Chapter 11
The COUNTREG Procedure (Experimental)

Chapter Contents

Chapter 11
The COUNTREG Procedure (Experimental)

Overview

The COUNTREG (Count Regression) procedure analyzes regression models in which the dependent variable takes nonnegative integer or count values. The dependent variable is usually an *event count*, which refers to the number of times an event occurs. For example, an event count might represent the number of ship accidents per year for a given fleet. In count regression, the conditional mean of the dependent variable, y, is assumed to be a function of a vector of covariates, $\mathbf{x}$.

PROC COUNTREG supports the following models for count data:

- Poisson regression
- negative binomial regression with quadratic (NEGBIN2) and linear (NEGBIN1) variance functions (Cameron and Trivedi 1986)
- zero-inflated Poisson (ZIP) model (Lambert 1992)
- zero-inflated negative binomial (ZINB) model

In recent years, count data models have been used extensively in economics, political science, and sociology. For example, Hausman, Hall, and Griliches (1984) examine the effects of R&D expenditures on the number of patents received by U.S. companies. Cameron and Trivedi (1986) study factors affecting the number of doctor visits. Greene (1994) studies the number of derogatory reports to a credit reporting agency for a group of credit card applicants. As a final example, Long (1997) analyzes the number of doctoral publications in the final three years of Ph.D. studies.

The COUNTREG procedure uses maximum likelihood estimation. When a model with a dependent count variable is estimated using linear ordinary least squares (OLS) regression, the count nature of the dependent variable is ignored. This leads to parameter estimates with undesirable properties in terms of efficiency, consistency, and unbiasedness unless the mean of the counts is high, in which case the Gaussian approximation and linear regression may be satisfactory. The Poisson (log-linear) regression model is the most basic model that explicitly takes into account the nonnegative integer-valued aspect of the outcome. With this model, the probability of an event count is determined by a Poisson distribution, where the conditional mean of the distribution is a function of a vector of covariates. However, the basic Poisson regression model is limited because it forces the conditional mean of the outcome, y, to equal the conditional variance. This assumption is often violated in real-life data. Negative binomial regression is an extension of Poisson regression in which the conditional variance may exceed the conditional mean. Also, an often encountered

characteristic of count data is that the number of zeros in the sample exceeds the number of zeros predicted by either the Poisson or negative binomial models. Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models explicitly model the production of zero counts to account for excess zeros and also allow the conditional variance of the outcome to differ from the conditional mean.

Getting Started

The COUNTREG procedure is similar in use to other regression model procedures in the SAS System. For example, the following statements are used to estimate a Poisson regression model:

```
proc countreg data=one type=poisson;
   model y = x1 ;
run;
```

The response variable y is numeric and has nonnegative integer values. You can also specify the negative binomial model as follows:

```
proc countreg data=one type=negbin;
```

The following example illustrates the use of PROC COUNTREG. The data are taken from Long (1997). This study examines how factors such as gender, marital status, number of young children, prestige of the graduate program, and the number of articles published by a scientist's mentor affect the number of articles published by the scientist.

```
data one;
   input fem ment phd mar kid5 art lnart;
   datalines;
      ... data lines are omitted ...
   ;
```

The first 10 observations are shown in Figure 11.1.

Obs	art	fem	mar	kid5	phd	ment
1	3	0	1	2	1.38000	8.0000
2	0	0	0	0	4.29000	7.0000
3	4	0	0	0	3.85000	47.0000
4	1	0	1	1	3.59000	19.0000
5	1	0	1	0	1.81000	0.0000
6	1	0	1	1	3.59000	6.0000
7	0	0	1	1	2.12000	10.0000
8	0	0	1	0	4.29000	2.0000
9	3	0	1	2	2.58000	2.0000
10	3	0	1	1	1.80000	4.0000

Figure 11.1. Article Count Data

The following SAS statements estimate the Poisson regression model:

```
proc countreg data=one type=poisson  method=qn;
   model art = fem mar kid5 phd ment ;
run;
```

The fit summary table is listed in Figure 11.2. First, PROC COUNTREG lists the estimation summary table. By default, the COUNTREG procedure uses the Newton-Raphson optimization technique. This table shows the maximum log-likelihood value as well as two information measures: Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (SBC).

The COUNTREG Procedure

Poisson Regression Estimates

Model Fit Summary

Dependent Variable	art
Number of Observations	915
Log Likelihood	-1651
Maximum Absolute Gradient	2.40759E-7
Number of Iterations	7
Optimization Method	Newton-Raphson
AIC	3314
Schwarz Criterion	3343

Figure 11.2. Estimation Summary Table for a Poisson Regression

The parameter estimates and standard errors are shown in Figure 11.3.

The COUNTREG Procedure

Poisson Regression Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|	Gradient
Intercept	1	0.3046	0.1030	2.96	0.0031	-5.25E-9
fem	1	-0.2246	0.0546	-4.11	<.0001	-1.04E-9
mar	1	0.1552	0.0614	2.53	0.0114	-4.25E-9
kid5	1	-0.1849	0.0401	-4.61	<.0001	-4.2E-9
phd	1	0.0128	0.0264	0.49	0.6271	-1.68E-8
ment	1	0.0255	0.002006	12.73	<.0001	-2.41E-7

Figure 11.3. Parameter Estimates of Poisson Regression

The negative binomial regression model is more general than the Poisson regression model. Whereas the Poisson regression model requires that the conditional mean and conditional variance be equal, the negative binomial regression model allows for overdispersion; that is, the conditional variance may exceed the conditional mean. The following statements fit the negative binomial regression model:

```
proc countreg data=one type=negbin method=qn;
   model art = fem mar kid5 phd ment ;
run;
```

The fit summary is shown in Figure 11.4, and parameter estimates are listed in Figure 11.5.

