



# PROC COUNTREG

Charlie Hallahan

October 14, 2009

# 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**  $E(y_i | x_i)$  of the dependent variable,  $y_i$ , is assumed to be a function of a vector of covariates,  $x_i$ .

The **Poisson (log-linear) regression model** is the most basic model that explicitly takes into account the nonnegative integer-valued aspect of the outcome.

\* This handout uses information and examples from the **PROC COUNTREG** documentation in the **SAS/ETS 9.2** manual. 2

# Overview

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 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.

# Overview

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 model.

**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 enable the conditional variance of the outcome to differ from the conditional mean.

Under **zero-inflated models**, additional zeros occur with probability  $\varphi_i$ , which is determined by a separate model,  $\varphi_i = F(\mathbf{z}'_i\boldsymbol{\gamma})$  where  $F$  is the normal or logistic distribution function resulting in a **probit** or **logistic** model, and  $\mathbf{z}_i$  is a set of covariates.

# Overview

**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

# Overview

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 negative predicted counts and to parameter estimates with undesirable properties in terms of statistical efficiency, consistency, and unbiasedness unless the mean of the counts is high, in which case the Gaussian approximation and linear regression may be satisfactory.

# Getting Started: COUNTREG Procedure

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 ;  
    model y = x / dist=poisson ;  
run;
```

The response variable  $y$  is numeric and has nonnegative integer values.

To allow for variance greater than the mean, specify the `dist=negbin` option to fit the negative binomial model instead of the Poisson.

# Getting Started: COUNTREG Procedure

The following example illustrates the use of **PROC COUNTREG**.

The data are taken from **Long (1997)**.

This **study examines how factors** such as

- gender (fem),
- marital status (mar),
- number of young children (kid5),
- prestige of the graduate program (phd), and
- number of articles published by a scientist's mentor (ment)

**affect the number of articles (art) published by the scientist.**

# Getting Started: COUNTREG Procedure

The first 10 observations are shown in Figure 10.1.

**Figure 10.1** Article Count Data

1st 10 observations of Article Count Data						
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

# Getting Started: COUNTREG Procedure

The following **SAS** statements estimate the **Poisson regression model**:

```
/*-- Poisson Regression --*/  
proc countreg data=long97data;  
    model art = fem mar kid5 phd ment / dist=poisson method=quanew;  
run;
```

The **Model Fit Summary**, shown in Figure 10.2, lists several details about the model.

By default, the **COUNTREG** procedure uses the **Newton-Raphson** optimization technique.

The **maximum log-likelihood value** is shown, as well as two **information measures**, **Akaike's information criterion (AIC)** and **Schwarz's Bayesian information criterion (SBC)**, which can be used to **compare competing Poisson models**.

**Smaller values** of these criteria indicate better models.

# Getting Started: COUNTREG Procedure

**Figure 10.2** Estimation Summary Table for a Poisson Regression

## The COUNTREG Procedure

### Model Fit Summary

Dependent Variable	art
Number of Observations	915
Data Set	SIGSTAT.LONG97DATA
Model	Poisson
Log Likelihood	-1651
Maximum Absolute Gradient	0.00181
Number of Iterations	13
Optimization Method	Quasi-Newton
AIC	3314
SBC	3343

# Getting Started: COUNTREG Procedure

The parameter estimates of the model and their standard errors are shown in Figure 10.3.

All covariates are significant predictors of the number of articles, except for the prestige of the program (phd), which has a p-value of 0.6271.

**Figure 10.3** Parameter Estimates of Poisson Regression

Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t	
Intercept	1	0.304617	0.102982	2.96	0.0031	
fem	1	-0.224594	0.054614	-4.11	<.0001	
mar	1	0.155243	0.061375	2.53	0.0114	
kid5	1	-0.184883	0.040127	-4.61	<.0001	
phd	1	0.012823	0.026397	0.49	0.6271	
ment	1	0.025543	0.002006	12.73	<.0001	

# Getting Started: COUNTREG Procedure

The following statements fit the **negative binomial model**.

While the Poisson model requires that the conditional mean and conditional variance be equal, the **negative binomial model allows for overdispersion**; that is, the conditional variance may exceed the conditional mean.

