Chapter 328

Zero-Inflated Negative Binomial Regression

Introduction

The zero-inflated negative binomial (ZINB) regression is used for count data that exhibit overdispersion and excess zeros. The data distribution combines the negative binomial distribution and the logit distribution. The possible values of *Y* are the nonnegative integers: 0, 1, 2, 3, and so on.

The results presented here are documented in the books by Cameron and Trivedi (2013) and Hilbe (2014) and in Garay, Hashimoto, Ortega, and Lachos (2011).

This program computes ZINB regression on both numeric and categorical variables. It reports on the regression equation as well as the confidence limits and likelihood. It performs a comprehensive residual analysis including diagnostic residual reports and plots.

The Zero-Inflated Negative Binomial Regression Model

Suppose that for each observation, there are two possible cases. Suppose that if case 1 occurs, the count is zero. However, if case 2 occurs, counts (including zeros) are generated according to the negative binomial model. Suppose that case 1 occurs with probability π and case 2 occurs with probability 1 - π. Therefore, the probability distribution of the ZINB random variable y_i can be written

$$\Pr(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)g(y_i = 0) & \text{if } j = 0 \\ (1-\pi_i)g(y_i) & \text{if } j > 0 \end{cases}$$

where π_i is the logistic link function defined below and $g(y_i)$ is the negative binomial distribution given by

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

The negative binomial component can include an 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 $b_1, b_2, \ldots, b_k$.

This logistic link function π_i is given by

$$\pi_i = \frac{\lambda_i}{1+\lambda_i}$$

where

$$\lambda_i = \exp(\ln(t_i) + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \cdots + \gamma_m z_{mi})$$

The logistic component includes an exposure time *t* and a set of *m* regressor variables (the *z's*). Note that the *z's* and the *x's* may or may not include terms in common.

Solution by Maximum Likelihood Estimation

The regression coefficients are estimated using the method of maximum likelihood. The logarithm of the likelihood function is

$$\mathcal{L} = L1 + L2 + L3 - L4$$

where

$$L1 = \sum_{\{i:y_i=0\}} \ln\left[\lambda_i + (1+\alpha\mu_i)^{-\alpha^{-1}}\right]$$

$$L2 = \sum_{\{i:y_i>0\}} \sum_{j=0}^{y_i-1} \ln(j + \alpha^{-1})$$

$$L3 = \sum_{\{i:y_i>0\}} \{-\ln(y_i!) - (y_i + \alpha^{-1})\ln(1+\alpha\mu_i) + y_i \ln(\alpha) + y_i \ln(\mu_i)\}$$

$$L4 = \sum_{i=1}^{n} \ln(1+\lambda_i)$$

The gradient of $\mathcal{L}$ is

$$\frac{\partial \mathcal{L}}{\partial \beta_r} = \sum_{\{i:y_i=0\}} \left[\frac{-\mu_i(1+\alpha\mu_i)^{-1-\alpha^{-1}}}{\lambda_i + (1+\alpha\mu_i)^{-\alpha^{-1}}}\right] x_{ir} + \sum_{\{i:y_i>0\}} \left[\frac{y_i - \mu_i}{1+\alpha\mu_i}\right] x_{ir}, \qquad r = 1, 2, \ldots, k$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_r} = \sum_{\{i:y_i=0\}} \left[\frac{\lambda_i}{\lambda_i + (1+\alpha\mu_i)^{-\alpha^{-1}}}\right] z_{ir} - \sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i} z_{ir}, \qquad r = 1, 2, \ldots, m$$

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

The second derivatives are

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \beta_s} = \sum_{\{i:y_i=0\}} \frac{x_{ir}x_{is}\,\mu_i\left[(\mu_i-1)\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}-1\right]}{(1+\alpha\mu_i)^2\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]^2} - \sum_{\{i:y_i>0\}} \frac{\mu_i(1+\alpha y_i)x_{ir}x_{is}}{(1+\alpha\mu_i)^2},$$

$$r,s = 1,2,\ldots,k$$

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma_r \partial \gamma_s} = \sum_{\{i:y_i=0\}} \frac{z_{ir}z_{is}\,\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}}{\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]^2} - \sum_{i=1}^{n} \frac{z_{ir}z_{is}\lambda_i}{(1+\lambda_i)^2}, \qquad r,s = 1,2,\ldots,m$$