The COUNTREG Procedure

Negative Binomial Regression Estimates

Model Fit Summary

Dependent Variable	art
Number of Observations	915
Log Likelihood	-1561
Maximum Absolute Gradient	0.0002304
Number of Iterations	9
Optimization Method	Newton-Raphson
AIC	3136
Schwarz Criterion	3170

Figure 11.4. Estimation Summary Table for a Negative Binomial Regression

The COUNTREG Procedure

Negative Binomial Regression Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|	Gradient
Intercept	1	0.2561	0.1386	1.85	0.0645	0.000038
fem	1	-0.2164	0.0727	-2.98	0.0029	9.695E-6
mar	1	0.1505	0.0821	1.83	0.0668	5.834E-7
kid5	1	-0.1764	0.0531	-3.32	0.0009	0.000017
phd	1	0.0153	0.0360	0.42	0.6718	0.000115
ment	1	0.0291	0.003470	8.38	<.0001	0.00023
ALPHA	1	0.4416	0.0530	8.34	<.0001	-0.00004

Figure 11.5. Parameter Estimates of Negative Binomial Regression

The parameter estimate for _Alpha of 0.4416 is an estimate of the dispersion parameter in the negative binomial distribution. A likelihood ratio test of $H_0 : \alpha = 0$ can be carried out: $-2(\mathcal{L}_P - \mathcal{L}_{NB}) = -2(-1651 + 1561) = 180$, which is highly significant. Thus, there is strong evidence of overdispersion.

Syntax

The COUNTREG procedure is controlled by the following statements:

```
PROC COUNTREG options ;
  BOUNDS bound1 [ , bound2 . . . ] ;
  BY variables ;
  INIT initvalue1 [ , initvalue2 . . . ] ;
  MODEL dependent variables = regressors / options ;
  RESTRICT options ;
```

Functional Summary

The statements and options used with the COUNTREG procedure are summarized in the following table:

Description	Statement	Option
Data Set Options		
specify the input data set	COUNTREG	DATA=
write parameter estimates to an output data set	COUNTREG	OUTEST=
Declaring the Role of Variables		
specify BY-group processing	BY	
Printing Control Options		
print the correlation matrix of the estimates	COUNTREG	CORRB
print the covariance matrix of the estimates	COUNTREG	COVB
print a summary iteration listing	COUNTREG	ITPRINT
suppress the normal printed output	COUNTREG	NOPRINT
request all printing options	COUNTREG	PRINTALL
Options to Control the Optimization Process		
specify the maximum number of iterations allowed	COUNTREG	MAXITER=
select the iterative minimization method to use	COUNTREG	METHOD=
set boundary restrictions on parameters	BOUNDS	
set initial values for parameters	INIT	
set linear restrictions on parameters	RESTRICT	
Model Estimation Options		
specify the type of model	COUNTREG	TYPE=
specify the type of covariance matrix	COUNTREG	COVEST=
suppress the intercept parameter	MODEL	NOINT
specify the offset variable	MODEL	OFFSET=

Description	Statement	Option
specify the P_OBS variable (probability of an event being observed, given that it occurred)	MODEL	P_OBS=
specify options specific to zero-inflated count regression	MODEL	ZI()
Output Control Options		
include covariances in the OUTEST= data set	COUNTREG	COVOUT

PROC COUNTREG Statement

PROC COUNTREG *options* **;**

The following options can be used in the PROC COUNTREG statement:

Data Set Options

DATA= SAS-data-set
specifies the input SAS data set. If the DATA= option is not specified, PROC COUNTREG uses the most recently created SAS data set.

Output Data Set Options

OUTEST= SAS-data-set
writes the parameter estimates to an output data set.

COVOUT
writes the covariance matrix for the parameter estimates to the OUTEST= data set. This option is valid only if the OUTEST= option is specified.

Printing Options

CORRB
prints the correlation matrix of the parameter estimates.

COVB
prints the covariance matrix of the parameter estimates.

ITPRINT
prints the objective function and parameter estimates at each iteration. The objective function is the negative log-likelihood function.

NOPRINT
suppresses all printed output.

PRINTALL
requests all printing options.

Estimation Control Options

TYPE= *value*
specifies a type of model to be analyzed. The supported model types are as follows:

POISSON	specifies a Poisson regression model
NEGATIVEBINOM1 | NEGBIN1	specifies a negative binomial regression model with a linear variance function
NEGATIVEBINOM | NEGBIN	specifies a negative binomial regression model with a quadratic variance function
ZIPOISSON | ZIP	specifies a zero-inflated Poisson regression
ZINEGBIN | ZINB	specifies a zero-inflated negative binomial regression

COVEST=*value*
The COVEST= option specifies the type of covariance matrix. When COVEST=OP is specified, the outer product matrix is used to compute the covariance matrix of the parameter estimates. The COVEST=HESSIAN option produces the covariance matrix using the Hessian matrix. The quasi-maximum likelihood estimates are computed with COVEST=QML. The supported covariance types are as follows:

OP	specifies covariance from outer product matrix
HESSIAN	specifies covariance from Hessian matrix
QML	specifies covariance from outer product and Hessian matrices

Options to Control the Optimization Process

The following options might be helpful when you experience a convergence problem:

MAXITER= *number*
sets the maximum number of iterations allowed. The default is MAXITER=100.

METHOD= *value*
specifies the iterative minimization method to use. METHOD=QN specifies the quasi-Newton method, METHOD=NRA specifies the Newton-Raphson method, and METHOD=TR specifies the trust region method. The default is METHOD=NRA.

BOUNDS Statement

BOUNDS *bound1 [, bound2 ...]* **;**

The BOUNDS statement imposes simple boundary constraints on the parameter estimates. BOUNDS statement constraints refer to the parameters estimated by the COUNTREG procedure. You can specify any number of BOUNDS statements.

Each *bound* is composed of variables, constants, and inequality operators:

item operator item [operator item [operator item . . .]]

Each *item* is a constant, the name of a regressor variable, or a list of regressor names. Each *operator* is '<', '>', '<=', or '>='.

You can use both the BOUNDS statement and the RESTRICT statement to impose boundary constraints; however, the BOUNDS statement provides a simpler syntax for specifying these kinds of constraints. See the "RESTRICT Statement" section on page 428 as well.

The following BOUNDS statement constrains the estimates of the coefficient of `z` to be negative and the coefficients of `x1` through `x10` to be between zero and one. This example illustrates the use of parameter lists to specify boundary constraints.

```
bounds z < 0,
       0 < x1-x10 < 1;
```

BY Statement

BY *variables* **;**

A BY statement can be used with PROC COUNTREG to obtain separate analyses on observations in groups defined by the BY variables.