```
/*-- Negative Binomial Regression --*/  
proc countreg data=long97data;  
    model art = fem mar kid5 phd ment / dist=negbin(p=2) method=quanew;  
run;
```

The fit summary is shown in Figure 10.4, and parameter estimates are listed in Figure 10.5.

# Getting Started: COUNTREG Procedure

**Figure 10.4** Estimation Summary Table for a Negative Binomial Regression

The COUNTREG Procedure

Model Fit Summary

Dependent Variable	art
Number of Observations	915
Data Set	SIGSTAT.LONG97DATA
<b>Model</b>	<b>NegBin</b>
Log Likelihood	-1561
Maximum Absolute Gradient	5.72276E-7
Number of Iterations	16
Optimization Method	Quasi-Newton
AIC	3136
SBC	3170

# Getting Started: COUNTREG Procedure

Figure 10.5 Parameter Estimates of Negative Binomial Regression

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	0.256144	0.138560	1.85	0.0645
fem	1	-0.216418	0.072672	-2.98	0.0029
mar	1	0.150489	0.082106	1.83	0.0668
kid5	1	-0.176415	0.053060	-3.32	0.0009
phd	1	0.015271	0.036040	0.42	0.6718
ment	1	0.029082	0.003470	8.38	<.0001
_Alpha	1	0.441620	0.052967	8.34	<.0001

# Getting Started: COUNTREG Procedure

The parameter estimate for `_Alpha` of 0.4416 is an **estimate of the dispersion** parameter in the negative binomial distribution.

A t test for the hypothesis  $H_0: \alpha = 0$  is provided. It is highly significant, **indicating overdispersion** ( $p < 0.0001$ ).

The null hypothesis  $H_0: \alpha = 0$  can be also tested against the alternative  $\alpha > 0$  by using the **likelihood ratio test**, as described by Cameron and Trivedi (1998, pp. 45, 77.78).

The **likelihood ratio test statistic** is equal to  $-2(L_p - L_{NB}) = -2(-1651 + 1561) = 180$ , which is highly significant, providing strong evidence of overdispersion.

# Syntax: COUNTREG Procedure

The **COUNTREG** procedure is controlled by the following statements:

**PROC COUNTREG** options ;

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

**BY** variables ;

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

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

**OUTPUT** options ;

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

**ZEROMODEL** dependent variable zero-inflated regressors / options ;

There can only be one **MODEL** statement. The **ZEROMODEL** statement, if used, must appear after the **MODEL** statement.

# Syntax: COUNTREG Procedure

(Only some **selected options** are shown here)

## 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.

# Syntax: COUNTREG Procedure

## Estimation Control Options

### **COVEST=value**

The **COVEST=** option specifies the type of covariance matrix of the parameter estimates.

The quasi-maximum likelihood estimates are computed with **COVEST=QML**. The **default is COVEST=HESSIAN**.

The supported covariance types are as follows:

**OP** specifies the covariance from the outer product matrix.

**HESSIAN** specifies the covariance from the Hessian matrix.

**QML** specifies the covariance from the outer product and Hessian matrices.

# Syntax: COUNTREG Procedure

## Options to Control the Optimization Process

**PROC COUNTREG** uses the **nonlinear optimization (NLO) subsystem** to perform nonlinear optimization tasks. All the NLO options are available. For details, see Chapter 6, “Nonlinear Optimization Methods.” In addition, the following option is supported on the PROC COUNTREG statement:

### **METHOD=value**

specifies the iterative minimization method to use. The **default** is **METHOD=NRA**.

- QN** specifies the quasi-Newton method.
- NRA** specifies the Newton-Raphson method.
- TR** specifies the trust region method.

# Syntax: COUNTREG Procedure

## 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 parameter names, constants, and inequality operators:

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

# Syntax: COUNTREG Procedure

Each item is a constant, a parameter name, or a list of parameter names.

Each operator is '<', '>', '<=', or '>='.

Parameter names are as shown in the **ESTIMATE** column of the “Parameter Estimates” table.

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.

# Syntax: COUNTREG Procedure

The following BOUNDS statement constrains the estimates of the parameter for  $z$  to be negative, the parameters for  $x_1$  through  $x_{10}$  to be between zero and one, and the parameter for  $x_1$  in the zeroinflation model to be less than one.

The following example illustrates the use of parameter lists to specify boundary constraints:

```
bounds z < 0,  
       0 < x1-x10 < 1,  
       Inf_x1 < 1;
```

# Syntax: COUNTREG Procedure

## MODEL Statement

**MODEL** dependent = regressors / options ;

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

The dependent count variable should take on only nonnegative integer values in the input data set. **PROC COUNTREG** rounds any positive noninteger count values to the nearest integer.

**PROC COUNTREG** discards any observations with a negative count.

Only one **MODEL** statement can be specified.

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

# Syntax: COUNTREG Procedure

**DIST=value, D=value**

specifies a type of model to be analyzed. If you specify this option in both the **MODEL** statement and the **PROC COUNTREG** statement, then only the value in the **MODEL** statement is used. The supported model types as follows:

POISSON   P	Poisson regression model
NEGBIN(P=1)	negative binomial regression model with a linear variance function
NEGBIN(P=2)   NEGBIN	negative binomial regression model with a quadratic variance function
ZIPOISSON   ZIP	zero-inflated Poisson regression
ZINEGBIN   ZINB	zero-inflated negative binomial regression

# Syntax: COUNTREG Procedure

## **OFFSET=**variable

specifies a variable in the input data set to be used as an offset variable. The offset variable appears as a covariate in the model with its parameter restricted to 1. The offset variable cannot be the response variable, the zero-inflation offset variable (if any), or one of the explanatory variables.

The **Model Fit Summary** gives the name of the data set variable used as the offset variable; it is labeled as “**Offset.**”

# Syntax: COUNTREG Procedure

## ZEROMODEL Statement

**ZEROMODEL** dependent variable zero-inflated regressors / options ;

The **ZEROMODEL** statement is required if either **ZIP** or **ZINB** is specified in the **DIST=** option in the **MODEL** statement.

If **ZIP** or **ZINB** is specified, then the **ZEROMODEL** statement must follow immediately after the **MODEL** statement.

The dependent variable in the **ZEROMODEL** statement must be the same as the dependent variable in the **MODEL** statement.

The zero-inflated (**ZI**) regressors appear in the equation that determines the probability ( $\phi_i$ ) of a zero count.

Each of these  $q$  variables has a parameter to be estimated in the regression.

# Syntax: COUNTREG Procedure

For example, let  $\mathbf{z}_i$  be the  $i$ th observation's  $1 \times (q + 1)$  vector of values of the  $q$  **ZI** explanatory variables ( $\mathbf{w}_0$  is set to 1 for the intercept term).

Then  $\phi_i$  will be a function of  $\mathbf{z}_i\boldsymbol{\gamma}$ , where  $\boldsymbol{\gamma}$  is the  $(q + 1) \times 1$  vector of parameters to be estimated.

(The zero-inflated intercept is  $\boldsymbol{\gamma}_0$ ; the coefficients for the  $q$  zero-inflated covariates are  $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q$ .)

If this option is omitted, then only the intercept term  $\boldsymbol{\gamma}_0$  is estimated.

The “**Parameter Estimates**” table in the displayed output gives the estimates for the **ZI** intercept and **ZI** explanatory variables; they are labeled with the prefix “**Inf\_**”.

For example, the **ZI** intercept is labeled “**Inf\_intercept**”.

If you specify “**Age**” (a variable in your data set) as a **ZI** explanatory variable, then the “Parameter Estimates” table labels the corresponding parameter estimate “**Inf\_Age**”.

# Syntax: COUNTREG Procedure

The following options can be specified in the **ZEROMODEL** statement following a slash (/):

**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.

If this option is omitted, then the **default ZI** link function is **logistic**.

# Syntax: COUNTREG Procedure

**OFFSET=variable**

specifies a variable in the input data set to be used as a zero-inflated (**ZI**) offset variable.

The **ZI offset** variable is included as a term, with coefficient restricted to 1, in the equation that determines the probability ( $\phi_i$ ) of a zero count.

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

The name of the data set variable used as the **ZI offset** variable is displayed in the “**Model Fit Summary**” output, where it is labeled as “**Inf\_offset**”.

# Details: COUNTREG Procedure

## Missing Values

Any observation in the input data set with a missing value for one or more of the regressors is ignored by **PROC COUNTREG** and not used in the model fit.

**PROC COUNTREG** rounds any positive noninteger count values to the nearest integer.

**PROC COUNTREG** ignores any observations with a negative count.

If there are observations in the input data set with missing response values but with nonmissing regressors, **PROC COUNTREG** can compute several statistics and store them in an output data set by using the **OUTPUT** statement.

For example, you can request that the output data set contain the estimates of  $x_i\beta$ , the expected value of the response variable, and the probability of the response variable taking on the current value.

# Details: COUNTREG Procedure

Furthermore, if a zero-inflated model was fit, you can request that the output data set contain the estimates of  $\mathbf{z}_i\boldsymbol{\gamma}$ , and the probability that the response is zero as a result of the zero-generating process.

Note that the presence of such observations (with missing response values) does not affect the model fit.

