

## Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models Using the COUNTREG Procedure

Donald Erdman, Laura Jackson, Arthur Sinko, SAS Institute Inc., Cary, NC

### ABSTRACT

Real-life count data are frequently characterized by overdispersion and excess zeros. Zero-inflated count models provide a parsimonious yet powerful way to model this type of situation. Such models assume that the data are a mixture of two separate data generation processes: one generates only zeros, and the other is either a Poisson or a negative binomial data-generating process. The result of a Bernoulli trial is used to determine which of the two processes generates an observation.

### 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 the number of times an event occurs. Some examples of event counts are:

- number of claims per year on a particular car owner's auto insurance policy
- number of workdays missed due to sickness of a dependent in a 4-week period
- number of papers published per year by a researcher

In count regression, the conditional mean  $E(y_i|\mathbf{x}_i)$  of the dependent variable,  $y_i$ , is assumed to be a function of a vector of covariates,  $\mathbf{x}_i$ . Possible covariates for the auto insurance example are:

- age of the driver
- type of car
- daily commuting distance

### MARGINAL EFFECTS IN COUNT REGRESSION

Marginal effects provide a way to measure the effect of each covariate on the dependent variable. The marginal effect of one covariate is the expected instantaneous rate of change in the dependent variable as a function of the change in that covariate, while keeping all other covariates constant. Unlike in linear models, the derivative of the conditional expectation with respect to  $x_{i,j}$  is no longer equal to  $\beta_j$ —that is,  $\partial E(y_i|\mathbf{x}_i)/\partial x_{i,j} \neq \beta_j$ . For example, for the Poisson regression with  $E(y_i|\mathbf{x}_i) = e^{\mathbf{x}_i\boldsymbol{\beta}}$  is

$$\frac{\partial E(y_i|\mathbf{x}_i)}{\partial x_{i,j}} = \beta_j e^{\mathbf{x}_i\boldsymbol{\beta}} = \beta_j E(y_i|\mathbf{x}_i) \quad (1)$$

Therefore the marginal effect of the change in covariate  $x_{i,j}$  depends not only on  $\beta_j$ , but also on all other estimated coefficients, and on all other covariate values. Another interpretation is that a one-unit change in the  $j$ th covariate leads to a proportional change in the conditional mean  $E(y_i|\mathbf{x}_i)$  of  $\beta_j$ .

### BASIC MODELS: POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS

The Poisson (log-linear) regression model is the most basic model that explicitly takes into account the nonnegative integer-valued aspect of the dependent count variable. In this model, the probability of an event count  $y_i$ , given the vector of covariates  $\mathbf{x}_i$ , is given by the Poisson distribution:

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

The mean parameter  $\mu_i$  (the conditional mean number of events in period  $i$ ) is a function of the vector of covariates in period  $i$ :

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

where  $\boldsymbol{\beta}$  is a  $(k + 1) \times 1$  parameter vector. (The intercept is  $\beta_0$ , and the coefficients for the  $k$  covariates are  $\beta_1, \dots, \beta_k$ .) Taking the exponential of  $\mathbf{x}_i' \boldsymbol{\beta}$  ensures that the mean parameter  $\mu_i$  is nonnegative. The name log-linear model is also used for the Poisson regression model because the logarithm of the conditional mean is linear in the parameters:

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

The Poisson regression model assumes that the data are equally dispersed—that is, that the conditional variance equals the conditional mean. The COUNTREG procedure uses maximum likelihood estimation to find the regression coefficients. The following statements demonstrate how the Poisson model can be estimated:

```
proc countreg data=a;
  model ypoizim=x1 x2/dist=poisson;
run;
```

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. The negative binomial regression model is a generalization of the Poisson regression model that allows for overdispersion by introducing an unobserved heterogeneity term for observation  $i$ . Observations are assumed to differ randomly in a manner that is not fully accounted for by the observed covariates. In the negative binomial model,