INIT Statement

INIT *initvalue1 [, initvalue2 . . .]* **;**

The INIT statement is used to set initial values for parameters in the optimization.

Each *initvalue* is written as a parameter or parameter list, followed by an optional equals sign (=), followed by a number:

parameter [=] number

MODEL Statement

MODEL *dependent = regressors / options* **;**

The MODEL statement specifies the dependent variable and independent regressor variables for the regression model.

The following options can be used in the MODEL statement after a slash (/).

NOINT
suppresses the intercept parameter.

OFFSET=*variable*
specifies a variable in the input data set to be used as an offset variable. The offset variable appears as a term in the link function with a coefficient of 1. In contrast, the jth explanatory variable (regressor) of the ith observation appears as a term in the link function with a coefficient of β_j. Thus, the regression does not estimate a coefficient for the offset variable; it is fixed at 1. The offset variable cannot be the response variable, zero-inflation offset variable (if any), or one of the explanatory variables.

P_OBS=*variable*
specifies a variable in the input data set that represents the probability of an event being observed (or counted), once it has occurred. A P_OBS value of 1 means that every event that occurred was also counted (observed). The P_OBS variable can only take on values that are strictly greater than 0 and less than or equal to 1. If the P_OBS= option is included on the model statement, and if an observation in the input data set contains an invalid (or missing) P_OBS value, then that observation will be excluded from the regression.

Zero-Inflated Count Data Regression Options

ZI(*option-list***)**
specifies options that are used for zero-inflated Poisson and negative binomial models.

The following options can be used in the ZI() option. The options are listed within parentheses and separated by commas.

LINK=*value*
specifies the distribution function used to compute probability of zeros. The supported distribution functions are as follows:

LOGISTIC	specifies logistic distribution
NORMAL	specifies standard normal distribution

OFFSET=*variable*
specifies a variable in the input data set to be used as a zero-inflated offset variable (ZI offset variable). The ZI offset variable is included as a term, with coefficient 1, in the equation that determines the probability (φ_i) of the observed count being zero.

The ZI offset variable cannot be the response variable, the offset variable (if any), or one of the explanatory variables.

VAR=*variables*

specifies the zero-inflated explanatory variables (ZI explanatory variables) that are used in the equation that determines the probability (φ_i) of the observed count being zero. Each of these q variables, $q \geq 0$, has a coefficient that must be estimated in the regression. For example, let $\mathbf{w}_i'$ be the ith observation's $1 \times q$ vector of values of the q ZI explanatory variables. Then φ_i will be a function of $\mathbf{w}_i'\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the $q \times 1$ vector of coefficients to be estimated.

RESTRICT Statement

RESTRICT *restriction1 [, restriction2 ...]* **;**

The RESTRICT statement is used to impose linear restrictions on the parameter estimates. You can specify any number of RESTRICT statements.

Each *restriction* is written as an expression, followed by an equality operator (=) or an inequality operator (<, >, <=, >=), followed by a second expression:

expression operator expression

The *operator* can be =, <, >, <=, or >=.

Restriction expressions can be composed of variable names, times ($*$) and plus ($+$) operators, and constants. Variables named in restriction expressions must be among the variables estimated by the model. The restriction expressions must be a linear function of the variables.

Lagrange multipliers are reported for all the active linear constraints. In the displayed output, the Lagrange multiplier estimates are identified with the names Restrict1, Restrict2, and so forth. The probability of the Lagrange multipliers are computed using a beta distribution (LaMotte 1994).

Details of Count Data Analysis

Poisson Regression

The most widely used model for count data analysis is Poisson regression. This assumes that y_i, given the vector of covariates $\mathbf{x}_i$, is independently Poisson distributed with

$$P(Y_i = y_i|\mathbf{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

and the mean parameter, that is, the mean number of events per period, is given by

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector. Taking the exponential of $\mathbf{x}_i'\boldsymbol{\beta}$ ensures that the mean parameter μ_i is nonnegative. It can be shown that the conditional mean is given by

$$E(y_i|\mathbf{x}_i) = \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

The name log-linear model is also used for the Poisson regression model since the logarithm of the conditional mean is linear in the parameters.

$$\ln[E(y_i|\mathbf{x}_i)] = \ln(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

Note that the conditional variance of the count random variable is equal to the conditional mean in the Poisson regression model.

$$V(y_i|\mathbf{x}_i) = E(y_i|\mathbf{x}_i) = \mu_i$$

The equality of the conditional mean and variance of y_i is known as *equidispersion*.

The marginal effect of a regressor is given by

$$\frac{\partial E(y_i|\mathbf{x}_i)}{\partial x_{ji}} = \exp(\mathbf{x}_i'\boldsymbol{\beta})\boldsymbol{\beta} = E(y_i|\mathbf{x}_i)\beta_j$$

Thus, a one unit change in the jth regressor leads to a *proportional* change in the conditional mean $E(y_i|\mathbf{x}_i)$ of β_j.

The standard estimator for the Poisson model is the maximum likelihood estimator (MLE). Since the observations are independent, the log-likelihood function is written

$$\mathcal{L} = \sum_{i=1}^{N}(-\mu_i + y_i \ln \mu_i - \ln y_i!) = \sum_{i=1}^{N}(-e^{\mathbf{x}_i'\boldsymbol{\beta}} + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln y_i!)$$

The gradient and the Hessian are

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N}(y_i - \mu_i)\mathbf{x}_i = \sum_{i=1}^{N}(y_i - e^{\mathbf{x}_i'\boldsymbol{\beta}})\mathbf{x}_i$$

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\sum_{i=1}^{N}\mu_i\mathbf{x}_i\mathbf{x}_i{}' = -\sum_{i=1}^{N}e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i\mathbf{x}_i'$$

The Poisson model has been criticized for its restrictive property that the conditional variance equals the conditional mean. Real-life data are often characterized by *overdispersion*, that is, the variance exceeds the mean. Allowing for overdispersion can improve model predictions since the Poisson restriction of equal mean and variance results in the underprediction of zeros when overdispersion exists. The most commonly used model that accounts for overdispersion is the negative binomial model.

