Chapter 327

Geometric Regression

Introduction

Geometric regression is a special case of negative binomial regression in which the dispersion parameter is set to one. It is similar to regular multiple regression except that the dependent (*Y*) variable is an observed count that follows the geometric distribution. Thus, the possible values of *Y* are the nonnegative integers: 0, 1, 2, 3, and so on.

Geometric 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.

Few books on regression analysis discuss geometric regression. We are aware of only one book that discusses it: Hilbe (2014). Most of the results presented here are obtained from that book.

This program computes geometric 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 Geometric 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*.

When the dispersion parameter α is set to one, the result is called the geometric distribution

The Geometric Regression Model

In geometric 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 geometric regression model for an observation *i* is written as

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

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

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

Rearranging gives

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

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

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

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

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

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

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

Distribution of the MLE's

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

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}} \\ \hat{\alpha}\end{bmatrix} \sim \mathrm{N}\left[\boldsymbol{\beta}, \mathrm{V}(\widehat{\boldsymbol{\beta}})\right]$$

where

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

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

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

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. Also, 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

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

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

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

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.

Example 1 – Geometric 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).

Setup

To run this example, complete the following steps:

1 Open the Koch36 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Koch36** and click **OK**.

2 Specify the Geometric Regression procedure options

- Find and open the **Geometric Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab	
Dependent Y	**Melanoma**
Exposure T	**Population**
Categorical X's	**Area(B;0) AgeGroup(B;<35)**
Terms	**1-Way**
Search Method	**None - No Search is Conducted**

Reports Tab	
Subset Selection Summary	**Unchecked**
Subset Selection Detail	**Unchecked**
All Other Reports	**Checked**
Incidence	**Checked**
Incidence Counts	**5 10 15 20 25**
Exposure Value	**100000**

Plots Tab	
All Available Plots	**Checked** (click the *Check All* button)

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Dependent Variable	Melanoma	Rows Processed	12
Exposure Variable	Population	Rows Used in Estimation	12
Frequency Variable	None		
Independent Variables Available	2		
Number of X's in the Model	6		
Log-Likelihood: Maximum Possible	-62.2353		
Log-Likelihood: Model	-62.2930		
Number of Likelihood Iterations	16 of 20		
Convergence Setting	1E-09		
Relative Log-Likelihood Change	2.583665E-10		
Subset Selection Method	None		

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.

Independent Variables Available

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

Number of X's in the Model

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

Log-Likelihood: Maximum Possible

This is the maximum possible value of the log-likelihood.

Log-Likelihood: Model

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

Number of Likelihood 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.

Relative Log-Likelihood Change

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

Subset Selection Method

This is the type of subset selection that was used in the analysis.

Rows Processed

This is the number of rows read from the database. Rows with missing values and filtered rows are not included in the analysis.

Rows Used in Estimation

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.

Model Summary

Model Summary

Model	DF Model	DF Error	Log-Likelihood	Deviance	AIC(1)	Pseudo-R^2
Intercept	1	11	-66.4364			
Model	7	5	-62.2930	0.1154	138.5859	0.9863
Maximum	12		-62.2353			

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 your 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

Means

Variable	Mean	Minimum	Maximum
Melanoma	68.66666	27	104
Population	554422.9	34233	2880262

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

Regression Coefficients

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Z-Test of H0: β(i) = 0 Z-Statistic	Z-Test of H0: β(i) = 0 Two-Sided P-Value	95% Confidence Interval Limits for β(i) Lower	95% Confidence Interval Limits for β(i) Upper
Alpha	1.00000					
Intercept	-10.64623	0.76969	-13.83	0.0000	-12.15479	-9.13767
(Area=1)	0.81356	0.58195	1.40	0.1621	-0.32705	1.95417
(AgeGroup="35-44")	1.79164	1.00728	1.78	0.0753	-0.18259	3.76587
(AgeGroup="45-54")	1.89784	1.00706	1.88	0.0595	-0.07595	3.87164
(AgeGroup="54-64")	2.22221	1.00711	2.21	0.0273	0.24831	4.19611
(AgeGroup="65-74")	2.38061	1.00879	2.36	0.0183	0.40342	4.35780
(AgeGroup=">74")	2.87695	1.00959	2.85	0.0044	0.89819	4.85571

