Paper 355-2011

Bivariate Count Data Regression Models – A SAS® Macro Program

Nan-Ting Chou, University of Louisville, Louisville, KY
David Steenhard, LexisNexis, Dayton, OH

ABSTRACT

Bivariate count models are used in situations where two count dependent variables are correlated and they need to be jointly estimated. Most research on count data regression models focus on univariate cases where the single dependent variable takes on non-negative integer. While Chou and Steenhard (2009) provided a SAS® macro program that handles a wide variety of univariate count data distributions, no study has provided a SAS program to estimate bivariate count regression models. This paper develops a SAS® macro program jointly estimate two correlated count data series---the bivariate count regression model. Our SAS® macro allows for ten bivariate count data distributions: bivariate Poisson, bivariate Poisson-LogNormal, bivariate negative binomial, bivariate generalized Poisson, bivariate Poisson inverse Gaussian, bivariate Borel-Tanner, bivariate Neyman Type A, bivariate generalized Waring, Poisson-Laguerre polynomial, and bivariate Poisson series expansion. In addition, the macro also provides estimation of the above bivariate distributions using a copula approach. Bivariate zero-inflated, bivariate hurdle, bivariate truncated, and bivariate censored regression models can also be estimated with this SAS® macro. We apply this SAS® macro procedure to a healthcare utilization data. The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) of these bivariate regression models are provided for model evaluation.

1. INTRODUCTION

Bivariate count data regression models are used when the event counts are jointly dependent; while the univariate count regression models estimate a single event count data (Chou and Steenhard, 2009). Applying two independent count regressions to paired joint event counts leads to inconsistent and inefficient estimators. Paired count events exhibiting correlation should be estimated jointly; and the bivariate count regression models are designed to handle such cases. Bivariate count regression models analyze correlated event counts such as: the number of doctor visit and non-doctor professional visit; the number of visits to general practitioners and visits to specialists; the number of insurance claims with and without bodily injuries; the number of voluntary and involuntary job changes. Studies that apply bivariate count models often assume a bivariate Poisson distribution which assumes the conditional mean of each count variable equals the conditional variance. Most of the paired events do not exhibit Poisson distribution. For the common case of overdispersed count data, the bivariate negative binomial, bivariate Poisson inverse Gaussian, or the Poisson-lognormal models are potentially useful. Another shortcoming of commonly used bivariate count models is they can only accommodate non-negative correlation between the paired counts. The bivariate Poisson-lognormal, bivariate Poisson-Laguerre polynomials and the bivariate Poisson-series expansion models can handle distributions that are more generalized than the bivariate Poisson or bivariate negative binomial models. Cameron and Trivedi (1998) provide an overview of standard bivariate count models. Our paper develops a SAS macro regression that is capable of handling various types of bivariate count data distributions.

There are studies apply bivariate count regression models to analyze correlated count events. Gurmu and Elder (2000) proposes a bivariate Poisson-Laguerre model to analyze an Australian healthcare utilization data: the number of visit with doctor and the number of visit with non-doctor health professionals. Wang (2003) examines the same dataset using a bivariate zero-inflated negative binomial model to account for excessive zeros. However, the bivariate model proposed by Wang (2003) restricts the correlation between the two count variables to be non-negative. Gurmu and Elder (2008) further develops a bivariate zero-inflated Poisson-Laguerre count regression model with an unrestricted correlation pattern to analyze the same data. Cameron et. al. (2004) uses copula functions to obtain a flexible bivariate parametric model. They applies this model to healthcare utilization data: the self-reported number of doctor visits and true number of doctor visits. Atella and Deb (2008) apply a bivariate negative-binomial regression model to examine the relationships between the number of visits to general practitioners and specialists using data from Italy. Riphahn et. al. (2003) applies a bivariate random effect count regression to test the adverse selection and moral hazard problems in German healthcare utilization. Mayer and Chappell (1992) use a bivariate Poisson model to examine the factors affect U.S. industries' entry and exit. Morata (2009) applies a bivariate Poisson distribution model to analyze automobile insurance ratemaking. Ho et. al. (2009) uses a bivariate zero-inflated negative binomial model to jointly examine two types of Japanese merger: M&A (merger and acquisition), and FDI (foreign direct investment) into the United States. Wang et. al. (2003) uses a bivariate zero-inflated Poisson regression model to analyze two types of occupational injuries. Lee et. al. (2005) further develops a bivariate zero-inflated Poisson autoregression model to analyze the same occupational injuries data.