Negative Binomial Regression

The Poisson regression model can be generalized by introducing an unobserved heterogeneity term for observation i. Thus, the individuals are assumed to differ randomly in a manner that is not fully accounted by the observed covariates. This is formulated as

$$E(y_i|\mathbf{x}_i, \tau_i) = \mu_i\tau_i = e^{\mathbf{x}_i'\boldsymbol{\beta}+\epsilon_i}$$

where the unobserved heterogeneity term $\tau_i = e^{\epsilon_i}$ is independent of the vector of regressors $\mathbf{x}_i$. Then the distribution of y_i conditional on $\mathbf{x}_i$ and τ_i is Poisson with conditional mean and conditional variance $\mu_i\tau_i$:

$$f(y_i|\mathbf{x}_i, \tau_i) = \frac{\exp(-\mu_i\tau_i)(\mu_i\tau_i)^{y_i}}{y_i!}$$

Let $g(\tau_i)$ be the probability density function of τ_i. Then, the distribution $f(y_i|\mathbf{x}_i)$ (no longer conditional on τ_i) is obtained by integrating $f(y_i|\mathbf{x}_i, \tau_i)$ with respect to τ_i:

$$f(y_i|\mathbf{x}_i) = \int_0^\infty f(y_i|\mathbf{x}_i, \tau_i)g(\tau_i)d\tau_i$$

An analytical solution to this integral exists when τ_i is assumed to follow a gamma distribution. This solution is the negative binomial distribution. When the model contains a constant term, it is necessary to assume that $E(e^{\epsilon_i}) = E(\tau_i) = 1$, in order to identify the mean of the distribution. Thus, it is assumed that τ_i follows a gamma(θ, θ) distribution with $E(\tau_i) = 1$ and $V(\tau_i) = 1/\theta$:

$$g(\tau_i) = \frac{\theta^\theta}{\Gamma(\theta)}\tau_i^{\theta-1}\exp(-\theta\tau_i)$$

where $\Gamma(x) = \int_0^\infty z^{x-1}\exp(-z)dz$ is the gamma function and θ is a positive parameter. Then, the density of y_i given $\mathbf{x}_i$ is derived as

$$\begin{aligned}
f(y_i|\mathbf{x}_i) &= \int_0^\infty f(y_i|\mathbf{x}_i, \tau_i)g(\tau_i)d\tau_i \\
&= \frac{\theta^\theta\mu_i^{y_i}}{y_i!\Gamma(\theta)}\int_0^\infty e^{-(\mu_i+\theta)\tau_i}\tau_i^{\theta+y_i-1}d\tau_i \\
&= \frac{\theta^\theta\mu_i^{y_i}\Gamma(y_i+\theta)}{y_i!\Gamma(\theta)(\theta+\mu_i)^{\theta+y_i}} \\
&= \frac{\Gamma(y_i+\theta)}{y_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_i}\right)^\theta\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i}
\end{aligned}$$

Making the substitution $\alpha = \frac{1}{\theta}$ $(\alpha > 0)$, the negative binomial distribution can then be rewritten as

$$f(y_i|\mathbf{x}_i) = \frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, \quad y_i = 0, 1, 2, \ldots$$

Thus, the negative binomial distribution is derived as a gamma mixture of Poisson random variables. It has conditional mean

$$E(y_i|\mathbf{x}_i) = \mu_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$$

and conditional variance

$$V(y_i|\mathbf{x}_i) = \mu_i[1 + \frac{1}{\theta}\mu_i] = \mu_i[1 + \alpha\mu_i] > E(y_i|\mathbf{x}_i)$$

The conditional variance of the negative binomial distribution exceeds the conditional mean. Overdispersion results from neglected unobserved heterogeneity. The negative binomial model with variance function $V(y_i|\mathbf{x}_i) = \mu_i + \alpha\mu_i^2$ that is quadratic in the mean is referred to as the NEGBIN2 model (Cameron and Trivedi 1986). (To estimate this model with the COUNTREG procedure, you must indicate the option TYPE=NEGBIN on the PROC COUNTREG statement.) The Poisson distribution is a special case of the negative binomial distribution where $\alpha = 0$. A test of the Poisson distribution can be carried out by testing the hypothesis that $\alpha = \frac{1}{\theta_i} = 0$ using the Wald or likelihood ratio test.

The log-likelihood function of the negative binomial regression model (NEGBIN2) is given by

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N}\left\{\sum_{j=0}^{y_i-1}\ln(j+\alpha^{-1}) - \ln(y_i!)\right. \\ &\qquad \left. -(y_i+\alpha^{-1})\ln(1+\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})) + y_i\ln(\alpha) + y_i\mathbf{x}_i'\boldsymbol{\beta}\right\}\end{aligned}$$

where use of the following fact is made:

$$\Gamma(y+a)/\Gamma(a) = \prod_{j=0}^{y-1}(j+a)$$

if y is an integer.

The gradient is

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{N}\frac{y_i-\mu_i}{1+\alpha\mu_i}\mathbf{x}_i$$

and

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

Cameron and Trivedi (1986) consider a general class of negative binomial models with mean μ_i and variance function $\mu_i + \alpha\mu_i^p$. The NEGBIN2 model, with $p = 2$, is the standard formulation of the negative binomial model. (To estimate this model with the COUNTREG procedure, you must indicate the option TYPE=NEGBIN on the PROC COUNTREG statement.) Models with other values of p have the same density $f(y_i|\mathbf{x}_i)$ except that α^{-1} is replaced everywhere by $\alpha^{-1}\mu^{2-p}$. The negative binomial model, NEGBIN1, which sets $p = 1$, has variance function $V(y_i|\mathbf{x}_i) = \mu_i + \alpha\mu_i$, which is linear in the mean. (To estimate this model with the COUNTREG procedure, you must indicate the option TYPE=NEGBIN1 on the PROC COUNTREG statement.)