Estimated Equation

Melanoma =
Exp(-10.6462305085772 + 0.813559984153889*(Area=1) + 1.79163734884751*(AgeGroup="35-44") +
1.89784065910583*(AgeGroup="45-54") + 2.22221282025253*(AgeGroup="54-64") +
2.3806125356114*(AgeGroup="65-74") + 2.87694709506365*(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.

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 b(i)

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 Sb(i)

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

Z-Statistic

This is the *z*-statistic for testing the null hypothesis that $\beta_i = 0$ against the two-sided alternative that $\beta_i \neq 0$.

The test statistic is calculated using

$$Z = \frac{b_i}{s'_{b_i}}$$

P-Value

The probability of obtaining a *z*-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.

95% Confidence Interval Limits for β(i) (Lower and Upper)

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 Ratios

Rate Ratios

Independent Variable	Regression Coefficient b(i)	Rate Ratio Exp(b(i))	95% Confidence Interval Limits for the Rate Ratio	
			Lower	Upper
(Area=1)	0.81356	2.256	0.721	7.058
(AgeGroup="35-44")	1.79164	5.999	0.833	43.201
(AgeGroup="45-54")	1.89784	6.671	0.927	48.021
(AgeGroup="54-64")	2.22221	9.228	1.282	66.428
(AgeGroup="65-74")	2.38061	10.812	1.497	78.085
(AgeGroup=">74")	2.87695	17.760	2.455	128.472

This report is mainly for binary (0-1) variables.

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 b(i)

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 Exp(b(i))

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).

95% Confidence Interval Limits for the Rate Ratio (Lower and Upper)

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

Lack-of-Fit Statistics

Statistic	Value
Log-Likelihood: Max Possible	-62.2353
Log-Likelihood: Model	-62.2930
Log-Likelihood: Intercept Only	-66.4364
Deviance	0.1154
AIC(1)	138.5859
AIC(n)	11.5488
BIC(R)	-12.3092
BIC(L)	141.9803
BIC(Q)	12.6524

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

Analysis of Deviance

Analysis of Deviance

Model Term	DF	Deviance	Increase From Model Deviance (Chi²)	P-Value
Intercept Only	1	8.4022		
Area	1	2.0032	1.89	0.1694
AgeGroup	5	6.9156	6.80	0.2359
(Full Model)	7	0.1154		

The p-value is for testing the significance of each term after considering all other terms.

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 and R^2 Report

Log-Likelihood and R²

Model Term	DF	If Term Omitted		If Term Included
		Log-Likelihood	Total R²	Increase in R²
Intercept Only	1	-66.4364		
Area	1	-63.2369	0.7616	0.2247
AgeGroup	5	-65.6931	0.1769	0.8093
(Full Model)	7	-62.2930		0.9863
(Perfect Fit)	12	-62.2353		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 Term is Omitted

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

Total R^2 if 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 Term is Included

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

Residuals

Residuals

Row	Melanoma (Y)	Predicted Value	Residual			Population (T)
			Raw	Pearson	Anscombe	
1	61	68.5224	-7.5224	-0.1090	-0.9262	2880262
2	76	80.5731	-4.5731	-0.0564	-0.5144	564535
3	98	94.1163	3.8837	0.0410	0.3976	592983
4	104	98.9513	5.0487	0.0508	0.5033	450740
5	63	69.6802	-6.6802	-0.0952	-0.8136	270908
6	80	68.3842	11.6158	0.1686	1.3675	161850
7	64	57.6540	6.3460	0.1091	0.8211	1074246
8	75	70.9658	4.0342	0.0565	0.4745	220407
9	68	70.9371	-2.9371	-0.0411	-0.3512	198119
10	63	66.4044	-3.4044	-0.0509	-0.4214	134084
11	45	41.0281	3.9719	0.0957	0.6105	70708
12	27	32.6297	-5.6297	-0.1699	-1.0161	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