# Details: COUNTREG Procedure

## 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

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

and the mean parameter  $\mu_i$ , that is, the mean number of events per period.

$\mu_i$  is given by  $\mu_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$  where  $\boldsymbol{\beta}$  is a  $k \times 1$  parameter vector. (The intercept is  $\beta_0$ ; the coefficients for the  $k$  regressors are  $\beta_1, \dots, \beta_k$ )

Taking the exponential of  $\mathbf{x}_i' \boldsymbol{\beta}$  ensures that the mean parameter  $\mu_i$  is nonnegative.

# Details: COUNTREG Procedure

## Poisson Regression

It can be shown that the conditional mean is given by  $E(y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ .

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

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

# Details: COUNTREG Procedure

## 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 for by the observed covariates.

This is formulated as  $E(y_i | \mathbf{x}_i, \tau_i) = \mu_i \tau_i = e^{\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i}$ .

where the unobserved heterogeneity term  $\tau_i = e^{\varepsilon_i}$  is independent of the vector of regressors  $\mathbf{x}_i$ .

Then the distribution of  $y_i$  conditional on  $\mathbf{x}_i$  and  $\tau_i$  is Poisson with conditional mean and

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

# Details: COUNTREG Procedure

## Negative Binomial Regression

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

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

An analytical solution to this integral exists when  $\tau_i$  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  $E(e^{\varepsilon_i}) = E(\tau_i) = 1$  in order to identify the mean of the distribution.

Thus, it is assumed that  $\tau_i$  follows a gamma( $\theta, \theta$ ) distribution with  $E(\tau_i) = 1$  and

$$V(\tau_i) = 1/\theta; \quad g(\tau_i) = \frac{\theta^\theta}{\Gamma(\theta)} \tau_i^{\theta-1} \exp(-\theta\tau_i).$$

# Details: COUNTREG Procedure

## Zero-Inflated Count Regression Overview

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 observation, 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 observation  $i$ , Process 1 is chosen with probability  $\varphi_i$  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. In general,

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

# Details: COUNTREG Procedure

## Zero-Inflated Count Regression Overview

Therefore, the probability of  $(Y_i = y_i)$  can be described as

$$P(y_i = 0 | \mathbf{x}_i) = \varphi_i + (1 - \varphi_i)g(0)$$

$$P(y_i | \mathbf{x}_i) = (1 - \varphi_i)g(y_i) \quad y_i > 0$$

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

When the probability  $\varphi_i$  depends on the characteristics of observation  $i$ ,  $\varphi_i$  is written as a function of  $\mathbf{z}'_i\boldsymbol{\gamma}$ , where  $\mathbf{z}'_i$  is the  $1 \times (q + 1)$  vector of zero-inflated covariates and  $\boldsymbol{\gamma}$  is the  $(q + 1) \times 1$  vector of zero-inflated coefficients to be estimated.

(The zero-inflated intercept is  $\gamma_0$ ; the coefficients for the  $q$  zero-inflated covariates are  $\gamma_1, \dots, \gamma_q$ . The function  $F$  relating the product  $\mathbf{z}'_i\boldsymbol{\gamma}$  (which is a scalar) to the

probability  $\varphi_i$  is called the **zero - inflated link function**,  $\varphi_i = F_i = F(\mathbf{z}'_i\boldsymbol{\gamma})$ .

# Details: COUNTREG Procedure

## Zero-Inflated Count Regression Overview

In the **COUNTREG** procedure, the zero-inflated covariates are indicated in the **ZEROMODEL** statement.

Furthermore, the zero-inflated link function  $F$  can be specified as either the

**logistic function**, 
$$F(\mathbf{z}'_i\boldsymbol{\gamma}) = \Lambda(\mathbf{z}'_i\boldsymbol{\gamma}) = \frac{\exp(\mathbf{z}'_i\boldsymbol{\gamma})}{1 + \exp(\mathbf{z}'_i\boldsymbol{\gamma})}$$

or the **standard normal cumulative distribution function** (also called the

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

The zero-inflated link function is indicated in the **ZEROMODEL** statement, using the **LINK=** option. The *default* **ZI** link function is the *logistic function*.

# Examples: COUNTREG Procedure

## Example 10.1: Basic Models

### Data Description and Objective

The data set **docvisit** contains information for approximately 5,000 Australian individuals about the number and possible determinants of doctor visits that were made during a two-week interval.

This data set contains a subset of variables taken from the **Racd3** data set used by Cameron and Trivedi (1998).

The variable **doctorco** represents doctor visits. Additional variables in the data set that you want to evaluate as determinants of doctor visits include

**sex** (coded 0=male, 1=female),

**age** (age in years divided by 100, with more than 70 coded as 72),

**illness** (number of illnesses during the two-week interval, with five or more coded as five),

**income** (annual income in Australian dollars divided by 1,000), and

**hscore** (a general health questionnaire score, where a high score indicates bad health).

# Examples: COUNTREG Procedure

**Summary statistics** for these variables are computed in the following statements and presented below.

In the rest of this example some possible applications of the **COUNTREG** procedure in this context are presented.