The log-likelihood function of the NEGBIN1 regression model is given by

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N}\left\{\sum_{j=0}^{y_i-1}\ln\left(j+\alpha^{-1}\exp(\mathbf{x}_i'\boldsymbol{\beta})\right)\right. \\ &\quad \left. -\ln(y_i!)-\left(y_i+\alpha^{-1}\exp(\mathbf{x}_i'\boldsymbol{\beta})\right)\ln(1+\alpha)+y_i\ln(\alpha)\right\}\end{aligned}$$

The gradient is

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{N}\left\{\left(\sum_{j=0}^{y_i-1}\frac{\mu_i}{(j\alpha+\mu_i)}\right)\mathbf{x}_i - \alpha^{-1}\ln(1+\alpha)\mu_i\mathbf{x}_i\right\}$$

and

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

Zero-Inflated Count Regression Models

The main motivation for zero-inflated count models is that real-life data frequently display overdispersion and excess zeros. Zero-inflated count models provide a way of modeling the excess zeros as well as allowing for overdispersion. In particular, for each time period, there are two possible data generation processes. The result of a Bernoulli trial is used to determine which of the two processes is used. For time period i, Process 1 is chosen with probability φ_i and Process 2 with probability $1 - \varphi_i$. Process 1 generates only zero counts. Process 2 generates counts from either a Poisson or a negative binomial model. The proportion probability φ_i is defined shortly. In general:

$$y_i \sim \begin{cases} 0 & \text{with probability} \quad \varphi_i \\ g(y_i) & \text{with probability} \quad 1-\varphi_i \end{cases}$$

Therefore, the probability of $\{Y_i = y_i\}$ can be described as

$$\begin{aligned} P(y_i = 0|\mathbf{x}_i) &= \varphi_i + (1-\varphi_i)g(0) \\ P(y_i|\mathbf{x}_i) &= (1-\varphi_i)g(y_i), \quad y_i > 0 \end{aligned}$$

where $g(y_i)$ follows either the Poisson or the negative binomial distribution.

When the proportion probability φ_i depends on the characteristics of the time period i, φ_i is written as a function of $\mathbf{w}_i'\boldsymbol{\gamma}$, where $\mathbf{w}_i'$ is the $1 \times q$ vector of zero-inflated covariates and $\boldsymbol{\gamma}$ is the $q \times 1$ vector of zero-inflated coefficients to be estimated. The function F relating the product $\mathbf{w}_i'\boldsymbol{\gamma}$ (which is a scalar) to the proportion probability φ_i is called the zero-inflated link function.

$$\varphi_i = F_i = F(\mathbf{w}_i'\boldsymbol{\gamma})$$

In the COUNTREG procedure, the zero-inflated covariates are indicated on the MODEL statement, within the ZI() option list, by using the VAR= option. Furthermore, the zero-inflated link function F can be specified as either the logistic function:

$$F(\mathbf{w}_i'\boldsymbol{\gamma}) = \Lambda(\mathbf{w}_i'\boldsymbol{\gamma}) = \frac{\exp(\mathbf{w}_i'\boldsymbol{\gamma})}{1+\exp(\mathbf{w}_i'\boldsymbol{\gamma})}$$

or the standard normal distribution function (also called the probit function):

$$F(\mathbf{w}_i'\boldsymbol{\gamma}) = \Phi(\mathbf{w}_i'\boldsymbol{\gamma}) = \int_0^{\mathbf{w}_i'\boldsymbol{\gamma}} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)du$$

In the COUNTREG procedure, the zero-inflated link function is indicated on the MODEL statement, within the ZI() option list, by using the LINK= option.

Zero-Inflated Poisson Regression Model

In the zero-inflated Poisson (ZIP) regression model, the data generation process referred to as Process 2 is

$$g(y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

where $\mu_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$. Thus the ZIP model is defined as

$$\begin{aligned} P(y_i = 0|\mathbf{x}_i, \mathbf{w}_i) &= F_i + (1-F_i)\exp(-\mu_i) \\ P(y_i|\mathbf{x}_i, \mathbf{w}_i) &= (1-F_i)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \quad y_i > 0 \end{aligned}$$

The conditional expectation and conditional variance of y_i is given by

$$E(y_i|\mathbf{x}_i, \mathbf{w}_i) = \mu_i(1-F_i)$$

$$V(y_i|\mathbf{x}_i, \mathbf{w}_i) = E(y_i|\mathbf{x}_i, \mathbf{w}_i)(1 + \mu_i F_i)$$

Note that the ZIP model (as well as the ZINB model) exhibits overdispersion since $V(y_i|\mathbf{x}_i, \mathbf{w}_i) > E(y_i|\mathbf{x}_i, \mathbf{w}_i)$.

In general, the log-likelihood function of the ZIP model is

$$\mathcal{L} = \sum_{i=1}^{N} \ln \left[P(y_i|\mathbf{x}_i, \mathbf{w}_i) \right]$$

Once a specific link function (either logistic or standard normal) for the proportion probability is chosen, it is possible to write the exact expressions for the log-likelihood function and the gradient.

ZIP Model with Logistic Link Function

First, consider the ZIP model in which the proportion probability is expressed with a logistic link function, namely

$$\varphi_i = \frac{\exp(\mathbf{w}_i'\boldsymbol{\gamma})}{1 + \exp(\mathbf{w}_i'\boldsymbol{\gamma})}$$

The log-likelihood function is

$$\begin{aligned} \mathcal{L} \;=\; & \sum_{\{i:y_i=0\}} \ln \left[\exp(\mathbf{w}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta})) \right] \\ & + \sum_{\{i:y_i>0\}} \left[y_i \mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \sum_{k=2}^{y_i} \ln(k) \right] \\ & - \sum_{i=1}^{N} \ln \left[1 + \exp(\mathbf{w}_i'\boldsymbol{\gamma}) \right] \end{aligned}$$

The gradient for this model is given by

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \sum_{\{i:y_i=0\}} \left[\frac{\exp(\mathbf{w}_i'\boldsymbol{\gamma})}{\exp(\mathbf{w}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))} \right] \mathbf{w}_i - \sum_{i=1}^{N} \left[\frac{\exp(\mathbf{w}_i'\boldsymbol{\gamma})}{1 + \exp(\mathbf{w}_i'\boldsymbol{\gamma})} \right] \mathbf{w}_i$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{\{i:y_i=0\}} \left[\frac{-\exp(\mathbf{x}_i'\boldsymbol{\beta}) \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}{\exp(\mathbf{w}_i'\boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))} \right] \mathbf{x}_i + \sum_{\{i:y_i>0\}} \left[y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}) \right] \mathbf{x}_i$$