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

where  $\tau_i$  follows a gamma( $\theta, \theta$ ) distribution with  $E(\tau_i) = 1$  and  $V(\tau_i) = 1/\theta$ . Conditional on both  $\mathbf{x}_i$  and  $\tau_i$ , the dependent count variable  $Y_i$  is still Poisson distributed:

$$P(Y_i = y_i | \mathbf{x}_i, \tau_i) = \frac{e^{-\mu_i \tau_i} (\mu_i \tau_i)^{y_i}}{y_i!}$$

However, conditional on only  $\mathbf{x}_i$ ,  $Y_i$  is distributed as a negative binomial:

$$P(Y_i = y_i | \mathbf{x}_i) = \frac{\theta^\theta \mu_i^{y_i} \Gamma(\theta + y_i)}{\Gamma(y_i + 1) \Gamma(\theta) (\mu_i + \theta)^{\theta + y_i}}$$

The distribution has conditional mean  $\mu_i$  and conditional variance  $\mu_i(1 + (1/\theta)\mu_i)$ . It is more straightforward to estimate  $\alpha = 1/\theta$  instead of  $\theta$ . With this substitution, the conditional variance is  $\mu_i(1 + \alpha\mu_i)$ . Negative binomial and Poisson models are nested because as  $\alpha$  converges to 0, the negative binomial distribution converges to Poisson. Cameron and Trivedi consider a general class of negative binomial models with mean  $\mu_i + \alpha\mu_i^p$ , where in general  $-\infty < p < \infty$  (Cameron and Trivedi 1986). PROC COUNTREG estimates two negative binomial models, corresponding to  $p = 2$  (with variance  $\mu_i + \alpha\mu_i^2$ ) and  $p = 1$  (with variance  $\mu_i + \alpha\mu_i$ ). The first is estimated with the option DIST=NEGBIN(p=2), and the second is estimated using DIST=NEGBIN(p=1). The following statements show how to estimate the first:

```
proc countreg data=a;
  model ypoizim=x1 x2/dist=negbin(p=2);
run;
```

## ADVANCED MODELS: ZERO-INFLATED MODELS

The main motivation for zero-inflated count models is that real-life data frequently display overdispersion and excess zeros (Lambert 1992; Greene 1994). Zero-inflated count models provide a way of modeling the excess zeros in addition to allowing for overdispersion. In particular, for each observation, there are two possible data generation processes; the result of a Bernoulli trial determines which process 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, whereas Process 2,  $g(y_i | \mathbf{x}_i)$ , generates counts from either a Poisson or a negative binomial model. In general:

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

The probability of  $\{Y_i = y_i | \mathbf{x}_i\}$  is

$$P(Y_i = y_i | \mathbf{x}_i, \mathbf{z}_i) = \begin{cases} \varphi(\boldsymbol{\gamma}'\mathbf{z}_i) & + \{1 - \varphi(\boldsymbol{\gamma}'\mathbf{z}_i)\}g(0|\mathbf{x}_i) & \text{if } y_i = 0 \\ \{1 - \varphi(\boldsymbol{\gamma}'\mathbf{z}_i)\}g(y_i|\mathbf{x}_i) & & \text{if } y_i > 0 \end{cases}$$

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 vector of zero-inflated covariates and  $\boldsymbol{\gamma}$  is the vector of zero-inflated coefficients to be estimated. The function  $F$  that relates the product  $\mathbf{z}_i'\boldsymbol{\gamma}$  (which is a scalar) to the probability  $\varphi_i$  is called the zero-inflated link function, and it can be specified as either the logistic function or the standard normal cumulative distribution function (the probit function).