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \gamma_s} = \sum_{\{i:y_i=0\}} \frac{x_{ir}z_{is}\,\mu_i\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}-1}}{\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]^2}, \qquad r = 1,2,\ldots,k; s = 1,2,\ldots,m$$

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \alpha} = \sum_{\{i:y_i=0\}} \frac{x_{ir}\mu_i\left\{\alpha\mu_i\left[\alpha\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+\alpha\right]-\lambda_i(1+\alpha\mu_i)^{1+\alpha^{-1}}\ln(1+\alpha\mu_i)\right\}}{\alpha^2(1+\alpha\mu_i)^2\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]^2}$$

$$+ \sum_{\{i:y_i>0\}} \frac{x_{ir}\mu_i(\mu_i-y_i)}{(1+\alpha\mu_i)^2}, \qquad r = 1,2,\ldots,k$$

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma_s \partial \alpha} = \sum_{\{i:y_i=0\}} -\frac{z_{is}\lambda_i(1+\alpha\mu_i)^{\frac{1}{\alpha}-1}[(1+\alpha\mu_i)\ln(1+\alpha\mu_i)-\alpha\mu_i]}{\alpha^2\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]^2}, \qquad s = 1,2,\ldots,m$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} = \sum_{\{i:y_i=0\}} \frac{F1+F2-F3}{F4} + \sum_{\{i:y_i>0\}} (F5+F6)$$

where

$$F1 = \alpha^2\mu_i\left\{2\,\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}} + \mu_i\,\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}} + 3\,\alpha\,\mu_i\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]+2\right\}$$

$$F2 = \lambda_i(1+\alpha\mu_i)^{2+1/\alpha}\ln^2(1+\alpha\mu_i)$$

$$F3 = 2\,\alpha\,(1+\alpha\mu_i)\ln(1+\alpha\mu_i)\left\{\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}} + (1+\alpha\mu_i)^{\alpha^{-1}}\,\mu_i\,\lambda_i + \alpha\mu_i\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]+1\right\}$$

$$F4 = \alpha^4(1+\alpha\mu_i)^2\left[\lambda_i(1+\alpha\mu_i)^{\alpha^{-1}}+1\right]^2$$

$$F5 = \frac{\alpha\left[(2-2\,\alpha\,y_i)\mu_i + 3\,\alpha\,\mu_i^2 - y_i\right] - 2\,(1+\alpha\mu_i)^2\ln(1+\alpha\mu_i)}{\alpha^3(1+\alpha\mu_i)^2}$$

$$F6 = \sum_{j=0}^{y_i-1} \frac{2\alpha j+1}{(\alpha^2 j+\alpha)^2}$$

Distribution of the MLE's

The asymptotic distribution of the maximum likelihood estimates is multivariate normal as follows

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\\ \widehat{\alpha}\end{bmatrix} \sim \mathrm{N}\left[\begin{matrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\\ \alpha\end{matrix}\middle|\left(\begin{matrix}-\frac{\partial^2\mathcal{L}}{\partial\beta_r\partial\beta_s} & -\frac{\partial^2\mathcal{L}}{\partial\beta_r\partial\gamma_s} & -\frac{\partial^2\mathcal{L}}{\partial\beta_r\partial\alpha}\\ -\frac{\partial^2\mathcal{L}}{\partial\beta_r\partial\gamma_s} & -\frac{\partial^2\mathcal{L}}{\partial\gamma_r\partial\gamma_s} & -\frac{\partial^2\mathcal{L}}{\partial\gamma_s\partial\alpha}\\ -\frac{\partial^2\mathcal{L}}{\partial\beta_r\partial\alpha} & -\frac{\partial^2\mathcal{L}}{\partial\gamma_s\partial\alpha} & -\frac{\partial^2\mathcal{L}}{\partial\alpha^2}\end{matrix}\right)^{-1}\right]$$

Akaike Information Criterion (AIC)

Hilbe (2014) mentions the Akaike Information Criterion (AIC) as one of the most commonly used fit statistics. It is calculated as follows

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

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

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 its expected value estimated by the model. Because we expect 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

$$r_i = y_i - \hat{\mu}_i(1 - \hat{\pi}_i)$$

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

$$p_i = \frac{y_i - \hat{\mu}_i(1 - \hat{\pi}_i)}{\sqrt{\hat{\mu}_i(1 - \hat{\pi}_i)[1 + \hat{\mu}_i(1 + \hat{\alpha})]}}$$

Variable Selection

Because of the complexity of the model, this routine does not have a direct variable selection capability. A reasonable stepwise strategy is as follows: remove the model term (other than the intercepts) with largest p-value over 0.200 and rerun. Repeat until all p-values are less than a threshold such as 0.20.