This paper develops a versatile SAS® macro to model ten different bivariate discrete distributions, this SAS® macro also includes bivariate zero-inflated, hurdle, truncated and censored models. By using the PROC NLMIXED procedure, one can jointly estimate bivariate count regression models through the log likelihood functions. We demonstrate the versatility of this macro by applying it to a healthcare utilization data. We further evaluate the performance of the different bivariate count regression models using AIC and BIC.

Section 2 briefly discusses five commonly used bivariate count data distributions and their respective log likelihood functions. It also briefly discusses zero-inflated, hurdle, truncated and censored models. Section 3 explains the capabilities of the SAS® macro-- %bicount. Section 4 applies our SAS® macro to a healthcare utilization data set where we further discuss and evaluate the results of these bivariate count regression models. Summary and concluding remarks are presented in section 5.

2. BIVARIATE COUNT MODELING DISTRIBUTIONS AND THEIR PROPERTIES

In this section we discuss five bivariate count data distributions and their respective log likelihood functions. The five data distributions are: bivariate Poisson, bivariate negative binomial, bivariate generalized Poisson, bivariate Poisson inverse Gaussian and the Poisson-Laguerre polynomial distribution. The bivariate zero-inflated, hurdle, truncated, and censored models are also briefly discussed. Other data distributions such as: bivariate Poisson-lognormal, bivariate Borel-Tanner, bivariate Neyman type A, bivariate generalized Waring, and the Poisson series expansion, are modeled into our SAS®macro procedure; but not discussed here due to the space limitation.

2.1 BIVARIATE POISSON MODEL

Among the bivariate count models, the bivariate Poisson regression model is the most widely used model. A well established approach is to generate the bivariate Poisson distribution by convolutions of Poisson random variables (Kocherladota and Kocherlakota 1992) is:

$y_{1i} = y_{1i}^{*} + \mu_i$

$y_{2i} = y_{2i}^{*} + \mu_i$

Where $y_i^{*} \sim Poisson(\theta_{1i}), y_{2i}^{*} \sim Poisson(\theta_{2i}), and\ \mu_i \sim Poisson(\theta_3)$ are independently distributed. The joint probability density function of the bivariate Poisson can be obtained as:

$$f(y_{1i}, y_{2i}|x_i) = \left[\prod_{j=1}^{2} \frac{exp(-\theta_{ji})\theta_{ji}^{y_{ji}}}{y_{ji}!}\right] exp(-\theta_3) \sum_{s=0}^{m} \binom{y_{1i}}{s}\binom{y_{2i}}{s} s! \left(\frac{\theta_3}{\theta_{1i}\theta_{2i}}\right)^{s}$$

Where $m = \min(y_{1i}, y_{2i})$ and $\theta_{ji} = \exp(x_{ji}\beta)$. this model allows only for non-negative correlation between the counts and restricts the mean to be equal to the variance for each of the respective marginal distributions. The marginal distributions of the model are still Poisson, and the correlation between the two count variables (conditioned on the covariates) is individual specific, being a function of the $\theta_{ji}\ and\ \theta_3$

$corr(y_{1i}, y_{2i}) = \theta_3 / \sqrt{(\theta_{1i})(\theta_{2i})}$

The log likelihood function of a bivariate Poisson model is derived by taking the log of the density function specified above.

2.2 BIVARIATE NEGATIVE BINOMIAL MODEL

As in the univariate case, bivariate count models can be generalized to allow for overdispersion. Consider a bivariate model with unobserved heterogeneity. Let $(y_{ji}|x_{ji}, v_i) \sim Poisson(\theta_{ji} v_i)$, where v_i is the unobserved heterogeneity component. The mixture bivariate density has the form:

$$f(y_{1i}, y_{2i}|x_i) = \int \left[\prod_{j=1}^{2} \frac{exp(-\theta_{ji} v_i)(\theta_{ji} v_i)^{y_{ji}}}{y_{ji}!}\right] g(v_i)\, dv_i$$

Where x_i is a vector of all regressors. If v_i has a gamma distribution with mean unity and variance $\frac{1}{\alpha}$ then you have the bivariate negative binomial regression model with a joint probability density function

$$f(y_{1i}, y_{2i}|x_i) = \left[\prod_{j=1}^{2} \frac{\theta_{ji}^{y_{ji}}}{y_{ji}!}\right] \frac{\Gamma(y_{.i} + \alpha)}{\Gamma(\alpha)} \alpha^{-y_{.i}} \left(1 + \frac{\theta_{.i}}{\alpha}\right)^{-(\alpha + y_{.i})}$$