To estimate a zero-inflated model with the COUNTREG procedure, use the ZEROMODEL statement with a dependent variable (the same dependent variable as in the MODEL statement), a vector of covariate variables  $\mathbf{z}_i$ , and a link function. The following statements demonstrate the use of the ZEROMODEL statement:

```
proc countreg data=a;
  model ypoizim=x1 x2/dist=poisson;
  zeromodel ypoizim ~ x3 /link=normal;
run;
```

The mean and variance of the zero-inflated Poisson model (ZIP) are:

$$\begin{aligned} E(y_i | \mathbf{x}_i, \mathbf{z}_i) &= \mu_i(1 - \varphi_i) \\ V(y_i | \mathbf{x}_i, \mathbf{z}_i) &= \mu_i(1 - \varphi_i)(1 + \mu_i\varphi_i) \end{aligned}$$

The mean and variance of the zero-inflated negative binomial model (ZINB) are:

$$\begin{aligned} E(y_i | \mathbf{x}_i, \mathbf{z}_i) &= \mu_i(1 - \varphi_i) \\ V(y_i | \mathbf{x}_i, \mathbf{z}_i) &= \mu_i(1 - \varphi_i)(1 + \mu_i(\varphi_i + \alpha)) \end{aligned}$$

Both zero-inflated models demonstrate overdispersion:  $V(y_i | \mathbf{x}_i, \mathbf{z}_i) > E(y_i | \mathbf{x}_i, \mathbf{z}_i)$ .

## SIMULATED EXAMPLE

In this section we generate four large ( $n = 10000$ ) data sets from each of the Poisson, negative binomial, zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) distributions. Then we try to fit each of these data sets with the four corresponding count regression models. The Poisson and negative binomial data sets are generated using the same conditional mean:

$$\mu_i = e^{1+0.3x_{1i}+0.3x_{2i}} \quad (2)$$

In addition, the negative binomial model further uses the parameter  $\theta = \alpha = 1$ . The zero-inflated models use  $\varphi_i = \Lambda(2x_{3i})$  (the standard normal cumulative distribution function) for the zero-inflated link function, such that the probability of  $\{Y_i = y_i | \mathbf{x}_i\}$  is:

$$P(Y_i = y_i | \mathbf{x}_i, \mathbf{z}_i) = \begin{cases} \Lambda(\boldsymbol{\gamma}'\mathbf{z}_i) & + \{1 - \Lambda(\boldsymbol{\gamma}'\mathbf{z}_i)\}g(0|\mathbf{x}_i) & \text{if } y_i = 0 \\ \{1 - \Lambda(\boldsymbol{\gamma}'\mathbf{z}_i)\}g(y_i|\mathbf{x}_i) & & \text{if } y_i > 0 \end{cases}$$

where  $g(\cdot)$  is either a Poisson distribution (with conditional mean  $\mu_i$ ) or a negative binomial distribution (with conditional mean  $\mu_i$  and parameter  $\theta = \alpha = 1$ ).

The following algorithm summarizes our method:

1. Generate 10000 count observations each using distribution  $i = 1, 2, 3, 4$ .
2. Estimate each count data set  $i$  by using four models  $j = 1, 2, 3, 4$ .
3. Compare the outcomes of the estimation with the actual values.

The first step is achieved with the following statements:

```

data a; /* generate the data */
call streaminit(1234);
do kk=1 to 10000;
  x1 = rannor(1234);
  x2 = rannor(1234);
  x3 = rannor(1234);
  theta = 1;
  mu = exp(1 + .3*x1 + .3*x2);
  parml = 1/(1+mu/theta);
  yneg = rand('NEGB', parml, theta);
  ypoi = ranpoi(1234, mu);
  pzero = cdf('LOGISTIC', x3*2);
  if ranuni(1234)>pzero then do;
    ynegzim = yneg;
    ypoizim = ypoi;
  end;
  else do;
    ynegzim = 0;
    ypoizim = 0;
  end;
  y=ynegzim;
  output ;
end ;
run;