Data Structure

At a minimum, datasets to be analyzed by ZINB regression must contain a dependent variable and one or more independent variables. Long (1990) presents a dataset of 915 rows that he uses as an example in his regression book: Long (1997). This dataset contains five independent variables (Female, MentorArts, Prestige, Married, Children) and one dependent variable (Articles).

Long 1990 Dataset

Female	MentorArts	Prestige	Married	Children	Articles
0	8	1.38	1	2	3
0	7	4.29	0	0	0
0	47	3.85	0	0	4
0	19	3.59	1	1	1
0	0	1.81	1	0	1
0	6	3.59	1	1	1
0	10	2.12	1	1	0
0	2	4.29	1	0	0
0	2	2.58	1	2	3
0	4	1.8	1	1	3

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 – Zero-Inflated Negative Binomial Regression using the Long 1990 Dataset

Long (1997) discusses a dataset used as an example of Zero-Inflated Negative Binomial regression. This dataset contains five independent variables (Female, MentorArts, Prestige, Married, Children) and one dependent variable (Articles). These variables are defined as follows

Articles	Number of articles published during the last 3 years of Ph.D.
Female	1 if female scientist; 0 if male scientist.
MentorArts	Number of articles published by the scientist mentor during the last 3 years.
Prestige	Prestige of the scientist's Ph.D. department.
Married	1 if married; 0 otherwise.
Children	Number of children 5 or younger.

The dataset can also be used to validate the program since the results of this model are given in Long (1997), page 246.

In this example, we will fit a Zero-Inflated Negative Binomial regression model to these data.

Setup

To run this example, complete the following steps:

1 Open the Long 1990 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Long 1990** and click **OK**.

2 Specify the Zero-Inflated Negative Binomial Regression procedure options

- Find and open the **Zero-Inflated Negative Binomial 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 Tab

Dependent Y	**Articles**
Numeric X's (Neg Bin Model Variables)	**Female, Married, Children, Prestige, MentorArts**
Numeric X's (Logistic Model Variables)	**Female, Married, Children, Prestige, MentorArts**

Model Tab

Terms (Neg Bin Regression Model)	**1-Way**
Terms (Logistic Regression Model)	**1-Way**

Reports Tab

All Available Plots ..... **Checked** (click the *Check All* button)
Incidence ..... **Checked**
Incidence Counts **0 1 2 3 4**
Exposure Value **1**

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	Articles	Rows Processed	915
Exposure Variable	None	Rows Used in Estimation	915
Frequency Variable	None	Observations with Y = 0	275 (30.1%)
Number of Parameters in Model	13		
Log-Likelihood	-1549.9915		
AIC(1)	3125.9830		
Number of Likelihood Iterations	12 of 100		
Convergence Setting	1E-09		
Relative Log-Likelihood Change	1.700178E-13		
Dispersion (Alpha)	0.37667		

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.

Number of Parameters in Model

This is the total number of parameters in the model. It includes those in the negative binomial portion and in the logistic portion. Note that some variables may be in both portions, but they will of course have different parameters.

Log-Likelihood

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

AIC(1)

This is Akaike's information criterion discussed above. 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.

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.

Dispersion (Alpha)

This is the estimated value of alpha, the dispersion parameter.

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.

Sum of Frequencies

This is the number of observations used by the estimation algorithm if you specified a Frequency Variable.

Observations with Y = 0

The gives the number and percentage of the observations in which Y is zero. Since this procedure is for the case in which there are too many zeros in the dataset, this value is important to consider.