Predicted Means

Row	Melanoma (Y)	Predicted Mean	Standard Error	95% Confidence Interval Limits for the Predicted Mean		Population (T)
				Lower	Upper	
1	61	68.5224	52.7410	-34.8481	171.8929	2880262
2	76	80.5731	61.9498	-40.8463	201.9924	564535
3	98	94.1163	72.3262	-47.6404	235.8730	592983
4	104	98.9513	76.0409	-50.0861	247.9888	450740
5	63	69.6802	53.6812	-35.5330	174.8934	270908
6	80	68.3842	52.7327	-34.9700	171.7384	161850
7	64	57.6540	44.3928	-29.3544	144.6623	1074246
8	75	70.9658	54.5760	-36.0013	177.9329	220407
9	68	70.9371	54.5403	-35.9599	177.8341	198119
10	63	66.4044	51.0654	-33.6820	166.4909	134084
11	45	41.0281	31.6520	-21.0088	103.0649	70708
12	27	32.6297	25.2176	-16.7959	82.0553	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 when Exposure = 100000

Incidence when Exposure = 100000

Row	Average Incidence Rate	Probability that Count is 5	10	15	20	25
1	2.3790	0.0512	0.0089	0.0015	0.0003	0.0000
2	14.2725	0.0467	0.0333	0.0237	0.0169	0.0120
3	15.8717	0.0437	0.0322	0.0237	0.0175	0.0129
4	21.9531	0.0349	0.0279	0.0223	0.0179	0.0143
5	25.7210	0.0309	0.0256	0.0211	0.0175	0.0144
6	42.2516	0.0206	0.0183	0.0163	0.0145	0.0129
7	5.3669	0.0668	0.0284	0.0121	0.0052	0.0022
8	32.1976	0.0259	0.0222	0.0190	0.0163	0.0140
9	35.8053	0.0237	0.0206	0.0180	0.0157	0.0136
10	49.5245	0.0179	0.0162	0.0147	0.0133	0.0120
11	58.0246	0.0156	0.0143	0.0131	0.0120	0.0111
12	95.3164	0.0099	0.0094	0.0089	0.0084	0.0080

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

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

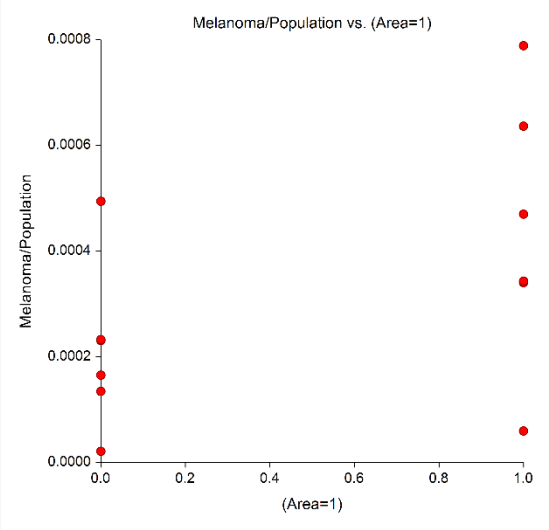
Note that the calculation is made for a specific exposure value, not the value of *T* on the database. This allows you to make valid comparisons of the incidence rates.

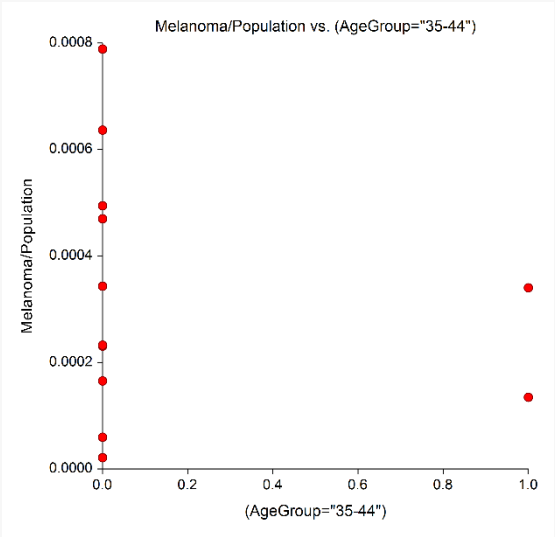
Probability that Count is Y

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.