```
proc means data=docvisit;  
    var doctorco sex age illness income hscore;  
run;
```

# Examples: COUNTREG Procedure

## The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
doctorco	5190	0.3017341	0.7981338	0	9.0000000
sex	5190	0.5206166	0.4996229	0	1.0000000
age	5190	0.4063854	0.2047818	0.1900000	0.7200000
illness	5190	1.4319846	1.3841524	0	5.0000000
income	5190	0.5831599	0.3689067	0	1.5000000
hscore	5190	1.2175337	2.1242665	0	12.0000000

# Examples: COUNTREG Procedure

## Poisson Models

These statements fit a Poisson model to the data by using the covariates SEX, ILLNESS, INCOME, and HSCORE:

```
/*-- Poisson Model --*/  
proc countreg data=docvisit;  
    model doctorco=sex illness income hscore / dist=poisson printall;  
run;
```

In this example, the **DIST=** option in the **MODEL** statement specifies the **POISSON** distribution.

In addition, the **PRINTALL** option displays the correlation and covariance matrices for the parameters, log-likelihood values, and convergence information in addition to the parameter estimates.

# Examples: COUNTREG Procedure

## Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	-1.855552	0.074545	-24.89	<.0001
sex	1	0.235583	0.054362	4.33	<.0001
illness	1	0.270326	0.017080	15.83	<.0001
income	1	-0.242095	0.077829	-3.11	0.0019
hscore	1	0.096313	0.009089	10.60	<.0001

# Examples: COUNTREG Procedure

Suppose that you suspect that the **population** of individuals can be viewed as **two distinct groups**:

a **low-risk group**, comprising individuals who never go to the doctor, and a **high-risk group**, comprising individuals who do go to the doctor.

You might suspect that the data have this structure both because the sample variance of **DOCTORCO** (0.64) exceeds its sample mean (0.30), which suggests **overdispersion**, and because a large fraction of the **DOCTORCO** observations (80%) have the value zero.

Estimating a **zero-inflated model** is one way to deal with **overdispersion** that results from **excess zeros**.

# Examples: COUNTREG Procedure

Suppose also that you suspect that the covariate **AGE** has an impact on whether an individual belongs to the low-risk group.

For example, younger individuals might have illnesses of much lower severity when they do get sick and be less likely to visit a doctor, all else being equal.

The following statements estimate a zero-inflated Poisson regression with **AGE** as a covariate in the zerogeneration process:

```
/*-- Zero-Inflated Poisson Model --*/  
proc countreg data=docvisit;  
    model doctorco=sex illness income hscore / dist=zip;  
    zeromodel doctorco ~ age;  
run;
```

# Examples: COUNTREG Procedure

In this case, the **ZEROMODEL** statement following the **MODEL** statement specifies that both an intercept and the variable **AGE** be used to estimate the likelihood of zero doctor visits.

## Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-1.033387	0.096973	-10.66	<.0001
sex	1	0.122511	0.062566	1.96	0.0502
illness	1	0.237478	0.019997	11.88	<.0001
income	1	-0.143945	0.087810	-1.64	0.1012
hscore	1	0.088386	0.010043	8.80	<.0001
<b>Inf_Intercept</b>	<b>1</b>	<b>0.986557</b>	<b>0.131339</b>	<b>7.51</b>	<b>&lt;.0001</b>
<b>Inf_age</b>	<b>1</b>	<b>-2.090924</b>	<b>0.270580</b>	<b>-7.73</b>	<b>&lt;.0001</b>

# Examples: COUNTREG Procedure

The estimates of the zero-inflated intercept (**Inf\_Intercept**) and the zero-inflated regression coefficient for **AGE** (**Inf\_age**) are approximately 0.99 and  $-2.09$ , respectively.

Therefore, you can estimate the probabilities for individuals of ages 20, 50, and 70 as follows:

$$20 \text{ years: } \frac{e^{(0.99 - 2.09 * -.20)}}{1 + e^{(0.99 - 2.09 * -.20)}} = 0.64$$

$$50 \text{ years: } \frac{e^{(0.99 - 2.09 * -.50)}}{1 + e^{(0.99 - 2.09 * -.50)}} = 0.49$$

$$70 \text{ years: } \frac{e^{(0.99 - 2.09 * -.70)}}{1 + e^{(0.99 - 2.09 * -.70)}} = 0.38$$

# Examples: COUNTREG Procedure

That is, the estimated probability of belonging to the low-risk group is about 0.64 for a 20-year-old individual, 0.49 for a 50-year-old individual, and only 0.38 for a 70-year-old individual.

This supports the suspicion that **older individuals** are **more likely** to have a positive number of doctor visits.

**Alternative models** to account for the overdispersion are the negative binomial and the zero-inflated negative binomial models, which can be fit using the **DIST=NEGBIN** and **DIST=ZINB** option, respectively.

# Examples: COUNTREG Procedure

## Example 10.2: ZIP and ZINB Models for Data Exhibiting Extra Zeros

In the study by **Long (1997)** of the number of published articles by scientists, the **observed proportion** of scientists publishing no articles is **0.3005**.

The following statements use **PROC FREQ** to compute the proportion of scientists publishing each observed number of articles.

```
proc freq data=sigstat.long97data;  
  table art / out=obs;  
run;
```

# Examples: COUNTREG Procedure

art	Frequency	Percent	Frequency	Percent
0	275	30.05	275	30.05
1	246	26.89	521	56.94
2	178	19.45	699	76.39
3	84	9.18	783	85.57
4	67	7.32	850	92.90
5	27	2.95	877	95.85
6	17	1.86	894	97.70
7	12	1.31	906	99.02
8	1	0.11	907	99.13
9	2	0.22	909	99.34
10	1	0.11	910	99.45
11	1	0.11	911	99.56
12	2	0.22	913	99.78
16	1	0.11	914	99.89
19	1	0.11	915	100.00

# Examples: COUNTREG Procedure

**PROC COUNTREG** is then used to fit Poisson and negative binomial models to the data.

For each model, the **PROBCOUNTS** macro computes the probability that the number of published articles is  $m$ , where  $m$  is a value in a list of nonnegative integers specified in the **COUNTS=** option.

The computations require the parameter estimates of the fitted model.

These are saved using the **ODS OUTPUT** statement as shown and passed to the **PROBCOUNTS** macro by using the **INMODEL=** option.

Variables containing the probabilities are created with names beginning with the **PREFIX=** string followed by the **COUNTS=** values and are saved in the **OUT=** data set.

# Examples: COUNTREG Procedure

For the **Poisson** model, the variables *poi0*, *poi1*, ... , *poi10* are created and saved in the data set **predpoi**, which also contains all of the variables in the **DATA=** data set.

The **PROBCOUNTS** macro is available from the Samples section at <http://support.sas.com>.

The following statements compute the estimates for Poisson and negative binomial models.

