Paper 3768-2019
Zero-Inflated and Zero-Truncated Count Data Models with the NLMIXED Procedure

Robin High, University of Nebraska Medical Center, Omaha, NE

SAS/STAT® and SAS/ETS® software have several procedures for analyzing count data based on the Poisson distribution or the negative binomial distribution with a quadratic variance function (NB-2). Count data may either have an excess number of zeros (inflation) or the situation where zero is not an outcome (truncation). Zero-inflated Poisson and negative binomial models are available with the COUNTREG, GENMOD, and FMM procedures. The FMM procedure also provides options for the zero-truncated Poisson and negative binomial distributions. Other types of count data models include the restricted and unrestricted generalized Poisson, negative binomial with a linear variance function (NB-1), and Poisson-Inverse Gaussian (P-IG) and likewise may be subject to zero-inflation or zero-truncation. Programming statements entered into the NLMIXED procedure in SAS/STAT® can model zero-inflated and zero-truncated count data with these distributions and may improve model fit which can be examined with the Vuong test or by comparing various fit statistics.

INTRODUCTION

For data having non-negative integer outcomes (count data), the two primary models available with SAS/STAT® software are based on the Poisson and negative binomial (NB-2) distributions. With count data, the outcome of zero may be the source of two problems:

- Inflation: excess zeros are present when compared to the expected number based on the count data distribution
- Truncation: zeros do not exist

These two situations are often ignored, perhaps due to lack of awareness how these conditions may affect results or lack of familiarity with or access to available software. With zero-inflation, a model can be developed that considers reasons why a zero is generated outside the count data model. A zero-truncated model acknowledges the reality that a zero does not exist.

In both situations other count data distributions can be examined in addition to the Poisson or negative binomial. Zero inflation and zero-truncation also contribute to overdispersion which affect inferences. The objective of this paper is to describe the coding process entered into the NLMIXED procedure to estimate both zero-inflated and zero-truncated count data models for several types of count data distributions. Other variations on these models exist, including k-inflation (Famoye and Singh, 2003) where one specific outcome is identified (e.g., k=y=1) which will have a greater number of responses than expected with the chosen distribution; also left-truncation can occur at the value C, an integer greater than or equal to 0; the most common situation is truncation at C=0.

Alternative parameter estimation methods for several count data models were described in High (2018). The methods to account for zero-inflation or zero-truncation follow directly from the log-likelihood equations for these models with a modifications necessary for their implementation. General formulas for the conditional means and variances of predicted values are provided. An overview of methods to assess and compare the fit of these various models with information criteria is described by Christensen (2018) and also with the Vuong test, usually applied to help decide whether the zero-inflated model is a preferred choice over the standard model. All estimation procedures for these distributions can be programmed with statements entered within the NLMIXED procedure.

COUNT DATA PROBABILITY DISTRIBUTIONS

Poisson (P)

The basic count data distribution is the Poisson with probability density function:

$$f(Y = y \mid \mu) = (\mu^{y} * e^{-\mu}) / y! \qquad \text{for } y = 0, 1, 2, ..$$

The mean and variance of the Poisson distribution are both equal to μ which implies it is frequently an unrealistic choice because of overdispersion (i.e., the variability of the data exceed the variability assumed by the model). This restrictive feature can be dealt with through other count data distributions which include a dispersion parameter (depending on the distribution, it is named delta, k, alpha, or tau). The Poisson distribution is a special case of these distributions since their probabilities are close or equal to the Poisson as the dispersion parameter either approaches or equals 0.

Unrestricted and Restricted Generalized Poisson (UGP / RGP)

The unrestricted generalized Poisson (UGP) probability density function is described by Consul (1989) and also Harris, Yang, and Hardin (2012). The formula in their notation includes the mean θ and a dispersion parameter δ in its unrestricted form (i.e., the mean and dispersion are independent):

$$\text{UGP: } f(y,\theta,\delta) = (\theta * (\theta + \delta y)^{(y-1)} * e^{-(\theta + \delta y)}) / y! \qquad y=0,1,2, ..$$

The restricted generalized Poisson (RGP) is derived from UGP pdf. The dispersion parameter δ of the UGP may be proportional to the mean, that is, let $\delta=\alpha\theta$ and the density function becomes:

$$\begin{aligned} f(y,\theta,\alpha\theta) &= (\theta * (\theta + \alpha\theta y)^{(y-1)} * e^{-(\theta + \alpha\theta y)}) / y! \\ &= (\theta^{y} * (1 + \alpha y)^{(y-1)} * e^{-\theta(1 + \alpha y)}) / y! \end{aligned}$$

Setting $\mu = \theta * (1-\alpha\theta)^{-1}$ (the expected value formula for UGP from Table 1) and solving yields $\theta=\mu/(1+\alpha\mu)$. The probability density function for the restricted generalized Poisson distribution (RGP) is obtained by substituting $\mu/(1+\alpha\mu)$ for θ into the UGP density function (Famoye, 1993):

$$\text{RGP: } f(y,\mu,\alpha)=(\mu/(1+\alpha\mu))^{y} * (1+\alpha y)^{(y-1)} * e^{[(-\mu*(1+\alpha y))/(1+\alpha y)]} / y! \qquad y=0,1,2,..$$

where μ is the mean and α is the dispersion parameter. Both the UGP and RGP can work with data having either over- or under-dispersion (though the amount of under-dispersion is limited). Both distributions equal the Poisson when their dispersion parameter equal 0.

Negative Binomial Distributions

The negative binomial distribution is a special case of a class of models defined by their variance functions identified with three parameters: μ, k, and P where the dispersion parameters k and P are both greater than 0. Since k must be positive, the negative binomial distribution can only deal with overdispersion. The Poisson distribution is a limiting case of these negative binomial distributions as k approaches 0 from the right (Hilbe, p. 221, 2011); that is, with small, positive k, results from the Poisson and the negative binomial distributions, both having a log link, are nearly the same.

