

## 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  $\pi$  and case 2 occurs with probability  $1 - \pi$ . 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  $\pi_i$  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

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

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

## Zero-Inflated Negative Binomial Regression

This logistic link function  $\pi_i$  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

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

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

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

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

The gradient of  $\mathcal{L}$  is

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

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

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

The second derivatives are

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

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

## Zero-Inflated Negative Binomial Regression

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

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \alpha} = & \sum_{\{i:y_i=0\}} \frac{x_{ir} \mu_i \{ \alpha \mu_i [\alpha \lambda_i (1 + \alpha \mu_i)^{\alpha^{-1}} + \lambda_i (1 + \alpha \mu_i)^{\alpha^{-1}} + \alpha] - \lambda_i (1 + \alpha \mu_i)^{1+\alpha^{-1}} \ln(1 + \alpha \mu_i) \}}{\alpha^2 (1 + \alpha \mu_i)^2 [\lambda_i (1 + \alpha \mu_i)^{\alpha^{-1}} + 1]^2} \\ & + \sum_{\{i:y_i>0\}} \frac{x_{ir} \mu_i (\mu_i - y_i)}{(1 + \alpha \mu_i)^2} \\ & r = 1, 2, \dots, k \end{aligned}$$

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

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

where

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

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

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

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

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

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

## Distribution of the MLE's

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

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \\ \hat{\alpha} \end{bmatrix} \sim N \left[ \begin{bmatrix} \beta \\ \gamma \\ \alpha \end{bmatrix}, \begin{bmatrix} -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \beta_s} & -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \gamma_s} & -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \alpha} \\ -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \gamma_s} & -\frac{\partial^2 \mathcal{L}}{\partial \gamma_r \partial \gamma_s} & -\frac{\partial^2 \mathcal{L}}{\partial \gamma_s \partial \alpha} \\ -\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \alpha} & -\frac{\partial^2 \mathcal{L}}{\partial \gamma_s \partial \alpha} & -\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} \end{bmatrix}^{-1} \right]$$

---

## 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.

## Zero-Inflated Negative Binomial Regression

## 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.

## Procedure Options

This section describes the options available in this procedure.

## Variables Tab

This panel specifies the variables that 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.

## Zero-Inflated Negative Binomial Regression

---

**Negative Binomial Model Variables**
**Negative Binomial Numeric X's**

Specify the numeric (continuous) independent variables in the negative binomial portion of the model. It is not necessary to specify any of these variables if at least one Negative Binomial Categorical X variable is specified.

These variables do not necessarily have to be the same as the Logistic Numerical X 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 Negative Binomial Model on 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.

**Negative Binomial Categorical X's**

Specify categorical (nominal or group) independent variables to be used in the negative binomial portion of the model. It is not necessary to specify any Negative Binomial Categorical X variables if at least one Negative Binomial Numeric X variable is specified.

These variables do not necessarily have to be the same as the Logistic Categorical X Variables.

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 be 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

## Zero-Inflated Negative Binomial Regression

- **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

## Zero-Inflated Negative Binomial Regression

- **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



## Zero-Inflated Negative Binomial Regression

- **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 New York. 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

## Zero-Inflated Negative Binomial Regression

- **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

## Zero-Inflated Negative Binomial Regression

- **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.

## Zero-Inflated Negative Binomial Regression

---

### Dispersion Parameter ( $\alpha$ )

#### $\alpha$ Input Type

Select the type of input you would like to use to specify the dispersion parameter  $\alpha$ .

The choices are:

#### Estimate $\alpha$ from the data

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

#### Enter $\alpha$ directly

Enter a fixed value for  $\alpha$ . You might do this for one of two reasons:

1. To speed up the convergence once a reasonable value of  $\alpha$  has been determined.
2. To force the fit of a geometric regression model which is the name of the special case in which  $\alpha$  is equal to one.

#### $\alpha$ (Dispersion)

The dispersion parameter  $\alpha$  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 Poisson distribution forces the unrealistic assumption that the variance is equal to the mean).

#### Range

$\alpha$  must be greater than zero and is usually less than 4.

#### Why specify $\alpha$

1. Maximum likelihood estimation may not converge if  $\alpha$  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  $\alpha$ .
3. Geometric Regression is a special case of negative binomial regression in which  $\alpha$  is set to one.

---

### Logistic (Zero Inflation) Model Variables

#### Logistic Numeric X's

Specify the numeric (continuous) independent variables in the logistic portion of the model. It is not necessary to specify any of these variables if at least one Logistic Categorical X variable is specified.