```
/*-- Poisson Model --*/  
proc countreg data=long97data;  
    model art=fem mar kid5 phd ment / dist=poisson;  
    ods output ParameterEstimates=pe;  
run;
```

# Examples: COUNTREG Procedure

```
%include probcounts;
```

```
%probcounts(data=long97data,  
            inmodel=pe,  
            counts=0 to 10,  
            prefix=poi, out=predpoi)
```

```
/*-- Negative Binomial Model --*/
```

```
proc countreg data=long97data;  
            model art=fem mar kid5 phd ment / dist=negbin(p=2);  
            ods output ParameterEstimates=pe;  
run;
```

```
%probcounts(data=predpoi,  
            inmodel=pe,  
            counts=0 to 10,  
            prefix=nb, out=prednb)
```

# Examples: COUNTREG Procedure

Parameter estimates for these two models are shown in the section “**Getting Started: COUNTREG Procedure**” on page 7.

For each model, the **predicted proportion of zero articles** can be calculated as the **average predicted probability** of zero articles across all scientists as shown in the **macro probcounts** in the following program.

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 ( $p < 0:0001$ ).

As a result, the **negative binomial** model is **preferred** to the **Poisson model**.

# Examples: COUNTREG Procedure

Another way to account for the large number of zeros in this data set is to fit a **zero-inflated Poisson (ZIP)** or a **zero-inflated negative binomial (ZINB)** model.

In the following statements, **DIST=ZIP** requests the **ZIP** model.

In the **ZEROMODEL** statement, you can specify the predictors,  $z$ , for the process that generated the additional zeros.

The **ZEROMODEL** statement also specifies the model for the probability  $\varphi$ .

By default, a **logistic model** is used for  $\varphi$ . The default can be changed using the **LINK=** option.

In this particular **ZIP** model, all variables used to model the article counts are also used to model  $\varphi$ .

# Examples: COUNTREG Procedure

```
proc countreg data=long97data;  
    model art = fem mar kid5 phd ment / dist=zip;  
    zeromodel art ~ fem mar kid5 phd ment;  
    ods output ParameterEstimates=pe;
```

```
run;
```

```
%probcntns(data=prednb,  
    inmodel=pe,  
    counts=0 to 10,  
    prefix=zip, out=predzip)
```

# Examples: COUNTREG Procedure

The parameters of the **ZIP** model are displayed below.

The first set of parameters gives the estimates of  $\varphi$  in the model for the **Poisson** process mean.

Parameters with the prefix “**lnf\_**” are the estimates of  $\gamma$  in the logistic model for  $\varphi$ .

## Model Fit Summary

Dependent Variable	art
Number of Observations	915
Data Set	SIGSTAT.LONG97DATA
<b>Model</b>	<b>ZIP</b>
<b>ZI Link Function</b>	<b>Logistic</b>
Log Likelihood	-1605
Maximum Absolute Gradient	2.08804E-7
Number of Iterations	16
Optimization Method	Newton-Raphson
AIC	3234
SBC	3291

# Examples: COUNTREG Procedure

## Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	0.640838	0.121306	5.28	<.0001
fem	1	-0.209145	0.063405	-3.30	0.0010
mar	1	0.103751	0.071111	1.46	0.1446
kid5	1	-0.143320	0.047429	-3.02	0.0025
phd	1	-0.006166	0.031008	-0.20	0.8424
ment	1	0.018098	0.002295	7.89	<.0001
Inf_Intercept	1	-0.577060	0.509383	-1.13	0.2573
Inf_fem	1	0.109747	0.280082	0.39	0.6952
Inf_mar	1	-0.354013	0.317611	-1.11	0.2650
Inf_kid5	1	0.217101	0.196481	1.10	0.2692
Inf_phd	1	0.001272	0.145262	0.01	0.9930
Inf_ment	1	-0.134114	0.045244	-2.96	0.0030

# Examples: COUNTREG Procedure

The **proportion of zeros** predicted by the **ZIP** model is 0.2986, which is much closer to the observed proportion than the Poisson model.

But the graph below shows that **both models deviate** from the observed proportions at one, two, and three articles.

The **ZINB** model is specified by the **DIST=ZINB** option.

All variables are again used to model both the number of articles and  $\varphi$ .

The **METHOD=QN** option specifies that the quasi-Newton method be used to fit the model rather than the default Newton-Raphson method.

These options are implemented in the following program.

# Examples: COUNTREG Procedure