Quadratic Negative Binomial Distribution (NB-2)

The most commonly applied form of the negative binomial distribution has a quadratic variance function (see Table 1) with mean μ and dispersion parameter k:

$$f(y,\mu,k) = (k*\mu)^{y} * (1 + (k*\mu))^{-(y + (1/k))} * \frac{\Gamma(y + (1/k))}{\Gamma(1/k) * \Gamma(y+1)}$$

For a non-negative integer y, mean=mu, and dispersion=k, the NB-2 probabilities can be generated with the following SAS® function:

```
      mk = 1/(1 + (mu*k));
 negbin2 = PDF('NEGBINOMIAL', y, mk, 1/k);
```

Linear Negative Binomial Distribution (NB-1)

The NB-1 distribution has a linear variance function (see Table 1) and is derived by replacing the dispersion parameter `k` in the NB-2 distribution with `k/mu` and removing the terms `mu/mu=1` or replacing `1/(k/mu)` with `mu/k`.

$$f(y,\mu,k) = k^{y} * (1+k)^{-(y + (\mu/k))} \frac{\Gamma(y + (\mu/k))}{\Gamma(\mu/k) * \Gamma(y+1)}$$

Probabilities for the NB-1 distribution can be generated with the following modifications to the negative binomial function for the NB-2 distribution:

```
      mk = 1/(1 + k);
 negbin1 = PDF('NEGBINOMIAL', y, mk, mu/k);
```

Three Parameter Negative Binomial Distribution (NB-P)

The pdf for NB-P is obtained from the NB-2 distribution with `k` replaced with $k*\mu^{(2-P)}$ (see Greene, 2008, p. 586 for details concerning the derivation of the negative binomial pdfs). The three-parameter negative binomial model (NB-P) allows more flexibility in working with overdispersion than is available with either the NB-1 or NB-2 distributions. The parameter P is the exponent in the distribution's variance function (see Table 1), thus the reason for naming it NB-P. The pdf for the NB-P looks intimidating in its complete form, but with the repetition of terms it reduces to an equation that resembles both the NB-1 and NB-2 distributions:

$$f(y,\mu,k,P) = (1-s)^{y} * s^{pm} \frac{\Gamma(y + pm)}{\Gamma(pm) * \Gamma(y+1)}$$

where

$$Q = 2-P$$
$$pm = (1/k) * \mu^{Q}$$
$$s = pm / (pm + \mu)$$

Though it is not immediately obvious from this formula, for a given mean and dispersion (μ,k), when P=1 (Q=1) the probabilities are the same as the NB-1 distribution. When P=2 (Q=0), the probabilities are the same as the NB-2 distribution.

Poisson-Inverse Gaussian Distribution (P-IG)

Applying the inverted Gaussian distribution for the mean of the Poisson distribution results in the Poisson-inverse Gaussian (P-IG) model. This model is especially relevant to work with extremely over-dispersed count data, beyond the situations appropriate for the negative binomial model (NB-2) or even the NB-P model with P > 2. The pdf for the Poisson-Inverse Gaussian distribution does not have a closed form as the other distributions described here. However, it does have a set of programmable equations (Zha, 2016, p. 23 and Dean, 1989, p. 173):

$$\begin{aligned} f(Y=0) &= EXP(\tau^{-1} * [1 - (1 + (2\tau\mu))^{1/2}]) \\ f(Y=1) &= [\mu * (1 + (2\tau\mu))^{-1/2}] * f(Y=0) \\ f(Y=y) &= [(2\tau\mu) / (1 + (2\tau\mu))] * [1 - (3/(2y))] * f(Y=y-1) \\ &\quad + [\mu^2 / (1+(2\tau\mu))] * [1 / (y*(y-1))] * f(Y=y-2) \qquad y=2,3,4.. \end{aligned}$$

where τ (tau) is the dispersion parameter. The computation of the probability for a given y progresses sequentially, starting with the probability of y=0 increasing by 1 up to y. For each value of y beginning with y=2, the probabilities of y-1 and y-2 are saved and appear in the third formula to compute f(Y=y).

Another derivation of the P-IG probability density function with $\tau=\mu^2/\eta$ with resulting variance $\mu + \mu\eta$ is shown in Guo and Trivedi (2002, p. 68) which has an equation for which the log likelihood can be programmed into NLMIXED; however, the gamma function gives a computational error (i.e., a missing value) for a response y greater than 76 (i.e., missing values result for (y+i) greater than or equal to 172 in the gamma function where i ranges from 0 to y-1 which is added to y).

ZERO-INFLATED COUNT DATA MODELS

Zero-inflated count data arise when excess zeros are observed in the data generating process when compared with the expected number of zeros that would be generated from the underlying process itself. The excess zeros are called "structural" zeros.

Suppose the density function for the count data model is f(Y=y) for y = 0, 1, 2, .. ∞; this function computes probabilities that sum to 1 for all integers greater than or equal to 0; the count data distributions presented above will be featured in this paper. An outcome of zero may occur due to factors outside the process that generates the data in which case a structural zero occurs with probability π $(0 < \pi < 1)$. Data from the count distribution are generated with probability (1 – π); the zeros from this source are called "sampling zeros." The zero-inflated probability density function for count data thus has the following general form:

$$\begin{aligned} Prob(Y = y) &= \Pi + (1 - \Pi) * f(Y=0) \qquad \text{for } y = 0 \\ &= \qquad (1 - \Pi) * f(Y=y) \qquad \text{for } y > 0 \end{aligned}$$

This density function sums to 1 for all values of y greater than or equal to 0. Just like the count data distribution, the zero-inflated distribution has a mean and variance; a general formula is given in a subsequent section.

The statements included in NLMIXED to run zero-inflated count data models requires the same types of statements applied with standard count data models (High, 2018):

```
PROC NLMIXED DATA =indat (rename=( < response >  = y ));
 PARMS < initial values for the coefficients of the two linear predictors > ;
etaZr = < Linear predictor for zero-inflation > ;
 p_zr = 1/(1 + exp(-etaZr)); * logit link, inverse for structural zeros ;
 etaN = < Linear predictor for the counts > ;
   mu = EXP(etaN);  * inverse function of the log link for the counts;

 lglk = < log likelihood statements for a zero-inflated model, see Appendix > ;

MODEL y ~ general( lglk ) ;
REPLICATE count;      * enter only if the same data rows are replicated by a count;
ESTIMATE "IRR" EXP(b1) ;  * estimate functions of the model parameters;
PREDICT mu   OUT=mu (KEEP= pred y rename=(pred=mu));  * mean and response;
PREDICT phi  OUT=phi(KEEP= pred   rename=(pred=phi)); * dispersion;
PREDICT p_zr OUT=pzr(KEEP= pred   rename=(pred=p_zr));* probability of structural 0;
RUN;
```