Where $\theta_{.i} = \theta_{1i} + \theta_{2i}$ and $y_{.i} = y_{1i} + y_{2i}$

The marginal distributions of this model are still negative binomial, and the correlation between the two count variables (conditional to the covariates) is individual specific,

$$Corr(y_{1i}, y_{2i}|x_i) = \frac{\theta_{1i}\theta_{2i}}{\alpha} \Big/ \sqrt{(\theta_{1i} + \theta_{1i}^2/\alpha)(\theta_{2i} + \theta_{2i}^2/\alpha)}$$

2.3 BIVARIATE GENERALIZED POISSON MODEL

This distribution is derived from the generalized Poisson distribution using a trivariate reduction method (Vernic, 1997):

$y_{1i} = y_{1i}^{*} + \mu_i$

$y_{2i} = y_{2i}^{*} + \mu_i$

Where $y_{1i}^{*} \sim GPoisson(\theta_{1i}, \lambda_1)$, $y_{2i}^{*} \sim GPoisson(\theta_{2i}, \lambda_2)$, $and\ \mu_i \sim Poisson(\theta_3, \lambda_3)$ are independently distributed. The joint probability density function of the bivariate generalized Poisson can be obtained as:

$$f(y_{1i}, y_{2i}|x_i) = \theta_{1i}\theta_{2i}\theta_{3i}\, exp(-\theta_{.i} - y_{1i}\lambda_1 - y_{2i}\lambda_2)$$
$$\sum_{k=0}^{m} \frac{1}{(y_{1i}\text{-}k)!(y_{2i}\text{-}k)!}(\theta_{1i}+(y_{1i}\text{-}k)\lambda_1)^{y_{1i}\text{-}k-1}(\theta_{2i}+(y_{2i}\text{-}k)\lambda_2)^{y_{2i}\text{-}k-1}(\theta_3+k\lambda_3)^{k-1}\exp\big(k(\lambda_1+\lambda_2\text{-}\lambda_3)\big)$$

Where $m = \min(y_{1i}, y_{2i})$ and $\theta_{ji} = \exp(x_{ji}\beta)$. This model allows only for non-negative correlation between the counts. The marginal distributions of the model are still generalized Poisson, and the correlation between the two count variables (conditioned on the covariates) is individual specific, being a function of the $\theta_{ji}\ and\ \theta_3$:

$$Corr(y_{1i}, y_{2i}|x_i) = \frac{\theta_3 M_3^2}{[(\theta_{1i}M_{1i}^3 + \theta_3 M_3^3)(\theta_{2i}M_{2i}^3 + \theta_3 M_3^3)]^{1/2}}$$

$$M_{ji} = \frac{1}{(1-\theta_{ji})}$$

2.4 BIVARIATE POISSON INVERSE GAUSSIAN

Several forms of the bivariate distribution can be developed by compounding the bivariate Poisson distribution with the inverse Gaussian distribution of the form discussed by Jorgensen (1987). For this paper we use the case of the inverse Gaussian distribution when $\gamma = -1/2$ (Brown et al. 2006). The joint probability density function of the bivariate Poisson inverse Gaussian can be obtained as:

$$f(y_{1i}, y_{2i}|x_i) = \frac{\theta_{1i}^{\,y_{1i}}}{y_{1i}!}\frac{\theta_{2i}^{\,y_{2i}}}{y_{2i}!}\left[\frac{\Psi^2}{\Psi^2 + 2\theta_{.i}}\right]^{\frac{y_{.i}}{2}} exp\left(\Psi^2 - \Psi\sqrt{(\Psi^2 + 2\theta_{.i})}\right) K_{y_{.i-1/2}}\left(\Psi\sqrt{(\Psi^2 + 2\theta_{.i})}\right)$$

Where $\theta_{.i} = \theta_{1i} + \theta_{2i}$ and $y_{.i} = y_{1i} + y_{2i}$ and $K_{n-1/2}(z)$ is the modified Bessel function of the second kind. This model allows only for non-negative correlation between the counts. The correlation between the two count variables (conditioned on the covariates) is individual specific, being a function of the $\theta_{ji}\ and\ \Psi$:

$$Corr(y_{1i}, y_{2i}|x_i) = \frac{\theta_{1i}\theta_{2i}}{\Psi^2} \Big/ \sqrt{(\theta_{1i} + \theta_{1i}^2/\Psi^2)(\theta_{2i} + \theta_{2i}^2/\Psi^2)}$$

2.5 BIVARIATE POISSON-LAGUERRE POLYNOMIAL MODEL

A major shortcoming of the commonly used multivariate models (Poisson, negative binomial, generalized Poisson, and Poisson inverse Gaussian) is that they do not allow for negative correlations between the count variables. Gurmu and Elder (2000) proposes semiparametric estimation models in which dependence between count variables is introduced through correlated unobserved heterogeneity components. These models allow for both positive and negative correlation between the two count variables.