ZIP Model with Standard Normal Link Function

Next, consider the ZIP model in which the proportion probability is expressed with a standard normal link function: $\varphi_i = \Phi(\mathbf{w}_i'\boldsymbol{\gamma})$. The log-likelihood function is

$$\begin{aligned}\mathcal{L} &= \sum_{\{i:y_i=0\}} \ln\left\{\Phi(\mathbf{w}_i'\boldsymbol{\gamma}) + \left[1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right]\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\right\} \\ &+ \sum_{\{i:y_i>0\}} \left\{\ln\left[\left(1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right)\right] - \exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i\mathbf{x}_i'\boldsymbol{\beta} - \sum_{k=2}^{y_i}\ln(k)\right\}\end{aligned}$$

The gradient for this model is given by

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\gamma}} &= \sum_{\{i:y_i=0\}} \frac{\phi(\mathbf{w}_i'\boldsymbol{\gamma})\left[1-\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\right]}{\Phi(\mathbf{w}_i'\boldsymbol{\gamma}) + \left[1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right]\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}\mathbf{w}_i \\ &- \sum_{\{i:y_i>0\}} \frac{\phi(\mathbf{w}_i'\boldsymbol{\gamma})}{\left[1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right]}\mathbf{w}_i\end{aligned}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} &= \sum_{\{i:y_i=0\}} \frac{-\left[1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right]\exp(\mathbf{x}_i'\boldsymbol{\beta})\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}{\Phi(\mathbf{w}_i'\boldsymbol{\gamma}) + \left[1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right]\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))}\mathbf{x}_i \\ &+ \sum_{\{i:y_i>0\}} \left[y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})\right]\mathbf{x}_i\end{aligned}$$

Zero-Inflated Negative Binomial Regression Model

The zero-inflated negative binomial (ZINB) model is obtained by specifying a negative binomial distribution for the data generation process called Process 2:

$$g(y_i) = \frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}$$

Thus the ZINB model is defined to be

$$\begin{aligned}P(y_i=0|\mathbf{x}_i,\mathbf{w}_i) &= F_i + (1-F_i)\left(1+\alpha\mu_i\right)^{-\alpha^{-1}} \\ P(y_i|\mathbf{x}_i,\mathbf{w}_i) &= (1-F_i)\frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} \\ &\times \left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, \quad y_i>0\end{aligned}$$

In this case, the conditional expectation and conditional variance of y_i are

$$E(y_i|\mathbf{x}_i,\mathbf{w}_i) = \mu_i(1-F_i)$$

$$V(y_i|\mathbf{x}_i, \mathbf{w}_i) = E(y_i|\mathbf{x}_i, \mathbf{w}_i)\,[1 + \mu_i(F_i + \alpha)]$$

Note: The ZINB model described here, like the ZIP model, exhibits overdispersion, because the conditional variance exceeds the conditional mean.

ZINB Model with Logistic Link Function

In this model, the proportion probability φ_i is given by the logistic function, namely

$$\varphi_i = \frac{\exp(\mathbf{w}_i'\boldsymbol{\gamma})}{1 + \exp(\mathbf{w}_i'\boldsymbol{\gamma})}$$

The log-likelihood function is

$$\begin{aligned}
\mathcal{L} &= \sum_{\{i:y_i=0\}} \ln\left[\exp(\mathbf{w}_i'\boldsymbol{\gamma}) + (1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}}\right] \\
&+ \sum_{\{i:y_i>0\}} \sum_{j=0}^{y_i-1} \ln(j + \alpha^{-1}) \\
&+ \sum_{\{i:y_i>0\}} \left\{-\ln(y_i!) - (y_i + \alpha^{-1})\ln(1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})) + y_i\ln(\alpha) + y_i\mathbf{x}_i'\boldsymbol{\beta}\right\} \\
&- \sum_{i=1}^{N} \ln\left[1 + \exp(\mathbf{w}_i'\boldsymbol{\gamma})\right]
\end{aligned}$$

The gradient for this model is given by

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\gamma}} &= \sum_{\{i:y_i=0\}} \left[\frac{\exp(\mathbf{w}_i'\boldsymbol{\gamma})}{\exp(\mathbf{w}_i'\boldsymbol{\gamma}) + (1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}}}\right]\mathbf{w}_i \\
&- \sum_{i=1}^{N} \left[\frac{\exp(\mathbf{w}_i'\boldsymbol{\gamma})}{1 + \exp(\mathbf{w}_i'\boldsymbol{\gamma})}\right]\mathbf{w}_i
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} &= \sum_{\{i:y_i=0\}} \left[\frac{-\exp(\mathbf{x}_i'\boldsymbol{\beta})(1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}-1}}{\exp(\mathbf{w}_i'\boldsymbol{\gamma}) + (1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}}}\right]\mathbf{x}_i \\
&+ \sum_{\{i:y_i>0\}} \left[\frac{y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]\mathbf{x}_i
\end{aligned}$$

$$\frac{\partial\mathcal{L}}{\partial\alpha} = \sum_{\{i:y_i=0\}} \frac{\alpha^{-2}\left[(1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))\ln(1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})) - \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]}{\exp(\mathbf{w}_i'\boldsymbol{\gamma})(1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))^{(1+\alpha)/\alpha} + (1 + \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))}$$

$$+ \sum_{\{i:y_i>0\}} \left\{ -\alpha^{-2} \sum_{j=0}^{y_i-1} \frac{1}{(j+\alpha^{-1})} + \alpha^{-2} \ln(1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta})) + \frac{y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})}{\alpha(1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta}))} \right\}$$