The primary differences from estimating the standard count data model are the addition of a second linear predictor (etaZr) for the binary component with its link function to model the structural zeros. Constructing these linear predictors follows the same guidelines as described in a previous SGF presentation (High and ElRayes, 2017). The loglikelihood equation for zero-inflated distributions includes separate components for the zeros and the counts greater than zero; when entered into NLMIXED it has the general form:

```
IF (y EQ 0) THEN lglk = LOG(  p_zr  + (1-p_zr)*( f(Y=0)) );
            ELSE lglk = LOG(1-p_zr) + LOG( f(Y=y) );
```

The first line of the IF / THEN statement accounts for the zeros (y EQ 0) as either due to zero-inflation (the structural component) or zeros generated by the count distribution. The second part (following ELSE) evaluates the counts greater than zero (y GE 1) multiplied by (1-p_zr). Whenever computationally possible, the log-likelihood is most efficiently computed by first taking the logs of the components of the PDF and summing them, rather than computing the probability and then taking the log (an exception to this rule is the Poisson-Inverse Gamma distribution). The formula to compute f(Y=0) requires fewer components than entering the complete pdf. Since a number multiplied by 0 is 0, or any number raised to the 0 power is 1, several terms of the pdf are usually not needed to express the probability of y=0. The minimal components to compute the f(Y=0) are given in Table 1. They are also included in the log-likelihood equations to be entered into PROC NLMIXED which are listed in the Appendix.

Zero-inflated count data models for two distributions, Poisson and negative binomial (NB-2), are available in the COUNTREG, GENMOD, and FMM procedures. For the zero-inflation component, the linear predictor and its inverse link (the default is the logit) estimate the probability of a structural 0. The FMM procedure works in the same manner; however, to match the signs of the coefficients from GENMOD and COUNTREG, the statement for the zero-inflated component is placed first followed by the MODEL statement for the count data distribution (see the Appendix for examples). The NLMIXED code presented here evaluates structural zeros as the outcome in this manner for all count data models.

ZERO-INFLATED MODEL COEFFICIENTS

Coefficients from these zero-inflated models usually have the same sign and values of similar magnitudes; the standard errors will differ depending on the extent of overdispersion. In particular, without a dispersion parameter, the zero-inflated Poisson coefficients tend to show smaller pvalues. Odds ratios can be computed from coefficients of the zero-inflated portion of the model (Hilbe, 2014, p., 206). For the structural zeros, the coefficients of the linear predictor etaZR predict membership in a category, that is, a positive coefficient indicates the variable generates zeros. The coefficients of the count data linear predictor (etaN) are associated with the magnitude of the counts, that is, a positive coefficient implies the counts increase as the associated variable increases. Thus, under this approach to model development, the coefficients of the same variable in in both the zero-inflation and the count linear predictors will usually have the opposite sign (assuming independence with other variables and model convergence). One exception may occur when applying these models with data sets having too few zeros (deflation). In this case, the probability of a structural zero, π, needs to be negative (Famoye and Singh, 2006), which cannot occur with the inverse link function, so this probability will always be bounded between 0 and 1. The intercept for zero-inflation takes on a relatively large negative value (on the logit scale) on order to estimate π close to 0 while the coefficient for zero-inflation may have the same sign as the coefficient for the same variable in count portion of the model resulting in estimation problems. With zero-deflation the binary component of the model does not estimate a probability (Hilbe, 2011, p. 371).

An important aid to estimate coefficients with the NLMIXED procedure (which is especially true with zero-inflation) is to begin the computations with feasible initial parameter estimates reasonably “close” so they will converge to the maximum likelihood solution. Starting values are especially important when estimating many parameters with complex distributions. The sign and magnitude of the intercept is often the most important initial value; with estimation on the log scale, small negative or positive values for the parameters are usually reasonable. The NLMIXED procedure assigns a default value of 1 for any parameter not listed on the PARMS statement which may give a calculation error during the first iteration, even at the first observation. To diagnose this problem, it is often helpful to enter initial values and extract the relevant NLMIXED code into a DATA step where printing results will usually indicate where computational problems exist. The PARMS statement also allows grid searches; however, entering initial values from an external data set may be preferred. Parameter estimates from the zero-inflated Poisson or negative binomial distributions (relatively easy to get with SAS/STAT procedures) often provide parameter estimates close enough, especially in sign and magnitude, such that models will converge to the maximum likelihood estimates for other count data distributions. If estimation issues still exist, a modification to the initial estimate of the respective dispersion parameter may overcome the problem (esp. with the ZI NB-P).

ZERO-TRUNCATED COUNT DATA MODELS

The zero-truncated count data model is characterized by a structural absence of zeros. The zero-truncated model is a special case of the left- or lower-truncated model with cutpoint C=0. The minimum outcome in this situation is y=1. Observing at least one event is required in order to generate a count. Thus, the zero-truncated count data probability distribution has the following general form:

```
  Prob(Y=y) = f(Y=y) / [ f(Y > 0)]
            = f(Y=y) / [1- f(Y=0)]    for y = 1, 2, 3, ...
```

The PDF of the zero-truncated distribution is normalized by dividing all probabilities for y greater than zero by `(1-py0)` where `py0=f(Y=0)`. Therefore, the cumulative distribution of the zero-truncated distribution probabilities sums to 1. For zero-truncated count data, the log-likelihood equation to enter in NLMIXED has the general form:

```
lglk = LOG( f(Y=y) / (1-f(Y=0)) );
     = LOG( f(Y=y) - LOG(1-f(Y=0)) );
     = < Log-likelihood of pdf >
      - LOG(1-py0);
```

The log-likelihood equation for truncated count data is adjusted by subtracting `LOG(1-py0)` from the Log of the pdf. In the log-likelihood statement entered into NLMIXED the adjustment can be placed on the last line for clarity, which indicates it is a calculation separate from the loglikelihood equation from the standard distribution. The log-likelihood equations to be entered into NLMIXED are listed in the Appendix.

The NLMIXED procedure to run truncated count data models includes the following statements:

```
  PROC NLMIXED DATA =indat (rename=( < response >  = y ));
  WHERE y GE 1;
  PARMS < initial estimates for the coefficients of the linear predictor > ;
  etaN = < linear predictor >;
  mu = exp(etaN);  * inverse function for the log link;

  < enter log likelihood statements for a truncated probability model >

  MODEL y ~ general( lglk ) ;
  PREDICT mu  OUT=mu (keep= pred y rename=(pred=mu) ); * mean and response;
  PREDICT phi OUT=phi(keep= pred   rename=(pred=phi)); * dispersion = phi ;
  RUN;
```

If there happens to be a stray outcome of y=0 in the data set, PROC FMM automatically omits it, whereas, as shown here with PROC NLMIXED, a WHERE statement ensures that an errant 0 does not enter into the calculations. It also is one way to document a zero-truncated model is applied in the statements that follow. Initial parameter estimates for any of the zero-truncated models described in the appendix can be found with the same process as the linear predictor for the counts in zero-inflated models. The FMM procedure will estimate coefficients for the truncated Poisson or negative binomial models. The process is essentially the same as saving parameter estimates from GENMOD, except enter FMM in the PROC statement and dist=tpoisson or dist=tnegbin in the MODEL statement.