```

The second step involves four estimation procedures for each of the four different dependent variables. We focus on two cases in detail. Our goal is to demonstrate how a fitted zero-inflated negative binomial model performs in the presence of model misspecification. In Case 1, a zero-inflated negative binomial model is fit to the data generated by the zero-inflated negative binomial distribution (dependent variable ynegzim). In Case 2, a zero-inflated negative binomial model is fit to the data generated by the plain negative binomial distribution (dependent variable yneg).

```

/*** Case 1 ***/
proc countreg data=a;
  model ynegzim=x1 x2 / dist=zinb method=qn;
  zeromodel ynegzim ~ x3;
  ods output ParameterEstimates=pe;
run;

/*** Case 2 ***/
proc countreg data=a;
  model yneg=x1 x2 / dist=zinb method=qn;
  zeromodel yneg ~ x3;
  ods output ParameterEstimates=pe;
run;

```

Figure 1 shows the output from Case 1, and Figure 2 shows the output from Case 2.

**Figure 1** PROC COUNTREG Results for ZINB Estimation (True Model is ZINB)

The COUNTREG Procedure	
Model Fit Summary	
Dependent Variable	ynegzim
Number of Observations	10000
Data Set	WORK.A
Model	ZINB
ZI Link Function	Logistic
Log Likelihood	-13144
Maximum Absolute Gradient	0.0004233
Number of Iterations	27
Optimization Method	Quasi-Newton
AIC	26301
SBC	26344

Figure 1 continued

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	1.026066	0.022038	46.56	<.0001
x1	1	0.279170	0.017555	15.90	<.0001
x2	1	0.266697	0.017215	15.49	<.0001
Inf_Intercept	1	0.046080	0.052786	0.87	0.3827
Inf_x3	1	1.989918	0.069677	28.56	<.0001
_Alpha	1	0.991183	0.049308	20.10	<.0001

Figure 2 PROC COUNTREG Results for ZINB Estimation (True Model is NB)

The COUNTREG Procedure					
Model Fit Summary					
Dependent Variable		yneg			
Number of Observations		10000			
Data Set		WORK.A			
Model		ZINB			
ZI Link Function		Logistic			
Log Likelihood		-21659			
Maximum Absolute Gradient		0.0006253			
Number of Iterations		35			
Optimization Method		Quasi-Newton			
AIC		43331			
SBC		43374			
Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	1.005908	0.017418	57.75	<.0001
x1	1	0.293607	0.011888	24.70	<.0001
x2	1	0.284540	0.011864	23.98	<.0001
Inf_Intercept	1	-4.354450	1.008171	-4.32	<.0001
Inf_x3	1	0.227890	0.325382	0.70	0.4837
_Alpha	1	0.995485	0.041769	23.83	<.0001

The main difference between the two estimations is the value of Inf\_Intercept. When this variable is statistically significant and significantly negative, it is a strong sign that a negative binomial specification is preferred to the zero-inflated negative binomial.

In addition, the negative binomial model (respectively, the zero-inflated negative binomial model) has a built-in test for whether the underlying data are Poisson (respectively, zero-inflated Poisson). Recall that the Poisson distribution possesses the property of equal dispersion (the mean is equal to the variance). When fitting a negative binomial model (respectively, a ZINB model), a test of whether `_Alpha` is significantly different from zero is a way to evaluate whether the true specification is Poisson (respectively, zero-inflated Poisson).

In Case 1, we can reject the zero-inflated Poisson model, because `_Alpha` is significantly different from zero (`_Alpha` = 0.991 with  $p$ -value < 0.0001). In Case 2, we also reject the zero-inflated Poisson model (`_Alpha` = 0.995 with  $p$ -value < 0.0001).

To accurately test whether the data used in Case 2 (dependent variable `yneg`, generated by the negative binomial) is Poisson, we must test it against the negative binomial model, not against the zero-inflated negative binomial. The statements below present Case 3, in which a negative binomial model is now fitted to the data used in Case 2 (that is, the model is now correctly specified). Figure 3 shows the output from Case 3.

```

/*** Case 3 ***/
proc countreg data=a;
  model yneg=x1 x2 / dist=negbin(p=2) method=qn;
  ods output ParameterEstimates=pe;
run;

```

Figure 3 presents the estimation results.