These variables do not necessarily have to be the same as the Negative Binomial Numerical X 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 Negative Binomial Model on 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.

## Zero-Inflated Negative Binomial Regression

### Logistic Categorical X's

Specify categorical (nominal or group) independent variables to be used in the negative binomial portion of the model. It is not necessary to specify any Logistic Categorical X variables if at least one Logistic Numeric X variable is specified.

These variables do not necessarily have to be the same as the Negative Binomial Categorical X Variables.

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 be 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

## Zero-Inflated Negative Binomial Regression

- **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

## Zero-Inflated Negative Binomial Regression

- **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 New York. Suppose the recoding scheme is specified as *Compare Each with Reference Value* with the reference value of *California*. You would enter

**State(R;California)**

## Zero-Inflated Negative Binomial Regression

## 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



## Zero-Inflated Negative Binomial Regression

- **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

## Zero-Inflated Negative Binomial Regression

- **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.

---

## Models Tab

This panel specifies the negative binomial portion of the regression model and the logistic portion of the regression model. These models are designated separately. They may or may not contain same variables.

---

### Negative Binomial and Logistic Regression Models

#### 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 *I-Way*.

## Zero-Inflated Negative Binomial Regression

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.

## Zero-Inflated Negative Binomial Regression

### Remove Intercept

Unchecked indicates that the intercept term,  $\beta_0$ , 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 two 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.

$$A|A|B|B \text{ (Max Term Order=2)} = A + B + A*A + A*B + B*B$$

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

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

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

---

## 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 100 iterations, so a value of one hundred 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 – Estimation

#### Regression Coefficients ... Rate Coefficients

Indicate whether to display these estimation 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.

## Zero-Inflated Negative Binomial Regression

### Incidence Counts

Up to five incidence counts may be entered. The probabilities of these counts under the fitted ZINB 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 $\geq$

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.

## Zero-Inflated Negative Binomial Regression

### 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 two 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.

## Zero-Inflated Negative Binomial Regression

- **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 – 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.

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 Zero-Inflated Negative Binomial Regression window.

### 1 Open the Long 1990 dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Long 1990.NCSS**.
- Click **Open**.

### 2 Open the Zero-Inflated Negative Binomial Regression window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Zero-Inflated 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 procedure window, select the **Variables tab**.
- Double-click in the **Dependent Y** box. This will bring up the variable selection window.
- Select **Articles** from the list of variables and click **Ok**. *Articles* will appear in the **Dependent Y** box.
- Leave the **Exposure T** variable box blank.
- Double-click in the **Numeric X's** box under the **Negative Binomial Model Variables** heading.
- Select **Female, Married, Children, Prestige, MentorArts** and then click **Ok**.
- Double-click in the **Numeric X's** box under the **Logistic (Zero Inflation) Model Variables** heading.
- Select **Female, Married, Children, Prestige, MentorArts** and then click **Ok**.

### 4 Specify the model.

- On the procedure window, select the **Models tab**.
- Set the **Terms** option to **1-Way** for both the negative binomial and the logistic models.

## Zero-Inflated Negative Binomial Regression

**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 here for documentation purposes.
- Set the **Incidence Counts** to **0 1 2 3 4**.
- Set the **Exposure Value** to **1.0**.

**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	Articles	Rows Used	915
Exposure Variable	None	Sum of Frequencies	915
Frequency Variable	None	Sum of Frequencies that Y = 0	275 (30.1%)
Parameters in Model	13	Iterations	12
Log Likelihood	-1549.9915	Convergence Setting	1E-09
AIC(1)	3125.9830	Relative LogLike Change	1.706045E-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.

**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.

**Dispersion (Alpha)**

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

**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.

**Sum of Frequencies that Y = 0**

This 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.

## Zero-Inflated Negative Binomial Regression

### 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 LogLike Change

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

## Means Report

Variable	Mean	Minimum	Maximum
Articles	1.693	0.000	19.000
NB_Female	0.460	0.000	1.000
NB_Married	0.662	0.000	1.000
NB_Children	0.495	0.000	3.000
NB_Prestige	3.103	0.755	4.620
NB_MentorArts	8.767	0.000	77.000
Lg_Female	0.460	0.000	1.000
Lg_Married	0.662	0.000	1.000
Lg_Children	0.495	0.000	3.000
Lg_Prestige	3.103	0.755	4.620
Lg_MentorArts	8.767	0.000	77.000

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 of Negative Binomial and Logistic Models

Parameter	Regression Coefficient b(j)	Standard Error Sb(i)	Z Value H0: $\beta=0$	Two-Sided P-Value	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
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
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 Regression Models

#### Negative Binomial Regression Model

Exp( 0.416173959916351 -0.195467563784848\*Female + 0.0976449160757307\*Married -0.151732624349505\*Children - 0.000518127894329069\*Prestige + 0.0247816025774684\*MentorArts )

#### Logistic Regression Model

Exp( -0.197434177010499 + 0.637002831769902\*Female -1.49805265244775\*Married + 0.628080059249325\*Children - 0.0360281307052683\*Prestige -0.882040310426485\*MentorArts )

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

## Zero-Inflated Negative Binomial Regression

### 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  $\geq$*  option in the Format tab.

### Regression Coefficient

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

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 Value H0: $\beta=0$

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.