Models for zero-truncation may only be necessary with when the mean of the distribution is relatively “small” (Hilbe, 2014). For example, the mean for count data from a long-tail distribution may be large enough that the omission of zero as an outcome has little practical difference when compared with a truncated distribution. If the observed counts include a substantial number of small values without zeros yet also contains a skewed distribution of much larger values, a zero-truncated model may still be relevant. Truncated distributions are also a feature of hurdle models (another method to deal with zero-inflation) in which all zeros are structural; the count data pdf is applied only for the positive outcomes, which may have a long tail. This type of model is not illustrated here; a hurdle model has a pdf and log-likelihood equation that combines features of the zero-inflated and zero-truncated models.

Estimating a count data model in which zero is not a possible outcome with the GENMOD procedure is not the same model as a truncated Poisson or negative binomial distribution when produced with either the FMM or the NLMIXED procedures. With zero-truncation, overdispersion produces biased and inconsistent estimates of the coefficients since the mean structure changes (Long, p. 241). The zero-truncated distributions presented here offer other approaches to deal with this source of overdispersion in count data.

CONDITIONAL MEANS AND VARIANCES

This section presents general formulas for the mean and variance of an observation from either a zero inflated or a zero-truncated distribution. Computations refer to the means and variance functions of the standard count distributions shown in Table 1. The probability that y=0, py0, is also included for each distribution.

Standard Distribution	Expected Value: E(y)	Variance Function: V(y)	Reduced equation of py0=f(Y=0)
Poisson	μ	μ	$py0 = e^{-\mu}$
NB-1	μ	μ + k*μ	$py0 = (1+k)^{(-mu/k)}$
NB-2	μ	$\mu + k*\mu^2$	$py0 = (1+(k*\mu))^{(-1/k)}$
NB-P	μ	$\mu + k*\mu^P$	$pm = (1/k)*\mu^Q$ where Q = P-2 $py0 = e^{[pm*LOG(pm/(pm + \mu))]}$
UGPS	μ/(1−δ)	$\mu/(1-\delta)^3$	$py0 = e^{-\mu}$
RGPS	μ	$\mu*(1+\alpha*\mu)^2$	$py0 = e^{[-\mu /(1 + (\alpha*\mu))]}$
P-IG	μ	$\mu + \tau*\mu^2$	$py0 = e^{[(1/\tau)*(1-SQRT(1+(2*\tau*\mu)))]}$

Table 1. Means, Variances, and f(Y=0) for Standard Count Data Distributions

Formulas for the variances of zero-inflated and zero-truncated distributions printed in the literature usually omit its underlying relationship with the mean and variance function of the standard distribution. The individual components of the general formulas for the zero-inflated and zero-truncated means and variances are taken from the components of the standard distribution and the probability of zero-inflation:

```
  E(Y) = mean of standard distribution
  V(y) = variance function of standard distribution
    Π  = probability of a structural 0 (p_zr in the log-likelihood equation)
   py0 = probability of 0 from the standard count model
```

The mean and variance of zero-inflated distributions have the general form:

$$\text{mean} = (1 - \Pi) * E(y)$$

$$\text{variance} = (1 - \Pi) * [V(y) + (\Pi * E(y)^2)]$$ (eq. 4.58, Cameron/Trevidi, 2013)

When there is no zero-inflation (Π =0), both equations reduce to the mean and variance of the standard distribution.

The mean and variance of zero-truncated distributions have the general form:

$$\text{mean} = E(y) / (1 - py0)$$

$$\text{variance} = (1 / (1 - py0)) * [V(y)-(py0/(1-py0))*E(y)^2]$$ (eq. 4.29, Cameron/Trevidi, 2013)

where py0=f(Y=0) is computed as if zeros exist as an outcome. When the probability of a zero, py0, becomes very small, both equations reduce to the mean and variance of the standard distribution which includes zero.

To compute the predicted means and variances for the observations with NLMIXED, the components p_zr, py0, and mu can be saved into SAS datasets with PREDICT statements entered after the MODEL statement. The additional parameters needed to compute V(y) also need to be made available; although the dispersion parameter is estimated in the model and can be accessed from the parameter estimates table, the easiest way to extract its value is with a PREDICT statement, as illustrated in a previous section. The means and variances of the predicted values are then computed in a subsequent DATA step by merging these files with multiple SET statements.

COMPARISONS OF COUNT DATA MODELS

Information criteria computed from these zero-inflated or zero-truncated models with NLMIXED (e.g., Akaike's AIC, Schwarz's BIC), as outlined by Christensen (2018), can compare the best fit for competing models, such as a zero-inflated negative binomial with a zero-inflated Poisson. They do not provide a statistical test of the comparison, that is, they do not indicate whether one model is significantly better than another. The criterion is to select the model with the smallest value.

The zero-inflated model does not reduce to the standard count data model when the coefficients of the zero-inflated linear predictor are zero. In this case the probability of a structural zero is inflated by $\Pi=0.5$ and therefore, the models are not nested (as is necessary when comparing models with the likelihood ratio test). To compare a zero-inflated count model with the corresponding standard model (such as a zero-inflated negative binomial with a standard negative binomial) a test for non-nested models is necessary (Long, 1997). The Vuong test (1989) may provide insight if a component for zero-inflation is an appropriate addition to the count data model or if a standard count model is preferred. It tests the null hypothesis that two models fit the data equally well; they do not need to be nested.