Means

Means

Variable	Mean	Minimum	Maximum
Articles	1.692896	0	19
NB_Female	0.4601093	0	1
NB_Married	0.6622951	0	1
NB_Children	0.495082	0	3
NB_Prestige	3.103372	0.755	4.62
NB_MentorArts	8.767213	0	77
Lg_Female	0.4601093	0	1
Lg_Married	0.6622951	0	1
Lg_Children	0.495082	0	3
Lg_Prestige	3.103372	0.755	4.62
Lg_MentorArts	8.767213	0	77

This report gives the mean, minimum, and maximum of each variable. These values let you quickly determine if any of the data values are outside a reasonable range.

Regression Coefficients

Regression Coefficients

Parameter	Regression Coefficient b(i)	Standard Error Sb(i)	Z-Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
			Z-Statistic	Two-Sided P-Value	Lower	Upper
Negative Binomial Regression Model						
NB_Alpha	0.37667	0.05103	7.38	0.0000	0.27665	0.47668
NB_Intercept	0.41617	0.14359	2.90	0.0038	0.13473	0.69761
NB_Female	-0.19547	0.07559	-2.59	0.0097	-0.34362	-0.04731
NB_Married	0.09764	0.08445	1.16	0.2476	-0.06788	0.26317
NB_Children	-0.15173	0.05421	-2.80	0.0051	-0.25798	-0.04549
NB_Prestige	-0.00052	0.03627	-0.01	0.9886	-0.07160	0.07057
NB_MentorArts	0.02478	0.00349	7.10	0.0000	0.01794	0.03163
Logistic Regression Model						
Lg_Intercept	-0.19743	1.32205	-0.15	0.8813	-2.78861	2.39374
Lg_Female	0.63700	0.84858	0.75	0.4529	-1.02619	2.30020
Lg_Married	-1.49805	0.93791	-1.60	0.1102	-3.33633	0.34022
Lg_Children	0.62808	0.44267	1.42	0.1559	-0.23954	1.49570
Lg_Prestige	-0.03603	0.30782	-0.12	0.9068	-0.63935	0.56729
Lg_MentorArts	-0.88204	0.31622	-2.79	0.0053	-1.50182	-0.26226

Estimated Models

Negative Binomial Regression Model
Exp(0.41617395991635 -0.195467563784848*Female + 0.0976449160757309*Married
-0.151732624349505*Children -0.00051812789432887 4*Prestige + 0.0247816025774684*MentorArts)

Logistic Regression Model
Exp(-0.1974341770105 + 0.637002831769903*Female -1.49805265244775*Married
+ 0.628080059249324*Children -0.0360281307052679*Prestige -0.882040310426485*MentorArts)

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 coefficients of the ZINB regression and associated statistics. It provides the main results of the analysis.

Variable Selection

This report can be used to reduce the number of terms in the model. One variable selection strategy is to remove the model term (other than the intercepts) with largest p-value over 0.200 and rerunning. This can be repeated until all p-values are less than a threshold such as 0.20. In our example, we would definitely remove Prestige since its p-value is very high.

Parameter

This item provides the name of the parameter shown on this line of the report. Parameters that begin with "NB" are in the negative binomial portion of the model. Parameters that begin with "Lg" are in the logistic portion of the model. 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 b(i)

These are the maximum-likelihood estimates of the regression coefficients. 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. 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.

Z-Statistic

This is the z-test statistic for testing the null hypothesis that $\beta_i = 0$ against the two-sided alternative that $\beta_i \neq 0$. This is a Wald-type statistic. This test has been found to follow the normal distribution only in large samples.

The test statistic is calculated using

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

Two-Sided P-Value

The probability of obtaining a z value greater in absolute value 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.

Estimated Regression Models

These give the negative binomial and logistic models in standard, full-precision format.

Rate Ratios

Rate Ratios

Parameter	Regression Coefficient b(i)	Rate Ratio Exp(b(i))	95% Confidence Interval Limits for the Rate Ratio	
			Lower	Upper
Negative Binomial Regression Model				
NB_Female	-0.19547	0.822	0.709	0.954
NB_Married	0.09764	1.103	0.934	1.301
NB_Children	-0.15173	0.859	0.773	0.956
NB_Prestige	-0.00052	0.999	0.931	1.073
NB_MentorArts	0.02478	1.025	1.018	1.032
Logistic Regression Model				
Lg_Female	0.63700	1.891	0.358	9.976
Lg_Married	-1.49805	0.224	0.036	1.405
Lg_Children	0.62808	1.874	0.787	4.462
Lg_Prestige	-0.03603	0.965	0.528	1.763
Lg_MentorArts	-0.88204	0.414	0.223	0.769

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

This report provides the rate ratio for each independent variable.