### 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.

### Estimated Regression Models

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

## Zero-Inflated Negative Binomial Regression

## Rate Report

Parameter	Regression Coefficient b(i)	Rate Ratio Exp(b(i))	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
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
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 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

These are the maximum-likelihood estimates of the regression coefficients,  $b_1, b_2, \dots, b_k$ . 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 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

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

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

## Zero-Inflated Negative Binomial Regression

### Residuals Report

Row	Articles (Y)	Conditional Mean of Y $E(Y X,Z)$	Raw Residual $Y - E(Y X,Z)$	Pearson Residual $[Y - E(Y X,Z)] / \sigma$	(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 Value Report

Row	Articles (Y)	Neg Bin Mean $\mu$	Logit CDF $Pr(Y=0)$ $\pi$	Conditional Mean of Y $E(Y X,Z)$	Std Error of $E(Y X,Z)$ $\sigma$	Lower 95.0% Conf Limit of $E(Y X,Z)$	Upper 95.0% Conf Limit of $E(Y X,Z)$
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 limits. It also provides the mean of the negative binomial portion of the model ( $\mu$ ) 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 Report when Exposure = 100000

Row	Average Incidence Rate	Prob that Count is 0	Prob that Count is 1	Prob that Count is 2	Prob that Count is 3	Prob that Count is 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.

## Zero-Inflated Negative Binomial Regression

### 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

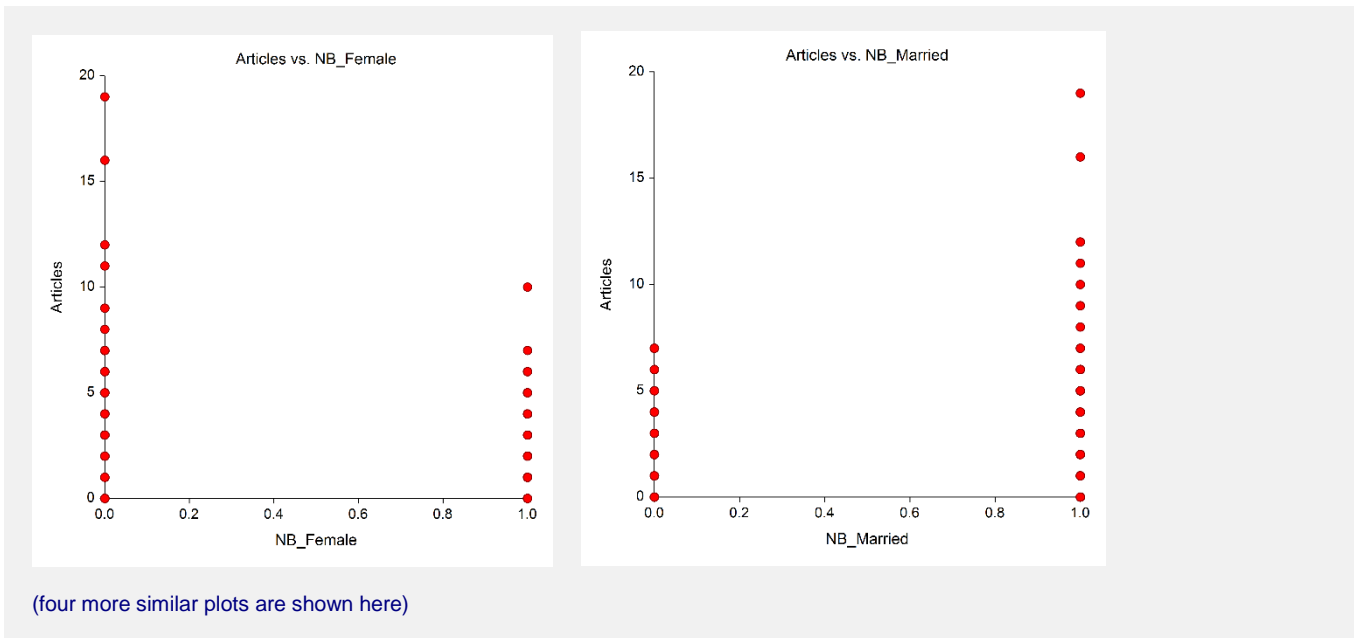
This is the predicted incidence rate. Note that the calculation is made for the specified exposure value, not the value of  $T$  on the database. This allows you to make valid comparisons of the incidence rates.