Gurmu and Elder (2000) modeled the dependence between y_{1i} and y_{2i} by means of correlated unobserved heterogeneity components v_1 and v_2. Each of the components is associated with only one of the event counts. Suppose $(y_{ji}|x_{ji}, v_{ji}) \sim Poisson(\theta_{ji}\, v_{ji})$ with (v_{1i}, v_{2i}) having a bivariate distribution $g(v_{1i}, v_{2i})$. The mixture density can be expressed as:

$$f(y_{1i}, y_{2i}|x_i) = \iint \left[\prod_{j=1}^{2} \frac{exp(-\theta_{ji}v_{ji})(\theta_{ji}v_{ji})^{y_{ji}}}{y_{ji}!}\right] g(v_{1i}, v_{2i})\, dv_{1i}\, dv_{2i} \qquad (1)$$

Let $M(-\theta_{1i}, -\theta_{2i}) = E_v\left[exp(-\theta_{1i}v_{1i,} - \theta_{2i}v_{2i})\right]$ denote the bivariate moment generating function of (v_{1i}, v_{2i}) evaluated at (y_{1i}, y_{2i}) then (1) takes the form of:

$$f(y_{1i}, y_{2i}|x_i) = \left[\prod_{j=1}^{2} \frac{(\theta_{ji})^{y_{ji}}}{y_{ji}!}\right] M^{(y_1, y_2)}(-\theta_{1i}, -\theta_{2i}) \qquad (2)$$

Where $M^{(y_1, y_2)}$ is the derivative of $M(-\theta_{1i}, -\theta_{2i})$ of order $y_{.} = y_1 + y_2$. Correlation based on the mixture model (2) can be positive or negative.

The form of the density (2) depends on the choice of the distribution of the unobservable $g(v_{1i}, v_{2i})$. If $g(v_{1i}, v_{2i})$ follows a bivariate lognormal distribution, you get the bivariate Poisson-lognormal distribution. If $g(v_{1i}, v_{2i})$ is approximated by Laguerre polynomial of order one, we obtain the bivariate Poisson-Laguerre polynomial density given by:

$$f(y_{1i}, y_{2i}|x_i) = \left[\prod_{j=1}^{2} \frac{(\theta_{ji})^{y_{ji}}}{y_{ji}!}\right] M^{(y_1,y_2)}(-\theta_{1i}, -\theta_{2i})$$

Where

$$M^{(y_1,y_2)}(-\theta_{1i}, -\theta_{2i}) = \left[\prod_{j=1}^{2} \frac{\Gamma(y_{ji} + \alpha_j)}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j} (\lambda_j + \theta_{ji})^{-(\alpha_j + y_{ji})}\right] \psi_i$$

With

$$\lambda_j = \frac{1}{1 + \rho_{11}^2}[\alpha_j + \rho_{11}^2(\alpha_j + 2)]$$

and

$$\psi_i = \frac{1}{1 + \rho_{11}^2}[1 + 2\rho_{11}\sqrt{\alpha_1\alpha_2}(1 - \eta_{1i})(1 - \eta_{2i}) + \rho_{11}^2\alpha_1\alpha_2(1 - 2\eta_{1i} + \eta_{1i}\xi_{1i})(1 - 2\eta_{2i} + \eta_{2i}\xi_{2i})]$$