```
proc countreg data=long97data;  
    model art=fem mar kid5 phd ment / dist=zinb method=qn;  
    zeromodel art ~ fem mar kid5 phd ment;  
    ods output ParameterEstimates=pe;  
run;  
  
%probcunts(data=predzip,  
    inmodel=pe,  
    counts=0 to 10,  
    prefix=zinb, out=predzinb)
```

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 ( $p < 0:0001$ ).

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.

# Examples: COUNTREG Procedure

## Model Fit Summary

Dependent Variable	art
Number of Observations	915
Data Set	SIGSTAT.LONG97DATA
<b>Model</b>	<b>ZINB</b>
<b>ZI Link Function</b>	<b>Logistic</b>
Log Likelihood	-1550
Maximum Absolute Gradient	0.00438
Number of Iterations	85
<b>Optimization Method</b>	<b>Quasi-Newton</b>
AIC	3126
SBC	3189

# Examples: COUNTREG Procedure

## Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	0.416747	0.143596	2.90	0.0037
fem	1	-0.195507	0.075592	-2.59	0.0097
mar	1	0.097583	0.084452	1.16	0.2479
kid5	1	-0.151733	0.054206	-2.80	0.0051
phd	1	-0.000700	0.036270	-0.02	0.9846
ment	1	0.024786	0.003493	7.10	<.0001
Inf_Intercept	1	-0.191694	1.322806	-0.14	0.8848
Inf_fem	1	0.635942	0.848911	0.75	0.4538
Inf_mar	1	-1.499466	0.938659	-1.60	0.1102
Inf_kid5	1	0.628432	0.442779	1.42	0.1558
Inf_phd	1	-0.037715	0.308004	-0.12	0.9025
Inf_ment	1	-0.882290	0.316222	-2.79	0.0053
_Alpha	1	0.376681	0.051029	7.38	<.0001

Note: I estimated this model using the default **Newton-Raphson** algorithm. Although the algorithm converged, **\_Alpha** was not estimated.

# Examples: COUNTREG Procedure

The following statements compute the **average predicted count probability** across all scientists for each count 0, 1, ..., 10.

The **averages for each model**, along with the observed proportions, are then arranged for plotting by **PROC SGPLOT**.

```
proc summary data=predzinb;  
  var poi0-poi10 nb0-nb10 zip0-zip10 zinb0-zinb10;  
  output out=mnpoi mean(poi0-poi10) =mn0-mn10;  
  output out=mnnb mean(nb0-nb10) =mn0-mn10;  
  output out=mnzip mean(zip0-zip10) =mn0-mn10;  
  output out=mnzinb mean(zinb0-zinb10)=mn0-mn10;  
run;  
  
data means;  
  set mnpoi mnnb mnzip mnzinb;  
  drop _type_ _freq_;  
run;
```

# Examples: COUNTREG Procedure

```
proc transpose data=means out=tmeans;  
run;
```

```
data allpred;  
  merge obs(where=(art<=10)) tmeans;  
  obs=percent/100;  
run;
```

```
proc sgplot;  
  yaxis label='Probability';  
  xaxis label='Number of Articles';  
  series y=obs x=art / name='obs' legendlabel='Observed'  
  lineattrs=(color=black thickness=4px);  
  series y=col1 x=art / name='poi' legendlabel='Poisson'  
  lineattrs=(color=blue);  
  series y=col2 x=art/ name='nb' legendlabel='Negative Binomial'  
  lineattrs=(color=red);  
  series y=col3 x=art/ name='zip' legendlabel='ZIP'  
  lineattrs=(color=blue pattern=2);  
  series y=col4 x=art/ name='zinb' legendlabel='ZINB'  
  lineattrs=(color=red pattern=2);  
  discretelegend 'poi' 'zip' 'nb' 'zinb' 'obs' / title='Models:'  
  location=inside position=ne across=2 down=3;
```

```
run;
```

# Examples: COUNTREG Procedure

For each of the four fitted models, the graph below shows the **average predicted count probability** for each article count across all scientists.

The **Poisson** model clearly **underestimates** the proportion of zero articles published, while the other three models are quite accurate at zero.

All of the models do well at the larger numbers of articles.

# Examples: COUNTREG Procedure