**Figure 3** PROC COUNTREG Results for NB Estimation (True Model is NB)

The COUNTREG Procedure					
Model Fit Summary					
Dependent Variable					yneg
Number of Observations					10000
Data Set					WORK.A
Model					NegBin
Log Likelihood					-21660
Maximum Absolute Gradient					0.0005555
Number of Iterations					13
Optimization Method					Quasi-Newton
AIC					43328
SBC					43357
Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	0.992781	0.011971	82.93	<.0001
x1	1	0.293645	0.011938	24.60	<.0001
x2	1	0.284071	0.011901	23.87	<.0001
_Alpha	1	1.032787	0.022156	46.61	<.0001

The results demonstrate that we can indeed reject the hypothesis that the process is Poisson, since  $\_Alpha = 1.033$  with  $p$ -value  $< 0.0001$ , and thus the variance of the process is larger than the mean. The graph in Figure 4 shows that the zero-inflated negative binomial model (NegBinZIM) describes the empirical probability distribution very well, even though they are not nested. The key to understanding this behavior lies in the intercept value of the zero-inflated part. A relatively large negative constant shows that the zero-inflated part is quite small and that the zero-inflated negative binomial model is observationally equivalent to the negative binomial model.

We turn now to the last step of the algorithm. One of the most popular approaches for comparing the performance of different models is to compare the sample probability distribution of the data to the average probability distributions predicted using the estimated models (Long 1997, p. 223)—that is, we have to compare  $\overline{Pr}(Y = y_i)$

$$\overline{Pr}(Y = m) = \frac{1}{N} \sum_{k=1}^N I(y_k - m) \tag{3}$$

$$I(y_k - m) = \begin{cases} 1 & \text{if } y_k = m \\ 0 & \text{otherwise} \end{cases}$$

with the average probabilities implied by the estimated models

$$\overline{\widehat{Pr}}(Y = m) = \frac{1}{N} \sum_{k=1}^N \widehat{Pr}(y_k = m | x_k) \tag{4}$$

Equations 3 and 4 can be evaluated in the following way. After fitting the data with each model, the PROBCOUNTS macro computes the probability that  $y_i$  is equal to  $m$ , 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 and passed to the PROBCOUNTS macro by using the INMODEL= option, as shown in the following statements. Variables containing the probabilities are created with names that begin with the PREFIX= string followed by the COUNTS= values and are saved in the OUT= data set. 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 the four models and construct average probability distributions.

```

proc countreg data=a;
  model y=x1 x2 / dist=zip;
  zeromodel y ~ x3;
  ods output ParameterEstimates=pe;
run;

%probcunts (data=prednb,
            inmodel=pe,
            counts=0 to 20,
            prefix=zip, out=predzip)

proc countreg data=a;
  model y=x1 x2 / dist=zinb method=qn;
  zeromodel y ~ x3;
  ods output ParameterEstimates=pe;
run;

%probcunts (data=predzip,
            inmodel=pe,
            counts=0 to 20,
            prefix=zinb, out=predzinb)

proc summary data=predzinb;
  var poi0-poi8 nb0-nb8 zip0-zip8 zinb0-zinb8;
  output out=mnpoi mean(poi0-poi8) =mn0-mn8;
  output out=mnnb mean(nb0-nb8) =mn0-mn8;
  output out=mnzip mean(zip0-zip8) =mn0-mn8;
  output out=mnzinb mean(zinb0-zinb8)=mn0-mn8;
run;

data means;
  set mnpoi mnnb mnzip mnzinb;
  drop _type_ _freq_;
run;

proc transpose data=means out=tmeans;
run;

```

The summarized results of the third step are shown in [Figure 4](#) and [Figure 5](#). [Figure 4](#) shows the averages of the estimated probability distributions (blue and red lines) in addition to the empirical probability distribution for the four different data generation processes. [Figure 5](#) presents the differences between the estimated (Equation 4) and the empirical (Equation 3) probability distributions. Since the sample is reasonably large ( $n = 10000$ ), we conclude that the empirical distributions are “close enough” to the population distributions. The same is true for the estimated models.