ZINB Model with Standard Normal Link Function

For this model, the proportion probability is specified with the probit function: $\varphi_i = \Phi(\mathbf{w}_i'\boldsymbol{\gamma})$. The log-likelihood function is

$$\begin{aligned}
\mathcal{L} &= \sum_{\{i:y_i=0\}} \ln \left\{ \Phi(\mathbf{w}_i'\boldsymbol{\gamma}) + \left[1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right] (1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}} \right\} \\
&+ \sum_{\{i:y_i>0\}} \ln \left[1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right] \\
&+ \sum_{\{i:y_i>0\}} \sum_{j=0}^{y_i-1} \left\{ \ln(j+\alpha^{-1}) \right\} \\
&- \sum_{\{i:y_i>0\}} \ln(y_i!) \\
&- \sum_{\{i:y_i>0\}} (y_i + \alpha^{-1}) \ln(1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta})) \\
&+ \sum_{\{i:y_i>0\}} y_i \ln(\alpha) \\
&+ \sum_{\{i:y_i>0\}} y_i \mathbf{x}_i'\boldsymbol{\beta}
\end{aligned}$$

The gradient for this model is given by

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \sum_{\{i:y_i=0\}} \left[\frac{\phi(\mathbf{w}_i'\boldsymbol{\gamma}) \left[1 - (1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}}\right]}{\Phi(\mathbf{w}_i'\boldsymbol{\gamma}) + [1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma})] (1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}}} \right] \mathbf{w}_i$$

$$- \sum_{\{i:y_i>0\}} \left[\frac{\phi(\mathbf{w}_i'\boldsymbol{\gamma})}{1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma})} \right] \mathbf{w}_i$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{\{i:y_i=0\}} \frac{-\left[1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma})\right] \exp(\mathbf{x}_i'\boldsymbol{\beta}) (1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-(1+\alpha)/\alpha}}{\Phi(\mathbf{w}_i'\boldsymbol{\gamma}) + [1 - \Phi(\mathbf{w}_i'\boldsymbol{\gamma})] (1+\alpha \exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-\alpha^{-1}}} \mathbf{x}_i$$

$$+ \sum_{\{i:y_i>0\}} \left[\frac{y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \alpha \exp(\mathbf{x}_i'\boldsymbol{\beta})} \right] \mathbf{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{\{i:y_i=0\}} \frac{[1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})]\,\alpha^{-2}\left[(1+\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))\ln(1+\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})) - \alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]}{\Phi(\mathbf{w}_i'\boldsymbol{\gamma})(1+\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))^{(1+\alpha)/\alpha} + [1-\Phi(\mathbf{w}_i'\boldsymbol{\gamma})]\,(1+\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))}$$

$$+ \sum_{\{i:y_i>0\}} \left\{ -\alpha^{-2} \sum_{j=0}^{y_i-1} \frac{1}{(j+\alpha^{-1})} + \alpha^{-2}\ln(1+\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})) + \frac{y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})}{\alpha(1+\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta}))} \right\}$$

Examples

Example 1: ZIP and ZINB Models for Data Exhibiting Extra Zeros

In the study by Long (1997) of the number of published articles by scientists (see the "Getting Started" section), the observed proportion of scientists publishing no articles is 0.3005. PROC COUNTEG is used to fit poisson and negative binomial models to these data. For each model, the predicted proportion of zero articles can be calculated as the average predicted probability of zero articles across all scientists. Under the poisson model, the predicted proportion of zero articles is 0.2092 which considerably underestimates the observed proportion. The negative binomial more closely estimates the proportion of zeros (0.3036). Also, the test of the dispersion parameter, _Alpha, in the negative binomial model indicates significant overdispersion in the poisson model ($p < .0001$). As a result, the negative binomial model is preferred to the poisson model.

Another way to account for the large number of zeros in these data is to fit a zero-inflated poisson (ZIP) or a zero-inflated negative binomial (ZINB) model. The following statements fit the ZIP model. The TYPE=ZIP option requests the ZIP model. The ZI option in the MODEL statement allows you to specify how the proportion probability, psi, is modeled. By default, a logistic model is used for psi. This can be changed using the LINK= option within the ZI option. The VAR= option within the ZI option specifies the linear predictor portion of the model for psi. In this ZIP model, all variables used to model the article counts are also used to model psi.

```
proc countreg data=one type=zip;
   model art =  fem mar kid5 phd ment /
         zi(var=fem mar kid5 phd ment);
run;
```

The parameters of the ZIP model are displayed below. The first set of parameters gives the estimates of beta in the model for the poisson mean. Parameters with the prefix "Inf_" are the estimates of gamma in the logistic model for psi.

Output 1: Parameter estimates of the ZIP model

Parameter Estimates

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	0.640838	0.121306	5.28	<.0001
FEM	−0.209145	0.063405	−3.30	0.0010
MAR	0.103751	0.071111	1.46	0.1446
KID5	−0.143320	0.047429	−3.02	0.0025
PHD	−0.006166	0.031008	−0.20	0.8424
MENT	0.018098	0.002295	7.89	<.0001
Inf_Intercept	−0.577060	0.509383	−1.13	0.2573
Inf_FEM	0.109747	0.280082	0.39	0.6952
Inf_MAR	−0.354013	0.317611	−1.11	0.2650
Inf_KID5	0.217101	0.196481	1.10	0.2692
Inf_PHD	0.001272	0.145262	0.01	0.9930
Inf_MENT	−0.134114	0.045244	−2.96	0.0030

The proportion of zeros predicted by the ZIP model is 0.2986 – much closer to the observed proportion than the poisson model. But Output 3 shows that both models deviate from the observed proportions at one, two, and three articles.

The ZINB model is specified by the TYPE=ZINB option. All variables are again used to model both the number of articles and psi. The METHOD=QN option specifies that the quasi-Newton method be used to fit the model rather than the default Newton-Raphson method.

```
proc countreg data=one type=zinb method=qn;
   model art =  fem mar kid5 phd ment /
         zi(var=fem mar kid5 phd ment);
run;
```

The estimated parameters of the ZINB model are shown below. The test for overdispersion again indicates a preference for the negative binomial version of the zero-inflated model (p<.0001). The ZINB model also does a good job of estimating the proportion of zeros (0.3119) and it follows the observed proportions well, though possibly not as well as the negative binomial model.