Parameter

This item provides the name of the parameter shown on this line of the report. Parameters that begin with “NB” are in the negative binomial portion of the model. Parameters that begin with “Lg” are in the logistic portion of the model.

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 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.

Residuals

Residuals

Row	Articles (Y)	Conditional Mean of Y E(Y|X,Z)	Residual Raw Y - E(Y|X,Z)	Residual Pearson Raw/σ	(T)
1	3	1.5028	1.4972	0.9756	1
2	0	1.7967	-1.7967	-1.0340	1
3	4	4.8497	-0.8497	-0.2295	1
4	1	2.2958	-1.2958	-0.6263	1
5	1	1.4251	-0.4251	-0.2601	1
6	1	1.6610	-0.6610	-0.4018	1
7	0	1.8381	-1.8381	-1.0421	1
8	0	1.7067	-1.7067	-1.0002	1
9	3	1.1767	1.8233	1.3263	1
10	3	1.5697	1.4303	0.8993	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This report provides the conditional mean (predicted value), the raw residual, and the Pearson residual. Large residuals indicate data points that were not fit well by the model.

Predicted Means

Predicted Means

Row	Articles (Y)	Negative Binomial Mean μ	Logit CDF Pr(Y = 0) π	Conditional Mean of Y E(Y|X,Z)	Standard Error of E(Y|X,Z) σ	95% Confidence Interval Limits for E(Y|X,Z) Lower	Upper
1	3	1.5036	0.0005	1.5028	1.5347	-1.5051	4.5108
2	0	1.7993	0.0015	1.7967	1.7376	-1.6089	5.2023
3	4	4.8497	0.0000	4.8497	3.7025	-2.4071	12.1065
4	1	2.2958	0.0000	2.2958	2.0691	-1.7595	6.3511
5	1	1.6701	0.1467	1.4251	1.6342	-1.7780	4.6281
6	1	1.6635	0.0015	1.6610	1.6450	-1.5631	4.8850
7	0	1.8382	0.0000	1.8381	1.7638	-1.6189	5.2952
8	0	1.7527	0.0262	1.7067	1.7064	-1.6378	5.0513
9	3	1.2951	0.0914	1.1767	1.3748	-1.5178	3.8712
10	3	1.5845	0.0094	1.5697	1.5905	-1.5477	4.6871
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

This report provides the predicted values along with their standard errors and confidence interval limits. It also provides the mean of the negative binomial portion of the model (μ) and the probability that Y = 0 from the logistic portion of the model.

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 = 1

Incidence when Exposure = 1

Row	Average Incidence Rate	Probability that Count is 0	1	2	3	4
1	1.5028	0.3042	0.2915	0.1926	0.1081	0.0552
2	1.7967	0.2542	0.2711	0.2001	0.1254	0.0716
3	4.8497	0.0634	0.1087	0.1284	0.1287	0.1176
4	2.2958	0.1912	0.2354	0.1995	0.1436	0.0941
5	1.4251	0.3803	0.2395	0.1690	0.1012	0.0553
6	1.6610	0.2759	0.2807	0.1976	0.1181	0.0643
7	1.8381	0.2474	0.2687	0.2009	0.1275	0.0738
8	1.7067	0.2797	0.2676	0.1945	0.1200	0.0675
9	1.1767	0.4078	0.2754	0.1650	0.0840	0.0389
10	1.5697	0.2953	0.2837	0.1938	0.1124	0.0594
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

This report gives the average incidence rate and estimated probabilities of 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