$\eta_{ji} = \frac{y_{ji}+\alpha_j}{\alpha_j}\left(1 + \frac{\theta_{ji}}{\lambda_j}\right)^{-1}$ and $\xi_{ji} = \frac{y_{ji}+1+\alpha_j}{\alpha_j}\left(1 + \frac{\theta_{ji}}{\lambda_j}\right)^{-1}$. Unlike the bivariate Poisson-lognormal distribution, the Poisson-Laguerre polynomial model has a closed form, and can be easily implemented within the likelihood framework. The correlation between the two count variables (conditional to the covariates) is:

$$Corr(y_{1i}, y_{2i}|x_i) = \frac{\theta_{ii}\theta_{2i}[M_a^{(1,1)}(0,0) - 1]}{\sqrt{[\theta_{1i} + \theta_{1i}^2(M_a^{(2,0)}(0,0) - 1)]\,[\theta_{2i} + \theta_{2i}^2(M_a^{(0,2)}(0,0) - 1)]}}$$

Where

$$M_a^{(1,1)}(0,0) = [\alpha_1\alpha_2 + 2\rho_{11}\sqrt{\alpha_1\alpha_2} + \rho_{11}^2(\alpha_1 + 2)(\alpha_2 + 2)]/\lambda_1\lambda_2$$

$$M_a^{(2,0)}(0.0) = \frac{(\alpha_1 + 1)[\alpha_1 + \rho_{11}^2(\alpha_1 + 6)]}{\lambda_1(1 + \rho_{11}^2)}$$

and for $M_a^{(0,2)}(0.0)$ just replace α_1 and λ_1 in the preceding equation by α_2 and λ_2 respectively. The conditional correlation can take on zero, positive or negative values.

2.6 BIVARIATE MODELS USING COPULA

Existing techniques of estimating joint distributions of nonlinear outcomes are very computationally demanding. Interest in copula approach arises from several reasons. First, one often possess more information about marginal distributions of related variables than their joint distribution. The copula approach is a useful method for deriving joint distributions given the marginal distributions, especially when the variables are nonnormal. Copulas are functions that connect multivariate distributions to their one-dimensional margins (Trivedi and Zimmer 2007). If F is an m-dimensional cumulative distribution function (cdf) with one-dimensional margins F_1,…, F_m then there exists an m-dimensional copula C such that $F(y_1, \ldots, y_m) = C(F_1(y_1, \ldots, F_m(y_m); \varphi)$ where φ is a parameter of the copula called the dependence parameter, which measures dependence between the marginals. Some common copulas used in modeling include:

Farlie-Gumbel-Morgenstern Copula	$C(u_1, u_2; \varphi) = u_1u_2(1 + \varphi(1 - u_1)(1 - u_2))$
Gaussian (Normal) Copula	$C(u_1, u_2; \varphi) = \Phi(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \varphi)$
Clayton Copula	$C(u_1, u_2; \varphi) = (u_1^{-\varphi} + u_2^{-\varphi} - 1)^{-1/\varphi}$
Frank Copula	$C(u_1, u_2; \varphi) = -\varphi^{-1}log\left\{1 + \frac{(e^{-\varphi u_1} - 1)(e^{-\varphi u_2} - 1)}{e^{-\varphi} - 1}\right\}$

We use Frank copula in our SAS® macro %bicount for the following reasons. First, unlike some other copulas, it permits negative dependence between the marginals. Second, dependence is symmetric in both tails, similar to the Gaussian copula. The bivariate joint distribution function using the Frank copula is:

$$C(F(y_1, y_2); \varphi) = -\varphi^{-1}log\left\{1 + \frac{(e^{-\varphi F_1(y_1)} - 1)(e^{-\varphi F_1(y_1)} - 1)}{e^{-\varphi} - 1}\right\}$$

The bivariate joint probability function corresponding to $C[F_1(y_1), F_2(y_2)]$ is obtained iteratively, as follows:

$$P(Y_1 = 0, Y_2 = 0) = C[F_1(0), F_2(0)]$$
$$P(Y_1 = y_1, Y_2 = 0) = C[F_1(y_1), F_2(0)] - C[F_1(y_1 - 1), F_2(0)] \qquad y_1 = 1,2,\dots,N$$
$$P(Y_1 = 0, Y_2 = y_2) = C[F_1(0), F_2(y_2)] - C[F_1(0), F_2(y_2 - 1)] \qquad y_2 = 1,2,\dots,N$$
$$P(Y_1 = y_1, Y_2 = y_2) = C[F_1(y_1), F_2(y_2)] - C[F_1(y_1 - 1), F_2(y_2)] - C[F_1(y_1), F_2(y_2 - 1)] + C[F_1(y_1 - 1), F_2(y_2 - 1)] \quad y_1, y_2 = 1,2,\dots,N$$