Let dl_i be the difference between the logs of the predicted probabilities of the zero inflated model and the standard model:

$$dl_i = LOG(Pr_1(y_i \mid x_i,\ z_i,\ \beta_1,\ \Upsilon)) - LOG(Pr_2(y_i \mid x_i,\ \beta_2))$$

where i ranges from 1 to the sample size, n. The first component of the equation is the log-likelihood from the zero-inflated model (1, where x_i & z_i are the variables and β_1 & Υ and are the coefficients). The second component is the log-likelihood from the standard count data model (2, where x_i and β_2 are the variables and coefficients). The Vuong test statistic is:

$$Vuong = [\ stddev(dl) * SQRT(n)]^{-1}\ * \sum dl_i$$

where $stddev(dl)$ is the standard deviation of dl_i and n is the sample size. The Vuong test statistic is asymptotically normally distributed by the central limit theorem (Vuong, 1989). Large positive values favor the zero-inflated model; large negative values favor the standard model. A non-significant pvalue indicates no preference for either model.

Since the number of parameters estimated in a zero-inflated model is larger than the regular model (i.e., the parameters included in the zero-inflation linear predictor), the Vuong test as computed here is biased toward supporting the zero-inflation model, even when no zero-inflation exists. In this case, modifications to the computation of the Vuong statistic with corrections based on the Akaike (AIC) and Bayesian (BIC) information criteria should be made (Demaris and Harden, 2013). The adjusted values for each difference in the log-likelihoods are:

$$\text{AIC: } dl_i^c = dl_i + ((p2 - p1) / n)$$
$$\text{BIC: } dl_i^c = dl_i + ((p2 - p1) * LOG(n)/(2*n))$$

where p1 and p2 are the total number of parameters from the zero-inflated model and the standard count model, respectively, with p2 < p1. The process to compute the Vuong test from the output files from PREDICT statements in PROC NLMIXED followed by a series of DATA steps and PROC MEANS is illustrated in the Appendix.

SAS statistical procedures, including GENMOD, COUNTREG, and FMM, currently do not offer the Vuong test as an option; however, a SAS macro to compute it is available from:

http://support.sas.com/kb/42/514.html

The purpose of this macro is to compare two nested or non-nested models which are fit by maximum likelihood. Utilization of this macro has much wider applications beyond those of the count data models described in this paper. It can compare a zero-inflated Poisson or negative binomial models with their standard distributions analyzed with the GENMOD procedure by saving predicted values with an OUTPUT statement (as illustrated in an example found on the support.sas.com website). The Appendix demonstrates how to run the test with the Vuong macro by using output files produced with PREDICT statements from the NLMIXED procedure; the AIC and BIC adjustments are also included. Also note the test may be suspect if over-dispersion is present, especially with the Poisson distribution which has no dispersion parameter.

To implement these analytical techniques, the couart data set (Long, 1997) from a study examining the number of published articles evaluated with a few predictors, is available in the online documentation for PROC COUNTREG (SAS/ETS® 15.1, Example 11.2). It provides a test of the calculations for both zero-inflated (using all the data) and also zero-truncated models (the sampling frame defined as persons who have published, by restricting the number of articles published to be one or more). Also see Usage Note 43522: Fitting truncated Poisson and negative binomial models under the section "Truncated Negative Binomial Model." Examples of these distributions are illustrated in the file of SAS code that accompanies this paper or is available from the author. The dispersion parameter for zero-inflated and zero-truncated distributions can also be modeled with a linear predictor which is illustrated with NLMIXED in the usage note.

REFERENCES

Cameron, A. Trivedi, P. (2013) Regression Analysis of Count Data, 2nd ed. Cambridge University Press, New York.

Christensen, Wendy (2018). Model Selection Using Information Criteria (Made Easy in SAS®). Proceedings of the SAS Global 2018 Conference, Paper 2587-2018.

The COUNTREG Procedure, SAS/ETS® 15.1 Users Guide, accessed March 18, 2019, https://documentation.sas.com/?docsetId=etsug&docsetTarget=etsug_countreg_examples06.htm&docsetVersion=15.1&locale=en

Consul, P. (1989) Generalized Poisson Distribution. Marcel Dekker, Inc. New York and Basel.

Consul, P. and F. Famoye (1992). Generalized Poisson regression model. Communications in Statistics – Theory and Methods 21: 89-109.

Dean, C., J. F. lawless, and G. E. Willmot, (1989). A mixed Poisson-inverse Gaussian regression model. The Canadian Journal of Statistics, vol. 17. No 2, pp. 171-181.

Desmarais, Bruce and Jeffrey J. Harden (2013). Testing for zero inflation in count models: Bias correction for the Vuong test. The Stata Journal 13, Number 4, pp. 810–835.

Famoye, F. (1993) Restricted Generalized Poisson Regression Model. Commun. Statist. – Theory Meth., 22(5), 1335-1354.

Famoye, F. and K. Singh (2003) On Inflated Generalized Poisson Regression Models. Advances and Aplications in Statistics, 3(2): 145-158.

Famoye, F. and K. Singh (2006) Zero-truncated Generalized Poisson Regression Model with an Application to Domestic Violence. Journal of Data Science 4: 117-130.

Guo, J. Q, P. Trivedi (2002) Flexible Parametric Models for Long-tailed Patent Count Distributions. Oxford Bulletin of Economics and Statistics, 64, 63-82.

High, R and W. ElRayes (2017) Fitting Statistical Models with PROCs NLMIXED and MCMC. Proceedings of the SAS Global 2017 Conference, Paper 902-2017.

High, R. (2018) Alternative Variance Parameterizations in Count Data Models with the NLMIXED Procedure. Proceedings of the SAS Global 2018 Conference, Paper 2694-2018.

Hilbe, J. (2011). Negative Binomial Regression, 2nd ed. Cambridge University Press, New York.

Hilbe, J. (2014). Modeling Count Data. Cambridge University Press, New York.

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

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57: 307–333.

Zha, Liteng, Dominique Lord, and Yajie Zou. (2016) The Poisson Inverse Gaussian (PIG) Generalized Linear Regression Model for Analyzing Motor Vehicle Crash Data. Journal of Transportation Safety and Security, Vol. 8, No. 1, 18-35.

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

Robin High
Department of Biostatistics
College of Public Health
University of Nebraska Medical Center
984375 Nebraska Medical Center
Omaha, NE 68198-4375
email: rhigh@unmc.edu

APPENDIX

Log-Likelihood Equations for Zero-Inflated and Zero-Truncated Models

The log-likelihood equations printed here are derived from the probability density functions defined in the text and placed in the logarithmic form for both zero-inflated and zero-truncated models. They assume the response variable is called y, either present in the data set or with the `RENAME=(<response> = y)` option attached to the data set name. Initial estimates for the dispersion parameter are placed in the PARMS statement. Each model assumes the mean is computed from the linear predictor for the counts, `mu=EXP(etaN)`.