Each figure contains four subplots. Each subplot corresponds to the estimation of the different data generation processes. The first row shows the estimation results for Poisson and zero-inflated Poisson (PoissonZIM) data, and the second row shows the same for the negative binomial (NegBin) and zero-inflated negative binomial (NegBinZIM) data. The results are easy to interpret. The first subplot shows how well Poisson data can be predicted using the count models we consider. It can be concluded that these models capture the features of Poisson data equally well. Analytically, it is straightforward to show that the Poisson model is a special case of the negative binomial model and the zero-inflated Poisson model is a special case of the zero-inflated negative binomial model.

In contrast, it is not possible to transform a zero-inflated Poisson model (respectively, a zero-inflated negative binomial model) to a plain Poisson (respectively, to a plain negative binomial model) by using any finite vector of coefficients (Greene 1994). The reasoning is the following: in order to reduce a zero-inflated model to its non-zero-inflated counterpart, it is necessary to have a cumulative distribution function  $F(z'_i \gamma) = 0$ . Since both the logistic and the standard normal cumulative distribution functions are strictly increasing and defined on the entire real line,  $F(z'_i \gamma) = 0$  if and only if  $z'_i \gamma = -\infty$ . However, as long as the vector of variables  $z_i$  contains an intercept or there is a linear combination of variables that is strictly negative or strictly positive, then  $\gamma$  can be chosen in a way that for all practical purposes  $\Phi(\delta'_i \gamma) = 0$ . The regression results shown in [Figure 2](#) support this assertion. The data generation process in this case is negative binomial, while the estimation model is zero-inflated negative binomial. They are not nested. However, in [Figure 4](#) they demonstrate observationally equivalent behavior. This feature occurs because the zero-inflated intercept is quite negative (Inf\_Intercept=  $-4.355$ ) and thus  $F(\text{Inf\_Intercept} + \text{Inf\_x3 } x_{3i})$  is sufficiently close to zero.