Incidence (Y/T) vs X Plot(s)

Incidence (Y/T) vs X Plot(s)



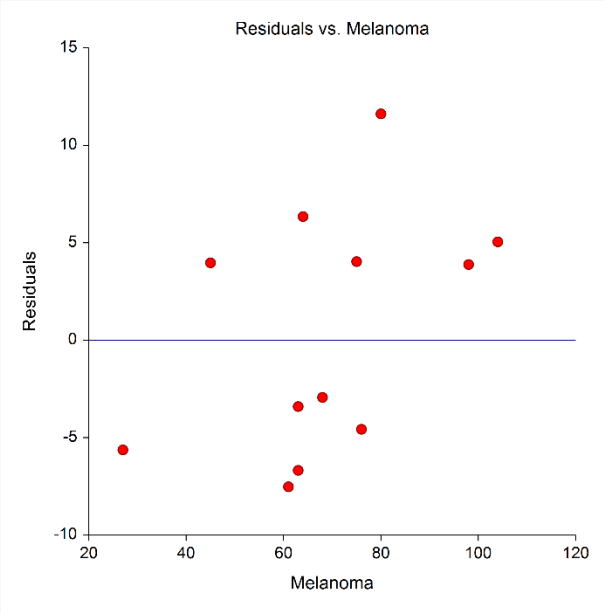


(four more similar plots are shown here)

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.

Residuals vs Y Plot

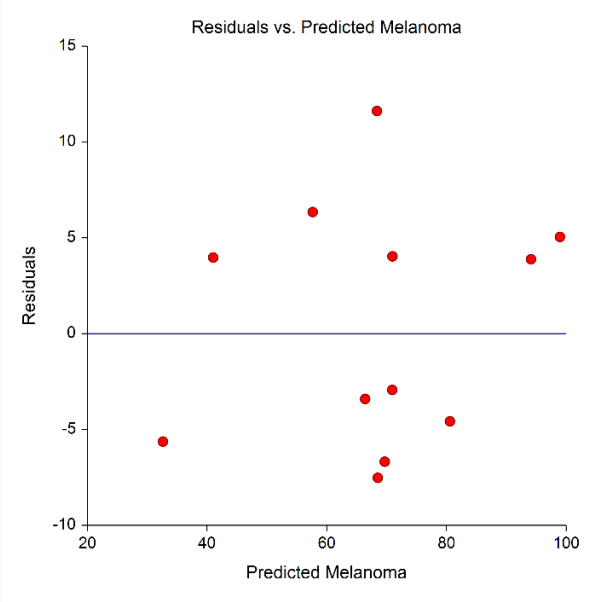
Residuals vs Y Plot



This plot shows the residuals versus the dependent variable. It can be used to spot outliers.