### Prob 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.

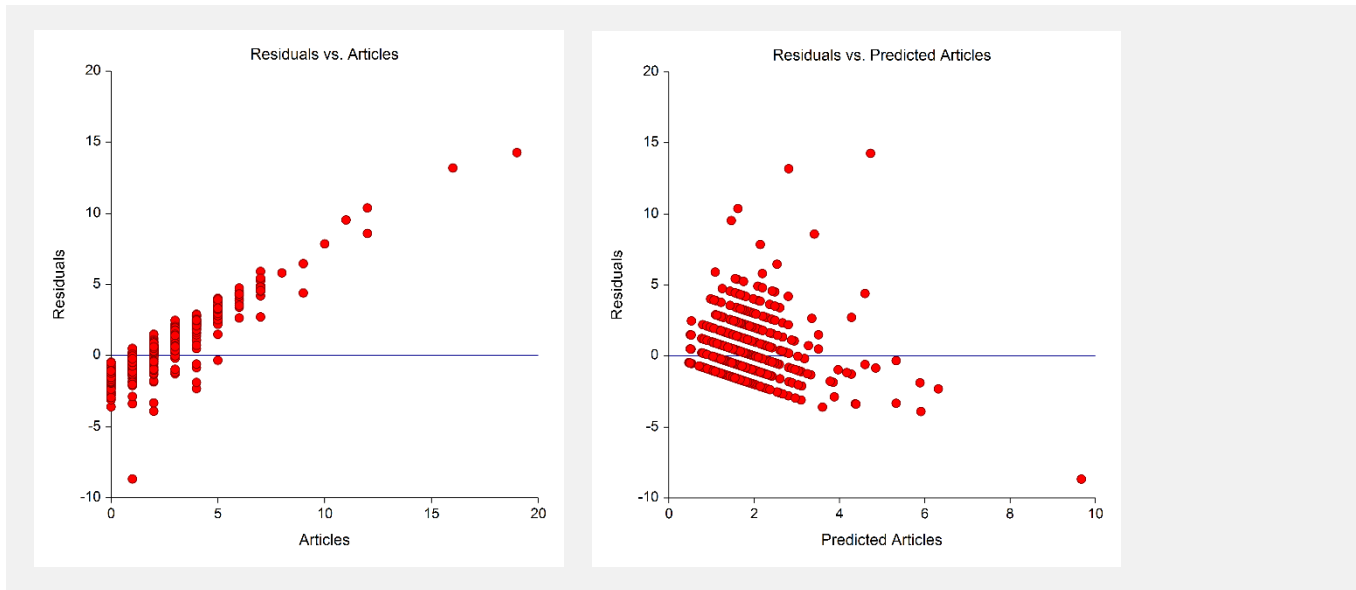
---

## Plots of Y vs Each Parameter

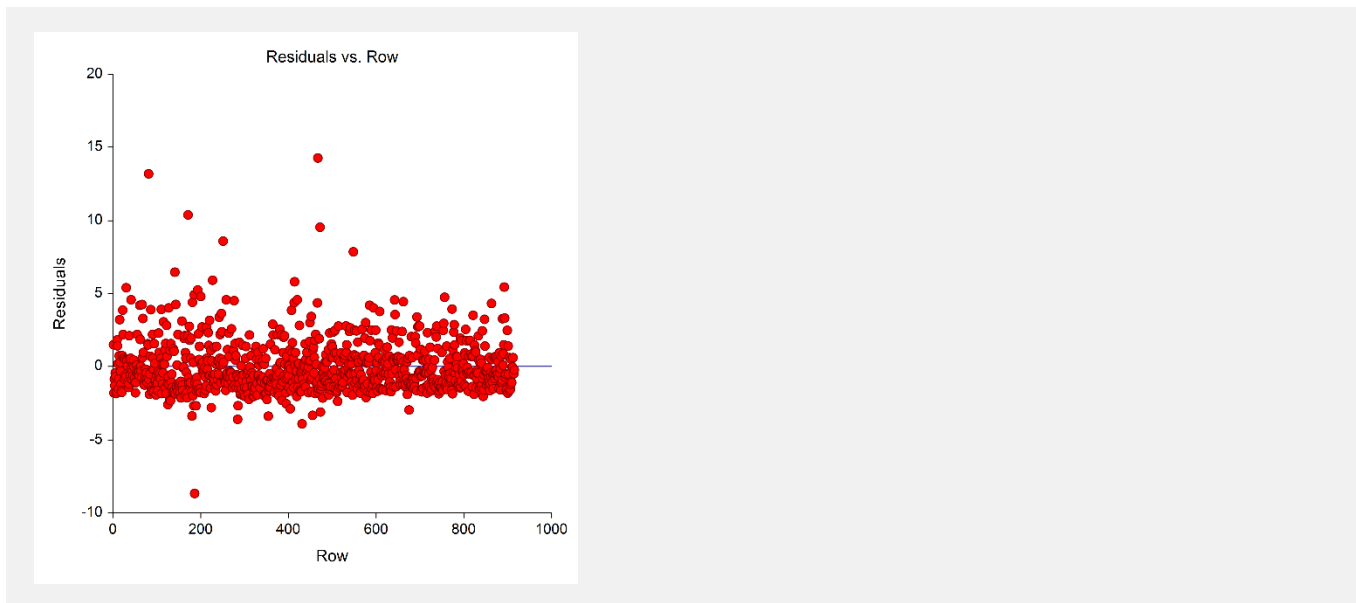


These plots show each of the independent variables plotted against the incidence as measured by  $Y/T$ . They should be scanned for outliers and curvilinear patterns.

