to **Count**.
- Set the **Exposure T** variable to **Time**.
- Set the **Numeric X's** variables to **X1-X10**.
- Set the **Frequencies** variable to **Cases.**

4 Specify the model.

- Set the **Terms** to **1-Way**.
- Set the **Search Method** to **Hierarchical Forward with Switching**.
- Set **Stop search when number of terms reaches** to **6**.
- The rest of this panel can be left at the default values.

5 Specify the reports.

- Select the **Reports tab**.
- Uncheck all of the reports and plots except **Run Summary**, **Subset Selection - Summary**, and **Subset Selection - Detail** (these should be checked).

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary

Item	Value	Item	Value
Dependent Variable	Count	Rows Used	40
Exposure Variable	Time	Sum of Frequencies	130
Frequency Variable	Cases	Iterations	21
Ind. Var's Available	10	Convergence Setting	1E-09
No. of X's in Model	5	Rel LogLike Change	0.005853259
LogLike: Max Possible	-373.1941	Subset Method	Hierarchical Forward/Switching
LogLike: Model	-373.5717	Alpha	0.17733

This report provides several details about the data and the MLE algorithm as it fit the best model found during the search. We note that, as expected, there were 40 rows used.

Subset Selection Summary

Number of Terms	Log Likelihood	Pseudo-R^2	Deviance	AIC(1)
1	-477.7489	0.0000		
2	-420.7844	0.5448	95.1806	845.5688
3	-393.5371	0.8054	40.6860	793.0741
4	-373.6209	0.9959	0.8535	755.2417
5	-373.5813	0.9963	0.7745	757.1626
6	-373.5717	0.9964	0.7553	759.1435

This report will help us determine an appropriate subset size. By scanning the *pseudo-R^2* column, we conclude that three variables are needed since this gets us up to 0.9959.

In this example, the four measures unanimously point to three as the appropriate subset size.

Number of Variables

This is the number of terms in the model including the intercept. Each line presents the results for the best model found for that subset size. The first line presents the results for the intercept-only model.

Log Likelihood

This is the value of the log likelihood function. Since the goal of maximum likelihood is to maximize this value, we want to select a subset size after which the log likelihood is not increased significantly.

In this example, after three terms are added (in addition to the intercept) the log likelihood does not change a great deal. The log likelihood points to a subset size of three terms plus the intercept for a total of four.

Pseudo-R^2

This is the value of pseudo R^2—a measure of the adequacy of the model. Since our goal is to maximize this value, we want to select a subset size after which the value is not increased significantly.

In this example, after four terms are included, the R^2 is 0.9959 and then it does not increase a great deal.

Deviance

Deviance is a measure of the lack of fit. Hence, we want to select a subset size after which the deviance is not significantly decreased.

In this example, after four terms are included, the Deviance is 0.7745 and it does not change a great deal. The Deviance values point to a subset size of four.

AIC(1)

These are the Akaike information criterion values for each subset size. This criterion measures both the lack of fit and the size of the regression model. Our goal is to minimize this value.

In this example, the subset size of four gives the lowest value AIC and is thus the subset size implied by this statistic.

Subset Selection Detail

Step	Action	No. of Terms	No. of X's	Log Likelihood	Pseudo-R^2	Term Entered	Term Removed
1	Add	1	1	-477.7489	0.0000	Intercept	
2	Add	2	2	-420.7844	0.5448	X3	
3	Add	3	3	-393.5371	0.8054	X2	
4	Add	4	4	-373.6528	0.9956	X6	
5	Switch	4	4	-373.6209	0.9959	X8	X2
6	Add	5	5	-373.5813	0.9963	X1	
7	Add	6	6	-373.5717	0.9964	X7	

This report shows the progress of the subset selection algorithm through its various steps. It shows the original term added at each step and any switching that was done.

Step

This is the number of the step in the subset selection process.

Action

Two actions are possible at each step: Add or Switch. *Add* means that the subset size was increased and the term entered as added to the set of active regressor variables. *Switch* means that the subset size remained the same while one active regressor was removed and another was activated.