The log-likelihood statements are entered into the NLMIXED code directly below the statement for mu. No matter how complex the linear predictor(s) or the number of variables/coefficients entered into the equations, the log-likelihood equation does not change, and thus could be supplied with a call to a macro, if desired. In fact, with this method one can quickly compare the fit of different models with minimal edits to the NLMIXED statements.

Log-Likelihood Statements for Zero-Inflated Models

The log-likelihood equations for zero-inflated models presented here contain two variables that refer to the probability of a 0:

```
p_zr = the probability of a structural zero,
       that is, a zero generated apart from the count data model
 py0 = the probability of a zero generated from the count data model
```

The probability of a structural zero, p_zr (notated as π in the loglikelihood formula), is computed from the linear predictor for zero-inflation, etaZr, and then back-transformed to the probability of zero-inflation based on the logit link:

```
   p_zr = 1 / (1 + EXP(-etaZr));
```

Other inverse links for p_zr can also be applied by entering the appropriate back-transformation formula (such as complementary log-log or probit) as a function of the zero-inflated linear predictor, etaZr; no other adjustments to the NLMIXED code are necessary.

ZI Poisson (ZI P)

```
py0= EXP(-mu);
IF y = 0 THEN lglk = LOG(p_zr + (1-p_zr)*py0 );
         ELSE lglk = LOG(1-p_zr) + y*LOG(mu) - mu - LGAMMA(y+1);
```

ZI Quadratic Negative Binomial Distribution (ZI NB-2)

```
py0= (1+(k*mu))**(-1/k);
IF y = 0 THEN lglk = LOG(  p_zr  + (1-p_zr)*py0 );
         ELSE lglk = LOG(1-p_zr) + y*LOG(k*mu) - (y+(1/k))*LOG(1+(k*mu))
                     + lgamma(y+(1/k)) - lgamma(1/k) - lgamma(y+1);
```

ZI Linear Negative Binomial Distribution (ZI NB-1)

The log-likelihood for NB-1 is formed by replacing the dispersion parameter `k` in the NB-2 log-likelihood with `k/mu`:

```
py0 = (1+k)**(-(mu/k));
IF y = 0 THEN lglk = LOG(  p_zr  + (1-p_zr)*py0   );
         ELSE lglk = LOG(1-p_zr) + y*log(k) - (y+(mu/k))*log(1+k)
                     + lgamma(y+(mu/k)) - lgamma(mu/k) - lgamma(y+1);
```

ZI Three Parameter Negative Binomial Distribution (ZI NB-P)

The estimate for P, the exponent in the NBP variance function, is `P=2-Q`. In the log-likelihood, let `Q=2-P`:

```
 pm = (1/k)*(mu**Q);
py0 = EXP( pm*LOG( pm /(pm + mu)));
IF y = 0 THEN lglk = LOG(  p_zr  + (1-p_zr)*py0 );
         ELSE lglk = LOG(1-p_zr)
                     + ( pm*log( pm / (pm+mu))) + (y*log(1 - (pm / (pm+mu))))
                     + (lgamma(y + pm ) - lgamma(pm) - lgamma(y+1)) ;
```

ZI Unrestricted Generalized Poisson Distribution (ZI UGP)

Using the notation of Consul and others, θ is entered as the mean of the generalized Poisson distribution. For the loglikelihood equation in NLMIXED, let mu = θ:

```
py0 = EXP(-mu);
IF y = 0 THEN lglk = LOG(  p_zr  + (1-p_zr)*py0 );
         ELSE lglk = LOG(1-p_zr) + log(mu) + (y-1)*log(mu + (phi*y))
                      - (mu + (phi*y)) - lgamma(y+1);
```

ZI Restricted Generalized Poisson Distribution (ZI RGP)

```
py0 = EXP( -mu/(1 + (alpha*mu)) );
IF y = 0 THEN lglk = LOG(  p_zr  + (1-p_zr)*py0 );
         ELSE lglk = LOG(1-p_zr) + y*log(mu/(1 + (alpha*mu)))
                     + (y-1)*log(1+(alpha*y))
                     + ((-mu*(1+(alpha*y)))/(1+(alpha*mu)))
                     - lgamma(y+1);
```

ZI Poisson-Inverse Gaussian Distribution (ZI P-IG)

The calculation of individual probabilities for the P-IG distribution does not need to use an ARRAY statement to store them as previously shown (High, 2018). Because of the sequential nature of the computation, the log-likelihood is computed differently than the other distributions: the probability of the outcome y is first calculated for each observation and then its log is computed.

```
py0 = EXP( (1/tau)*(1 - SQRT(1 + (2*tau*mu))) );
IF y = 0 then py = py0;
py1 = py0 * mu * (1/SQRT(1 + (2*tau*mu))) ;
IF y EQ 1 then py = py1;
pm1 = py1; pm2 = py0; * store f(Y=1 and f(Y=0);
IF y GE 2 then
DO; DO k = 2 to y ;
      py = ((2*tau*mu/(1+ (2*tau*mu))) * (1 - (3/(2*k))) * pm1 )
          + ((mu**2)/(1 + (2*tau*mu))) * (1/(k*(k-1))) * pm2;
      pm2=pm1; pm1=py;
      END; END;
IF y = 0 THEN lglk = LOG(  p_zr  + (1-p_zr)*py0 );
         ELSE lglk = LOG(1-p_zr) + LOG(py);
```

Log-Likelihood Equations for Truncated Count Data Models

The loglikelihood equations for zero truncated models contain the probability of a 0 from the standard model:

```
  py0= the probability of a zero from the estimated parameters of the count data model
```

When omitting zero as a possible outcome, the probabilities from the distribution are divided by `1-py0` so that their sum is 1. With the log-likelihood equation, this is equivalent to subtracting LOG(1-py0).

Truncated Poisson (TP)

For zero-truncated count data, PROC FMM computes the truncated Poisson:

```
  PROC FMM DATA =indata;
  CLASS group;
  MODEL y = group / DIST=tpoisson link=log;
  TITLE 'FMM: Zero Truncated Poisson';
  RUN;
```

With zero-truncation with the Poisson distribution, the log-likelihood for PROC NLMIXED can be coded in two ways:

```
    py0 = EXP(-mu);
   lglk = y*LOG(mu) - mu  - lgamma(y+1)
          - LOG(1 - py0);   * subtract LOG(1 - f(y EQ 0));

   can also apply the LOGSDF function;
   lglk = y*LOG(mu) - mu  - lgamma(y+1)
         - LOGSDF('Poisson', 0, mu);  * subtract LOG(PR(y GE 1)) ;
```

LOGSDF is the LOG survival function which computes f(Y > 0) for the Poisson distribution; since the survival function estimates the cumulative probability greater than the value given, the greater than sign (>) gives the results needed for (Y GE 1). For truncated distributions, the final line of the `lglk` statement subtracts of LOG(1-py0), the log of the probability that (y > 0).