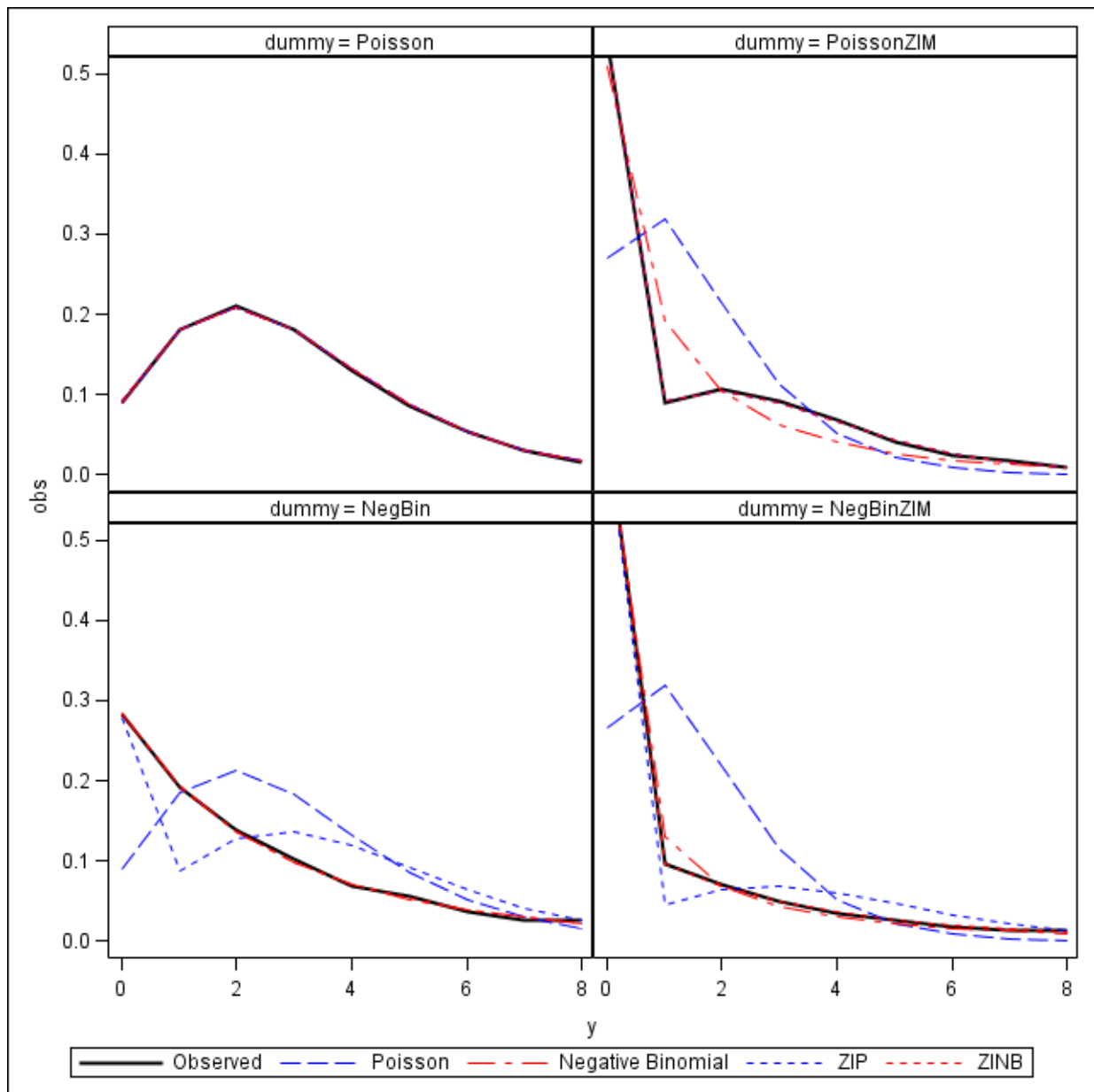
Finally, we summarize the performance of each of the four fitted models when fitted to each of the four types of

generated data:

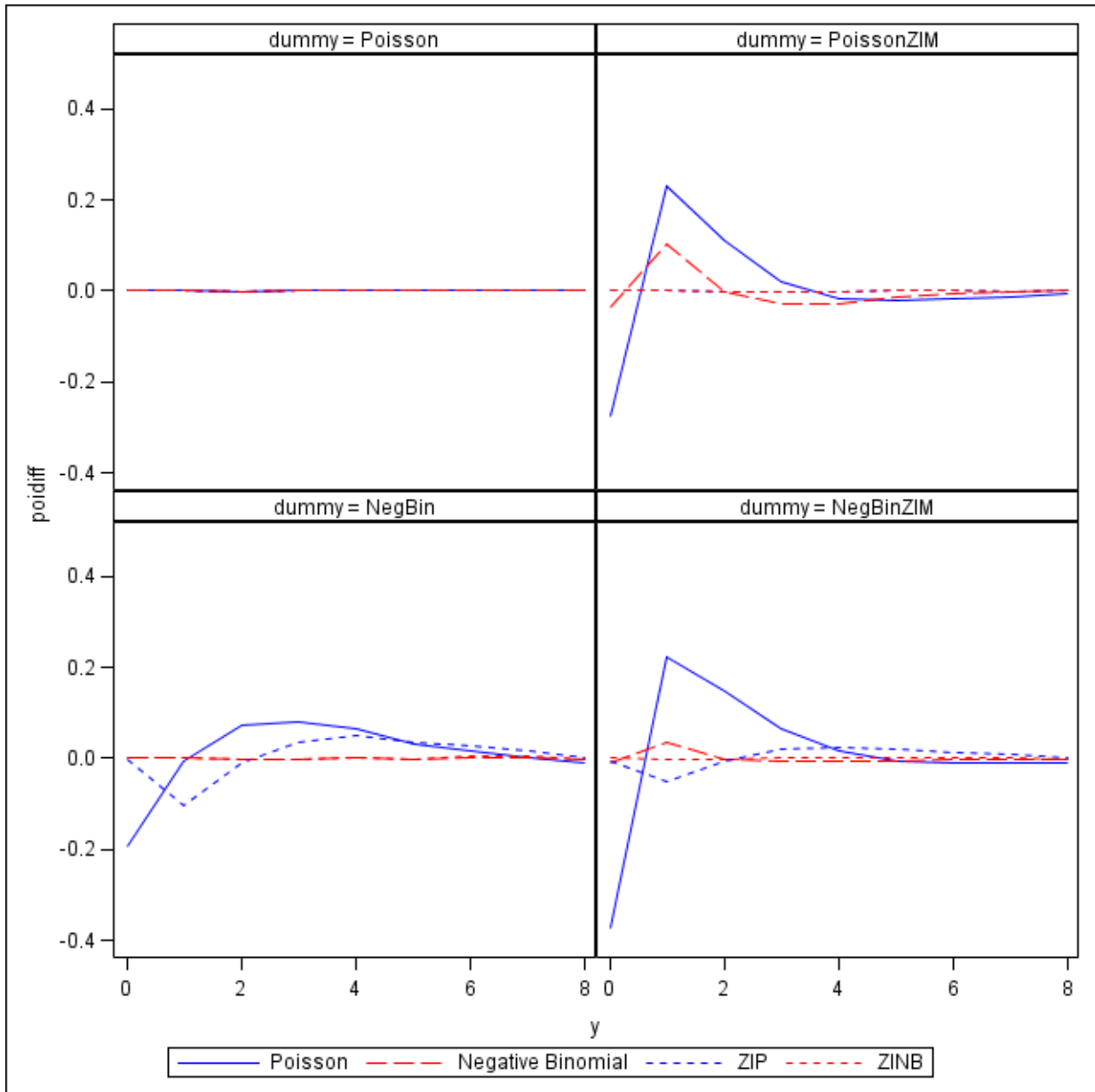
- The data generated by the Poisson distribution can be predicted equally well by each of the four models that we consider.
- The data generated by the zero-inflated Poisson can be predicted most accurately using either a zero-inflated Poisson or a zero-inflated negative binomial model. The negative binomial model performs next best. The Poisson model fares the worst: it significantly underpredicts the number of zeros and overpredicts the number of ones.
- The data generated by the negative binomial process can be predicted equally well by either a negative binomial or a zero-inflated negative binomial model. These models are followed by the zero-inflated Poisson and the Poisson.
- The data generated by the zero-inflated negative binomial model can be predicted best by a zero-inflated negative binomial, followed by a negative binomial, a zero-inflated Poisson, and a Poisson.

Notice that the Poisson model provides the worst fit in all cases other than in the case of Poisson-generated data. Thus, a Poisson model should be used only in cases where there is strong evidence that it is the correct specification. As long as data sample is reasonably large, a slight loss of efficiency is, on average, more preferable compared to model misspecification.

**Figure 4** Relative Performance of Different Models, Average Probability Distribution over the Sample



**Figure 5** Relative Performance of Different Models, Deviations from the Empirical Probability Distribution



## CONCLUSION

This paper studies the performance of different count models on a simulated example. The results demonstrate that among the count models we consider, in many cases a Poisson model tends to be overly restrictive. If model specification is unknown, it is safer to start from more general model (for example, zero inflated negative binomial) and then test whether this model specification can be reduced to more restrictive ones.

## REFERENCES

- Cameron, A. C. and Trivedi, P. K. (1986), "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators," *Journal of Applied Econometrics*, 1, 29–53.
- Greene, W. H. (1994), *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*, Technical report.

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

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

## **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at:

Arthur Sinko  
SAS Institute Inc.  
100 SAS Campus Drive, R5214  
Cary, NC 27513  
(919) 531-2133  
Arthur.Sinko@sas.com  
www.sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.