Residuals vs Yhat (Predicted Y) Plot

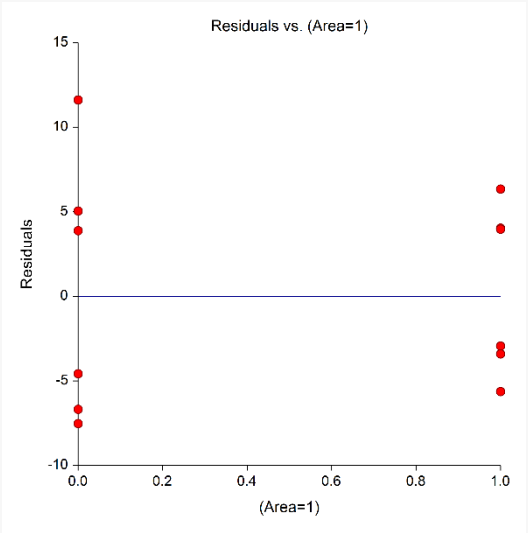
Residuals vs Yhat (Predicted Y) Plot

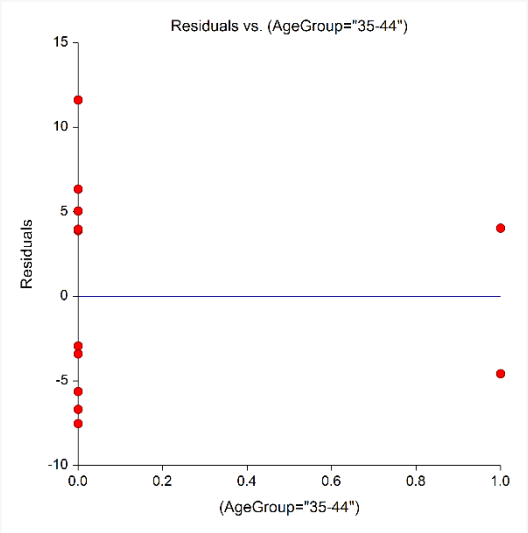


This plot shows the residuals versus the predicted value (Yhat) of the dependent variable. It can show outliers.

Residuals vs X Plots

Residuals vs Row Plot



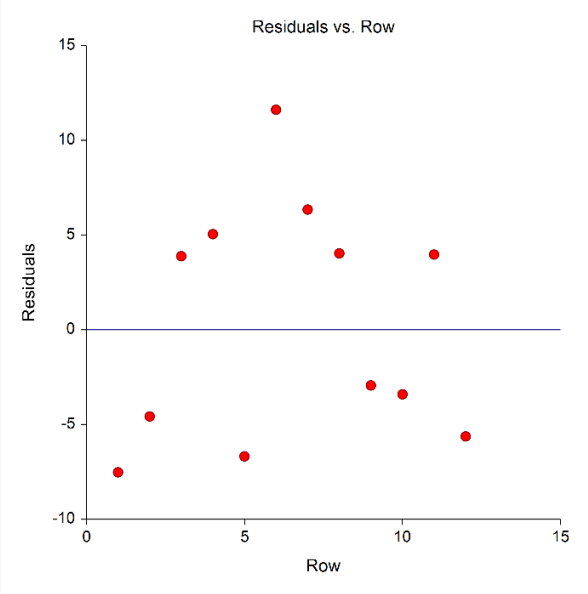


(four more similar plots are shown here)

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.

Residuals vs Row Plot

Residuals vs Row Plot



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

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

$$Count = \text{Int}[Time \times \text{Exp}(0.6 + 0.1X1 + 0.2X2 + 0.3X3)]$$

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.

Setup

To run this example, complete the following steps:

1 Open the PoisReg example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **PoisReg** and click **OK**.

2 Specify the Geometric Regression procedure options

- Find and open the **Geometric Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2a** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab	
Dependent Y	**Count**
Exposure T	**Time**
Numeric X's	**X1-X10**
Frequencies	**Cases**
Terms	**1-Way**
Search Method	**Hierarchical Forward with Switching**
Stop search when number of terms Reaches	**6**

Reports Tab	
Run Summary	**Checked**
Subset Selection Summary	**Checked**
Subset Selection Detail	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Dependent Variable	Count	Rows Processed	40
Exposure Variable	Time	Rows Used in Estimation	40
Frequency Variable	Cases	Sum of Frequencies	130
Independent Variables Available	10		
Number of X's in the Model	5		
Log-Likelihood: Maximum Possible	-471.6051		
Log-Likelihood: Model	-471.7136		
Number of Likelihood Iterations	8 of 20		
Convergence Setting	1E-09		
Relative Log-Likelihood Change	1.177855E-11		
Subset Selection Method	Hierarchical Forward with Switching		

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

Subset Selection Summary

Subset Selection Method = Hierarchical Forward with Switching

Number of Terms	Log-Likelihood	Pseudo-R^2	Deviance	AIC(1)
1	-499.7159	0.0000		
2	-481.0421	0.6643	18.8740	966.0842
3	-475.1461	0.8740	7.0820	956.2922
4	-471.7258	0.9957	0.2414	951.4516
5	-471.7150	0.9961	0.2198	953.4300
6	-471.7136	0.9961	0.2170	955.4272

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 0.9957.