## Zero-Inflated Negative Binomial Regression

**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.

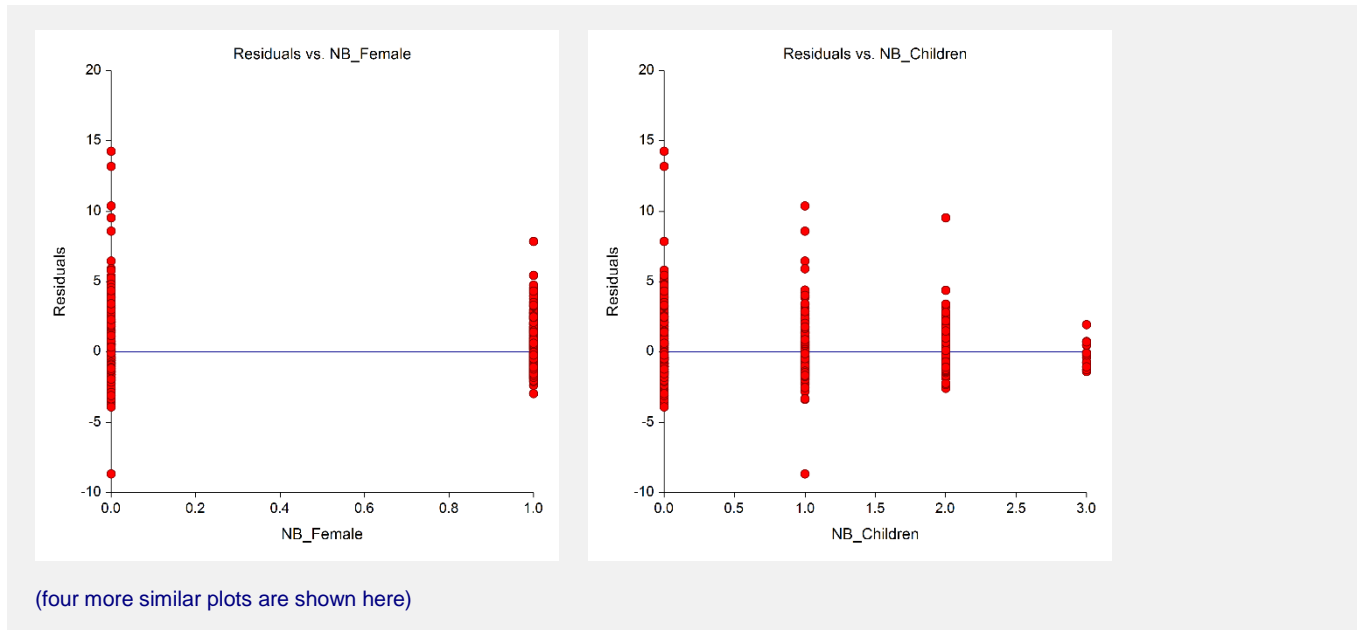
**Plots of Residuals and Row**

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



## Zero-Inflated Negative Binomial Regression

## Plots of Residuals and X's



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.