Truncated Quadratic Negative Binomial (TNB-2)

PROC FMM has the zero-truncated negative binomial (NB2) distribution invoked with a MODEL statement option.

```
PROC FMM DATA=indata;
CLASS group;
MODEL y = group / dist=tnegbin link=log;
run;
```

For PROC NLMIXED with truncation of y=0, the log-likelihood of the NB-2 distribution is coded:

```
py0 = (1 + (k*mu))**(-1/k);
lglk = (y*log(k*mu) - (y+(1/k))*log(1+(k*mu))
       + lgamma(y+(1/k)) - lgamma(1/k) - lgamma(y+1)  )
       - log(1 - py0);
```

Truncated Linear Negative Binomial Distribution (TNB-1)

```
 py0 = (1+k)**(-mu/k);
lglk = (y*log(k) - (y+(mu/k))*log(1+k)
        + lgamma(y+(mu/k)) - lgamma(mu/k) - lgamma(y+1) )
        - LOG(1-py0) ;
```

Truncated Three Parameter Negative Binomial Distribution (TNB-P)

```
pm = (1/k)*(mu**Q);
py0 = EXP( pm*LOG( pm /(pm + mu)));
lglk = ( pm *log( pm / ( pm + mu))) + (y*log(1 - ( pm / ( pm + mu))))
        + (lgamma(y + pm) - lgamma(pm) - lgamma(y+1) )
        - LOG( 1 - py0);
```

The exponent in the NBP variance function, is P = 2-Q. To compute P, after the MODEL statement enter:

```
ESTIMATE 'P=' 2-Q;
```

Truncated Unrestricted Generalized Poisson Distribution (TGP)

```
py0 = EXP(-mu) ;
lglk = ( LOG(mu) + (y-1)*log(mu + (phi*y)) - (mu + (phi*y)) - lgamma(y+1) )
        - LOG(1 - py0);
```

Truncated Restricted Generalized Poisson Distribution (TRGP)

```
 py0 = EXP( -mu/(1+(alpha*mu)) );
lglk = (y*log(mu/(1 + (alpha*mu))) + (y-1)*log(1+(alpha*y)) + ((-mu*(1+(alpha*y)))
        / (1 + (alpha*mu))) - lgamma(y+1) )
        - LOG(1 - py0);
```

Truncated Poisson-Inverse Gaussian Distribution (TP-IG)

The truncated Poisson-Inverse Gaussian distribution may not be applied as often as other distributions; however, it occurs with hurdle models (another way to work with count data having zero-inflation) and also when the counts are extremely skewed with a substantial number of the smallest values equal or close to 1, yet no zeros in the data set.

The probability y=0 for the P-IG model is:

```
py0 = EXP( (1/tau)*(1 - SQRT(1 + (2*tau*mu))) );
```

The individual probabilities for y = 1, 2, 3, … denoted as py are computed with the same NLMIXED code as for the zero-inflated model before the zero-inflation computation with the following adjustment at the end:

```
pTy = py / (1-py0);  * divide py by (1 - f(y=0));
lglk = log(pTy);
```

INITIAL PARAMETER ESTIMATES WITH ZERO-INFLATION

To enter initial estimates from a data set, a necessary portion of the task is to enter parameter names that match the coefficient names which are placed on the zero inflation and count linear predictor equations (etaZr and etaN in the NLMIXED code). For the linear predictor of the counts, one option is to enter coefficient names that match the variable names with an underscore _ attached at the end (to avoid any unlikely conflict with internal variables in NLMIXED that begin with the underscore) such that the linear predictor looks like:

```
etaN = intercept_ + gender_*(gender=”F”) + age_*age ;
```

Since the same input variables can appear in both linear predictors, etaZr must have different coefficient names than etaN for the same explanatory variable. In this case the coefficients for etaZr can be given the variable name with _P placed at the end:

```
etaZr = intercept_P + gender_P*(gender=”F”) + age_P*age ;
```

With this coefficient naming convention, initial estimates of the parameters for the linear predictors of the components for zero-inflation and counts may be obtained with the GENMOD procedure with either a zero-inflated Poisson or a zero-inflated negative binomial model:

```
ODS OUTPUT parameterestimates=gnmdprms(KEEP= parameter estimate
                                      WHERE=(parameter NE 'Scale'))
           ZeroParameterEstimates=zprms(KEEP=parameter estimate);
PROC GENMOD DATA=indata ;
MODEL y = gender age / dist = zip;    * do not enter dist=poisson here;
ZEROMODEL gender age / link = logit;
RUN;
```

```
DATA zprms;
LENGTH parameter $25;
SET zprms(rename=(parameter=zprm));
DROP zprm;
parameter = LOWCASE(CATT(zprm,'_P')); * add _P at the end of the variable name;
RUN;

DATA gnmdprms;
LENGTH parameter $25;
SET gnmdprms(rename=(parameter=prm));
parameter = LOWCASE(CATT(prm,'_'));  * add _ at the end of the variable name;
RUN;

DATA initprms;
SET  zprms gnmdprms;
RUN;

PROC PRINT DATA= initprms NOobs n;
VAR parameter estimate;
TITLE 'Initial Parameter Estimates for NLMIXED';
RUN;

PROC NLMIXED DATA = dsn(rename=( <response> = y) ) ;
PARMS phi .1 / DATA= initprms ; * phi represents the actual dispersion parameter;
< enter the NLMIXED code > ;
RUN;
```

Estimates of zero-inflation coefficients can also be obtained with the FMM procedure. To produce coefficients with signs that match GENMOD, a reverse ordering of the MODEL statement for the zero-inflation component (i.e., to estimate structural zeros) with the MODEL statement for the count data coefficients:

```
ODS OUTPUT parameterestimates=FMMprms(KEEP=parameter estimate)
                  MixingProbs=zprm(KEEP=parameter estimate);

PROC FMM DATA= indata ;
MODEL      +   / dist=constant;
MODEL y = gender age / dist= poisson;
PROBMODEL gender age;
TITLE 'FMM: zi poisson coefficients';
RUN;
```

Processing the two output files of the coefficients to produce the file of initial parameter estimates for NLMIXED proceeds in the same manner as described above.