No. of Terms

This is the number of active terms (including the intercept) at the end of this step.

No. of X's

This is the number of active variables (excluding the intercept) at the end of this step. This reminds you of how many *X* variables were generated for each term involving a categorical variable.

Log Likelihood

This is the value of the log likelihood after this step was completed.

Pseudo-R^2

This is the pseudo R^2 value after this step was completed.

Variable Entered

This is the name of the regressor that was added to the list of active regressor variables.

Variable Removed

In switching steps, this is the name of the variable that was removed from the list of active regressor variables.

Example 2b – Subset Selection Continued

Example 2a completed the first step in the subset selection process by indicating that a subset of four terms is appropriate. Now, a second run must be made to find those terms.

The instructions provide here assume that you have just completed Example 2a. If you have not, you must complete it first since we will only tell you what needs to be changed.

You may follow along here by making the appropriate entries or load the completed template **Example 2b** by clicking on Open Example Template from the File menu of the Negative Binomial Regression window.

1 Specify the model.

- On the Negative Binomial Regression window, select the **Variables, Model tab**.
- Set the **Stop search when number of terms reaches** to **4**.
- The rest of this panel can be left at the default values.

2 Specify the reports.

- Select the **Reports tab**.
- Uncheck all of the reports and plots except **Run Summary**, **Subset Summary**, **Subset Detail**, and **Regression Coefficients** (these should be checked).

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top) or press the F9 function key.

Run Summary Report

Parameter	Value	Parameter	Value
Dependent Variable	Count	Rows Used	40
Exposure Variable	Time	Sum of Frequencies	130
Frequency Variable	Cases	Iterations	21
Ind. Var's Available	10	Convergence Setting	1E-09
No. of X's in Model	3	Rel LogLike Change	0.005854715
LogLike: Max Possible	-373.2209	Subset Method	Hierarchical Forward/Switching
LogLike: Model	-373.6209	Alpha	0.17744

We note that the final model converged in 21 iterations and the relative log-likelihood change is 0.00585. This means that the algorithm terminated normally.

Subset Selection Summary

Number of Terms	Log Likelihood	Pseudo-R^2	Deviance	AIC(1)
1	-477.7489	0.0000		
2	-420.7844	0.5450	95.1270	845.5688
3	-393.5371	0.8056	40.6324	793.0741
4	-373.6209	0.9962	0.7999	755.2417

This report again shows us that a subset size of four is a reasonable choice.

Subset Selection Detail

Step	Action	No. of Terms	No. of X's	Log Likelihood	R^2	Term Entered	Term Removed
1	Add	1	1	-477.7489	0.0000	Intercept	
2	Add	2	2	-420.7844	0.5450	X3	
3	Add	3	3	-393.5371	0.8056	X2	
4	Add	4	4	-373.6528	0.9959	X6	
5	Switch	4	4	-373.6209	0.9962	X8	X2

This report shows the algorithm's journey through the maze of possible models. During the process, three variables were switched in order to achieve a better model.

Regression Coefficients Report

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Z Value H0: β=0	Two-Sided P-Value	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
Alpha	0.17744	0.00559	31.75	0.0000	0.16648	0.18839
Intercept	-0.15892	0.19009	-0.84	0.4032	-0.53150	0.21366
X3	0.01064	0.00078	13.63	0.0000	0.00911	0.01217
X6	0.00352	0.00067	5.27	0.0000	0.00221	0.00483
X8	0.00679	0.00092	7.41	0.0000	0.00499	0.00859

This report provides the details of the model that was selected. We note the X3, X6, and X8 were included in the model. We assume that X8 is taking the place of X2 and X6 is taking the place of X1. In fact, we ran a Poisson regression with X1, X2, and X3 in the model. The log likelihood for this model was -373.6839, which is slightly less than the -373.6209 achieved by our best model. This concludes our discussion of this example. Usually, we would go on to study the residual plots and complete the analysis by making a third run with only the variables X3, X6, and X8 specified.