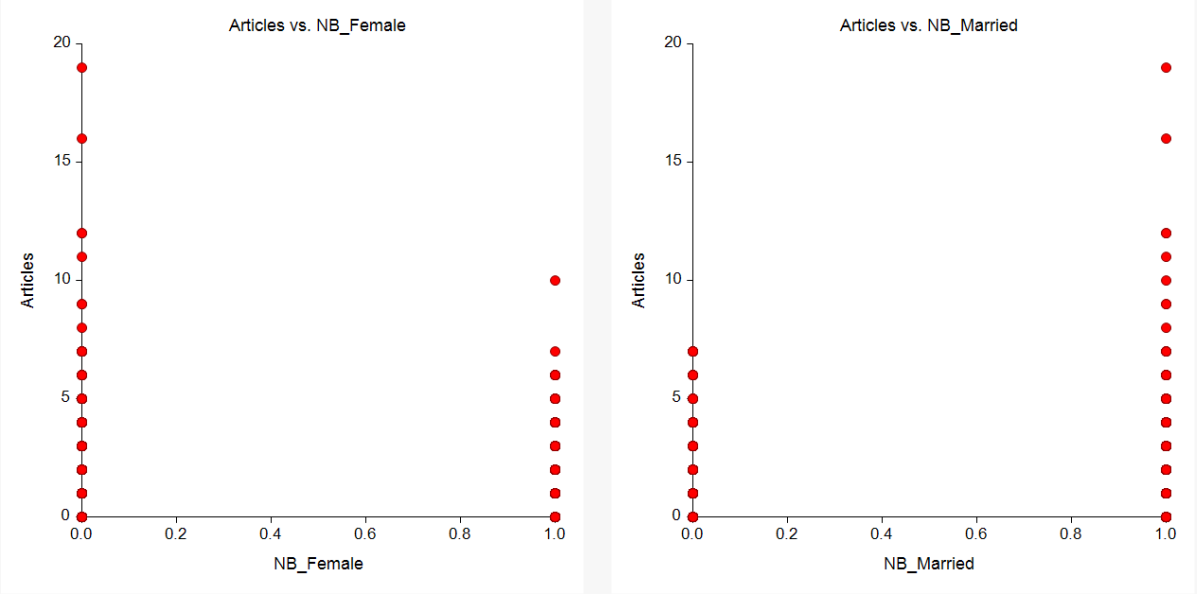
This is the predicted incidence rate. Note that the calculation is made for the specified exposure value, not the value of *T* on the database. This allows you to make valid comparisons of the incidence rates.

Probability that Count is Y

Using the ZINB model, the probability of obtaining exactly *Y* events during the exposure 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)



(three more similar plots are shown)

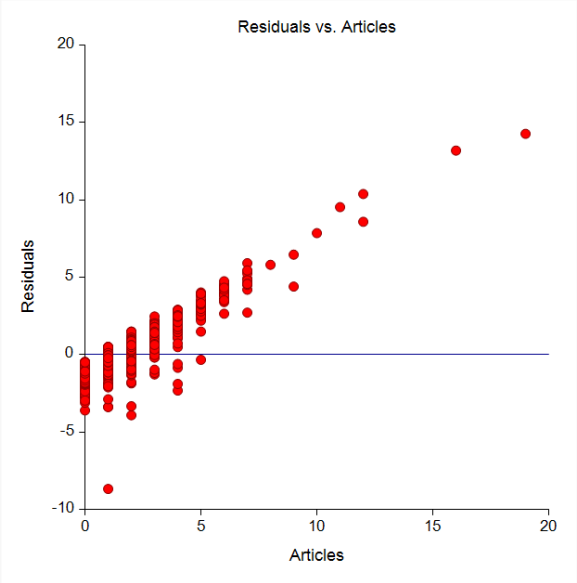
These plots show each of the independent variables plotted against the incidence as measured by Y/T. They should be scanned for outliers and curvilinear patterns.

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

These plots are similar to the incidence versus X plots and are used for the same purpose, so we have not shown them here. They should be scanned for outliers and curvilinear patterns.