When parameters for the count data distribution are not provided an initial estimate (such as the dispersion parameter), they can be initialized with a value entered on the PARMS statement as indicated above, such as the respective dispersion parameter indicated here with phi; otherwise, they will be given the default value of 1. If the error message “no valid parameter points found” appears in the LOG window, it may be resolved with an adjustment to the initial estimate of the dispersion parameter.

VUONG TEST WITH NLMIXED

After running the standard and zero-inflated models with NLMIXED, the Vuong test helps to decide which of the two models is preferred. The data processing to compute the Vuong test begins with the zero-inflated distribution (Model 1) by adding this PREDICT statement after the MODEL statement in NLMIXED:

```
PREDICT lglk OUT=LL_zi(keep=pred rename=(pred=LL_zi));
```

A similar statement is added to a separate set of NLMIXED statements for the standard distribution (Model 2):

```
PREDICT lglk OUT=LL_c (keep=pred rename=(pred=LL_c));
```

Since these two count data models must be derived from the same input data set (assuming identical omission of observations due to missing data), the output files will match row by row, and can be merged with two SET statements:

```
p1= 7     Parameters in the zero-inflated model (etaZR, etaN, and dispersion)
p2= 3     Parameters in the standard model (etaN and dispersion)
n = 3874  Number of observations in dataset

DATA LL_diff;
SET LL_zi; SET LL_c ;      * merge with two SET statements;
dl     =  ll_zi - ll_c ;   * model 1 Log-likelihood – model 2 Log-likelihood;
dl_AIC = dl + ( (3-7)/ 3874 );  * 3-7 is p2-p1 ;
dl_BIC = dl + (((3-7)*LOG(3874))/(2*3874) );
             * see p. 814 "Testing for Zero Inflation" Desmarais/Harden (2013);
keep ll_zi ll_c dl: ;
RUN;
PROC PRINT DATA=LL_diff(obs=10);
RUN;

PROC MEANS DATA=ll_diff vardef=n noprint;
VAR dl dl_AIC dl_BIC;
OUTPUT OUT=vuong_stats(drop=_type_ RENAME=(_freq_=n))
       sum=  SM_dl  SM_dl_AIC  SM_dl_BIC
       std= std_dl std_dl_AIC std_dl_BIC;
RUN;

DATA vuong_stats;
SET vuong_stats;
Vuong     = ( 1/(std_dl * sqrt(n))) * SM_dl;
  pvalue    = 2*(1-probnorm(ABS(vuong)));
Vuong_aic = ( 1/(std_dl_aic * sqrt(n))) * SM_dl_aic;
  pvalue_a  = 2*(1-probnorm(ABS(vuong_aic)));
Vuong_bic = ( 1/(std_dl_bic * sqrt(n))) * SM_dl_bic;
  pvalue_b  = 2*(1-probnorm(ABS(vuong_bic)));
RUN;
```

If the test statistic is relatively large and positive, the data suggest model 1 (the zero-inflated model) is considered the preferred model. If the test statistic is relatively large and negative, the data suggest model 2 (the standard model) is the preferred model. A test statistic arbitrarily close to 0 is inconclusive. If the dispersion is not adequately modeled, the results of the Vuong test may indicate the zero-inflated model is preferred, even when structural zeros are not present. This is of particular concern for the Poisson distribution which has no dispersion parameter. Further evaluation is needed for other count data distributions if over-dispersion remains after running the standard and zero-inflated models.

The Vuong macro can compare any of the zero-inflated models presented here with their respective standard model, such as a zero-inflated P-IG (Model 1) with the standard P-IG (Model 2). To do so, the probabilities of structural zeros and the predicted probabilities are saved from the zero-inflated model; enter these two statements into the NLMIXED code following the MODEL statement:

```
   PREDICT p_zr      OUT=pzr   (keep= pred RENAME=(pred = p_zr));
   PREDICT EXP(lglk) out=prb_zi(keep= pred RENAME=(pred = prbzi));
```

For the standard P-IG model, enter this statement into NLMIXED for the predicted probabilities:

```
   PREDICT EXP(lglk) out=prb_c(keep=y pred RENAME=(pred=prbc));
```

Merge the three output files with SET statements:

```
   DATA prd;
   SET prb_zi; SET pzr; SET prb_c;
   RUN;
```

For the Vuong macro, the choices for the two distributions, dist1= and dist2= , must be selected from:

```
   NOR, BIN, MULT, GAM, IG, NB, POI, ZIP, ZINB, OTH
```

When a zero-inflated model is compared with its standard model, if the distribution option does not exist in the macro (as the case for both ZIP-IG and P-IG), then dist1=OTH and dist2=OTH are entered. The computed probabilities from the two models are the inputs, along with the probability of a structural zero from the zero-inflated model:

```
   %vuong(data=prd, response=y, test=Vuong,
          model1=ZIPIG, p1=prbzi, dist1=OTH, nparm1=7, pzero1=p_zr,
          model2=PIG,   p2=prbc,  dist2=OTH, nparm2=3)

                                                 Preferred
   Vuong Statistic            Z     Pr>|Z|       Model

   Unadjusted             2.367     0.0180       ZIPIG
   Akaike Adjusted        1.496     0.1348       ZIPIG
   Schwarz Adjusted      -1.109     0.2674       PIG
```

In the macro output, if Z > 0 then ZIPIG (Model 1) is listed as the preferred model. If Z < 0 then PIG (Model 2) is listed as the preferred model. Notice also how the Unadjusted pvalue prefers Model 1 (p=0.018) while the pvalues for “Adjusted” indicate neither model is preferred.

Comparing two zero-inflated models with the Vuong macro proceeds in a similar manner, (e.g, ziUGP vs ziRGP) although in this situation the number of parameters does not need to be entered, since they are the same for both models:

```
   %vuong(data=prd, response=y, test=Vuong,
          model1=ziUGP, p1=prb_zigp,  dist1=OTH, pzero1=pzr_zigp,
          model2=ziRGP, p2=prb_zirgp, dist2=OTH, pzero2=pzr_zirgp);
```

where the variables to enter for the p1= and p2= options are the predicted probabilities from the model and the pzero1= and pzero2= options are the probabilities of a structural zero from each distribution.