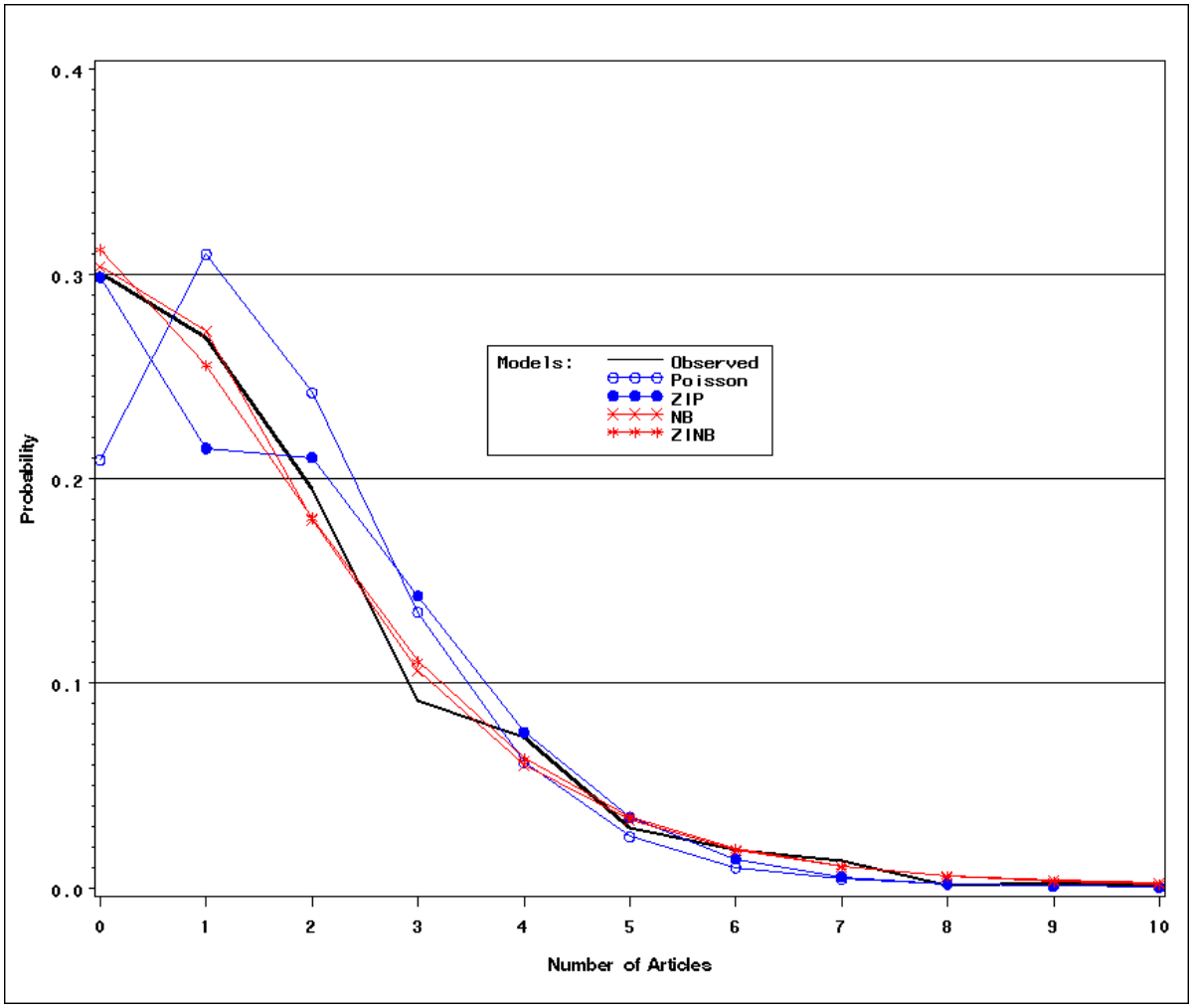
Output 2: Parameter estimates of the ZINB model

Parameter Estimates

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	0.416746	0.143596	2.90	0.0037
FEM	−0.195506	0.075592	−2.59	0.0097
MAR	0.097582	0.084452	1.16	0.2479
KID5	−0.151732	0.054206	−2.80	0.0051
PHD	−0.000700	0.036270	−0.02	0.9846
MENT	0.024786	0.003493	7.10	<.0001
Inf_Intercept	−0.191716	1.322807	−0.14	0.8848
Inf_FEM	0.635957	0.848913	0.75	0.4538
Inf_MAR	−1.499458	0.938655	−1.60	0.1102
Inf_KID5	0.628428	0.442780	1.42	0.1558
Inf_PHD	−0.037710	0.308004	−0.12	0.9026
Inf_MENT	−0.882298	0.316225	−2.79	0.0053
_Alpha	0.376681	0.051029	7.38	<.0001

For each of the four fitted models, the graph in Output 3 shows the average predicted probability for each article count across all scientists. The poisson model clearly underestimates the proportion of zero articles published while the other three model are quite accurate at zero. All of the models do well at the larger number of articles.

Output 3: Average predicted probabilities of article counts for poisson, negative binomial, ZIP, and ZINB models



References

Abramowitz, M. and Stegun, A. (1970), *Handbook of Mathematical Functions*, New York: Dover Press.

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge: Harvard University Press.

Cameron, A. C. and Trivedi, P. K. (1986), "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Some Tests," *Journal of Applied Econometrics*, 1, 29–53.

Cameron, A. C. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press.

Godfrey, L. G. (1988), *Misspecification Tests in Econometrics*, Cambridge: Cambridge University Press.

Greene, W. H. (1994), "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models," *Working Paper No. 94-10*, New York: Stern School of Business, Department of Economics, New York University.

Greene, W. H. (2000), *Econometric Analysis*, Upper Saddle River, N.J.: Prentice Hall.

Hausman, J. A., Hall, B. H., and Griliches, Z. (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52, 909–938.

King, G. (1989a), "A Seemingly Unrelated Poisson Regression Model," *Sociological Methods & Research*, 17, 235–255.

King, G. (1989b), *Unifying Political Methodology: The Likelihood Theory and Statistical Inference*, Cambridge: Cambridge University Press.

Lambert, D. (1992), "Zero-inflated Poisson Regression with an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

LaMotte, L. R. (1994), "A Note on the Role of Independence in t Statistics Constructed from Linear Statistics in Regression Models," *The American Statistician*, 48, 238–240.

Long, J. S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks: Sage Publications, Inc.

Winkelmann, R. (2000), *Econometric Analysis of Count Data*, Berlin: Springer-Verlag.