Residuals vs Y Plot

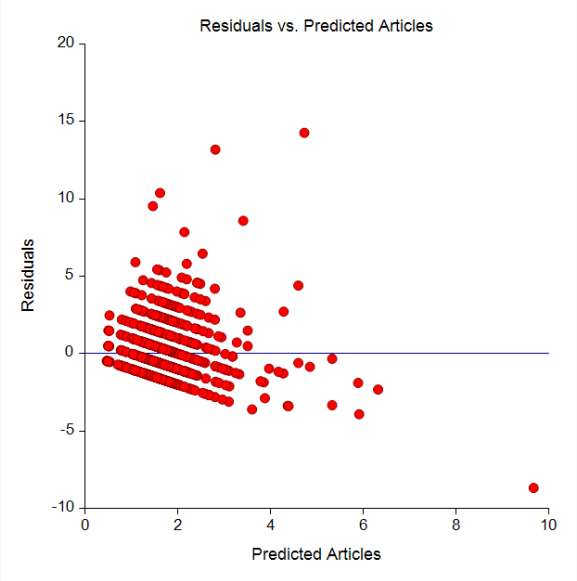
Residuals vs Y Plot



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

Residuals vs Yhat (Predicted Y) Plot

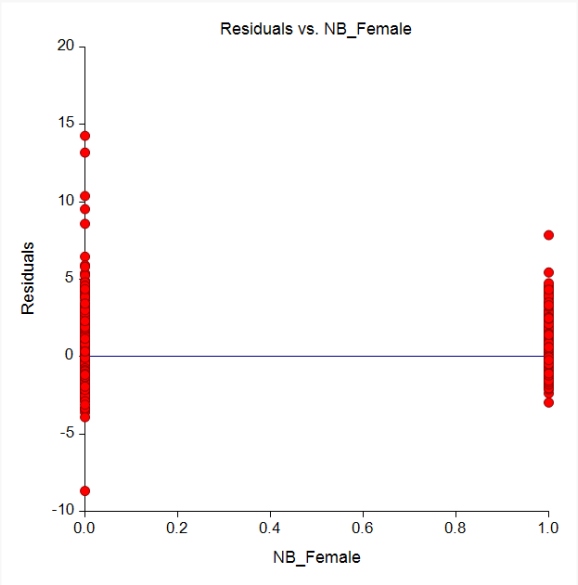
Residuals vs Yhat (Predicted Y) Plot

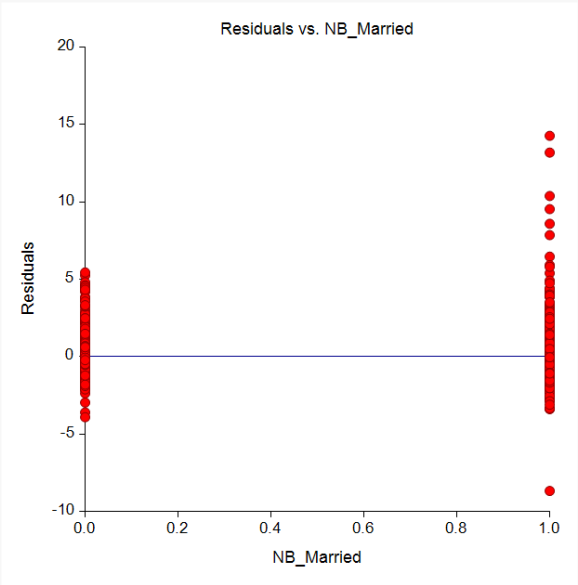


This plot shows the residuals versus the predicted value (Yhat) of the dependent variable. It can show outliers.

Residuals vs X Plots

Residuals vs X Plot(s)





(three more similar plots are shown)

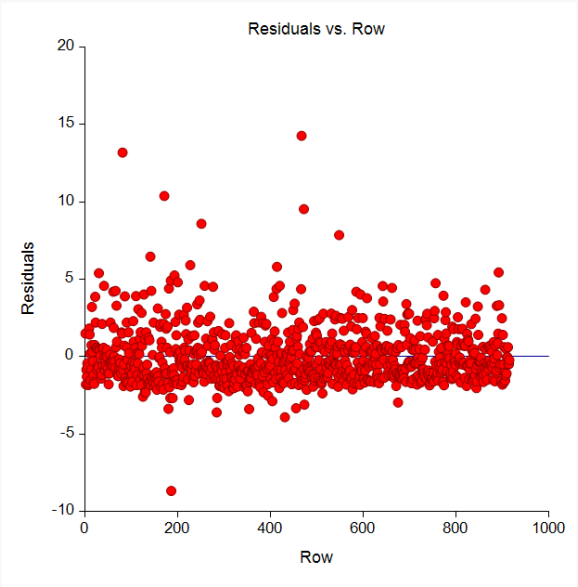
These plots show the residuals plotted against the independent variables. They are used to spot outliers. They are also used to find curvilinear patterns that are not represented in the regression model.

Residuals vs Z Plot(s)

These plots are similar to the residual versus X plots and are used for the same purpose, so we have not shown them here. They should be scanned for outliers and curvilinear patterns. 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.