In this example, the four measures unanimously point to three (not including the intercept) 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

Subset Selection Detail

Subset Selection Method = Hierarchical Forward with Switching

		Number of				Term	
Step	Action	Terms	X's	Log-Likelihood	Pseudo-R²	Entered	Removed
1	Add	1	1	-499.7159	0.0000	Intercept	
2	Add	2	2	-481.0421	0.6643	X3	
3	Add	3	3	-475.1461	0.8740	X2	
4	Add	4	4	-471.7296	0.9956	X6	
5	Switch	4	4	-471.7258	0.9957	X8	X2
6	Add	5	5	-471.7150	0.9961	X5	
7	Add	6	6	-471.7136	0.9961	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.

Number of Terms

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

Number 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.

Term Entered

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

Term 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 provided 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.

Setup

To run this example, complete the following steps:

1 Open the PoisReg example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **PoisReg** and click **OK**.

2 Specify the Geometric Regression procedure options

- Find and open the **Geometric Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2b** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab	
Dependent Y	**Count**
Exposure T	**Time**
Numeric X's	**X1-X10**
Frequencies	**Cases**
Terms	**1-Way**
Search Method	**Hierarchical Forward with Switching**
Stop search when number of terms Reaches	**4**

Reports Tab	
Run Summary	**Checked**
Subset Selection Summary	**Checked**
Subset Selection Detail	**Checked**
Regression Coefficients	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Dependent Variable	Count	Rows Processed	40
Exposure Variable	Time	Rows Used in Estimation	40
Frequency Variable	Cases	Sum of Frequencies	130
Independent Variables Available	10		
Number of X's in the Model	3		
Log-Likelihood: Maximum Possible	-471.6051		
Log-Likelihood: Model	-471.7258		
Number of Likelihood Iterations	8 of 20		
Convergence Setting	1E-09		
Relative Log-Likelihood Change	8.228771E-12		
Subset Selection Method	Hierarchical Forward with Switching		

We note that the final model converged in 8 iterations and the relative log-likelihood change is 0.00585. This means that the algorithm terminated normally.

Subset Selection Summary

Subset Selection Summary

Subset Selection Method = Hierarchical Forward with Switching

Number of Terms	Log-Likelihood	Pseudo-R^2	Deviance	AIC(1)
1	-499.7159	0.0000		
2	-481.0421	0.6643	18.8740	966.0842
3	-475.1461	0.8740	7.0820	956.2922
4	-471.7258	0.9957	0.2414	951.4516

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

Subset Selection Detail

Subset Selection Detail

Subset Selection Method = Hierarchical Forward with Switching

Step	Action	Number of Terms	Number of X's	Log-Likelihood	Pseudo-R^2	Term Entered	Term Removed
1	Add	1	1	-499.7159	0.0000	Intercept	
2	Add	2	2	-481.0421	0.6643	X3	
3	Add	3	3	-475.1461	0.8740	X2	
4	Add	4	4	-471.7296	0.9956	X6	
5	Switch	4	4	-471.7258	0.9957	X8	X2

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

Regression Coefficients

Regression Coefficients

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Z-Test of H0: β(i) = 0 Z-Statistic	Z-Test of H0: β(i) = 0 Two-Sided P-Value	95% Confidence Interval Limits for β(i) Lower	95% Confidence Interval Limits for β(i) Upper
Alpha	1.00000					
Intercept	-0.17780	0.35284	-0.50	0.6143	-0.86936	0.51376
X3	0.01073	0.00151	7.12	0.0000	0.00778	0.01368
X6	0.00356	0.00135	2.64	0.0084	0.00091	0.00620
X8	0.00680	0.00187	3.64	0.0003	0.00314	0.01047

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.