2.7 BIVARIATE HURDLE AND ZERO-INFLATED MODEL

Zero-modified count models are used when the observed data displays a high frequency of the zero-zero state, $(y_1 = 0, y_2 = 0)$. There are two ways of handling this situation. First, the bivariate hurdle model (Mullahy 1986) or two-part model where the first part is a binary outcome model (logit or probit) and the second part is a bivariate truncated count model. Such partition permits the interpretation that positive observations arise from crossing the zero-zero hurdle or threshold. The bivariate hurdle model is appealing because it reflects a two-part decision-making process. The probability density function of the bivariate hurdle model is given by:

$$h(y_{1i}, y_{2i}|x_i) = \begin{cases} \pi_i & y_{1i} = 0, y_{2i} = 0 \\ (1 - \pi_i)\frac{f(y_{1i}, y_{2i}|x_i)}{1 - f(y_{1i}=0, y_{2i}=0)} & y_{1i} > 0, y_{2i} > 0 \end{cases}$$

Where $\pi_i = \Pr(y_{1i} = 0, y_{2i} = 0)$ is the cumulative density function (CDF) of the logit or probit regression selection model and $\frac{f(y_{1i}, y_{2i}|x_i)}{1 - f(y_{1i}=0, y_{2i}=0)}$ is the probability density function of a bivariate truncated count regression model. Based on the above probability density function, we can derive the log likelihood function of bivariate hurdle model.

Another way to model excess zeros in the count data is the bivariate zero-inflated count models (Lambert 1992). A bivariate zero-inflated model is a special case of a finite mixture model. Bivariate zero-inflated model assumes that the zero counts come from two sources not one source as in the bivariate hurdle model. A logit or probit model is used to determine the probability of counts being the zero-zero state. The bivariate zero-inflated probability density function is given by

$$h(y_{1i}, y_{2i}|x_i) = \begin{cases} \pi_i + (1 - \pi_i) f(y_{1i} = 0, y_{2i} = 0) & y_{1i} = 0, y_{2i} = 0 \\ (1 - \pi_i) f(y_{1i}, y_{2i}|x_i) & y_{1i} > 0, y_{2i} > 0 \end{cases}$$

Where $\pi_i = \Pr(y_{1i} = 0, y_{2i} = 0)$ is the CDF of the logit or probit regression, and $f(y_{1i}, y_{2i}|x_i)$ is the density function. The log likelihood function of a bivariate zero-inflated count regression model is derived based on the above density function.

2.8 BIVARIATE TRUNCATED MODEL

The bivariate truncated models are used if the observations $(y_{1i}, \boldsymbol{x}_{1i})$ or $(y_{2i}, \boldsymbol{x}_{2i})$ or both in some range are totally lost and the joint distribution of observed counts is restricted. A series may be truncated from below (left truncated) or truncated from above (right truncated). The most popular truncated model is the zero-truncated model (Gurmu and Elder, 2008), where the zero class is missing for both dependent variables. The zero truncated bivariate distribution takes the form:

$$\frac{f(y_1, y_2)}{\phi} \quad for\ \phi = 1 - f(y_1 = 0) - f(y_2 = 0) + f(y_1 = 0, y_2 = 0)$$

This approach can be easily extended to the case where only a single variable is truncated at zero, for example, if only $y_i = 0$, then $\phi = 1 - f(y_i = 0)$.

2.9 BIVARIATE CENSORED MODEL

A censored model is required when one set of the observations $(y_{1i}, \boldsymbol{x}_{1i})$ or $(y_{2i}, \boldsymbol{x}_{2i})$ or both are available for restricted range (y_{1i}, y_{2i}), but those for the explanatory variables (x_{1i}, x_{2i}) are always observed. A series maybe censored from below (left censored) or censored from above (right censored). Censored samples may result when high counts are not observed, or may be imposed by survey design. Thus right censoring is the most common form in the analysis of bivariate count models. Given the bivariate counts are right censored at $r = (r_1, r_2)$ so that $y_{ji} = 1,2,\dots,r_j$ for $j = 1,2$. Letting $f(y_{1i}, y_{2i}; \varphi)$ denote the complete bivariate density (Gurmu and Elder, 2000), the log-likelihood function for the right-censored bivariate count model is:

$$LL(y_1 y_2|\varphi) = \sum_{i=1}^{n} d_i[lnf(y_{1i}, y_{2i}; \varphi)] + [1 - d_i] ln\left[1 - \sum_{l=0}^{r_1-1}\sum_{m=0}^{r_2-1} f(y_{1i} = l, y_{2i} = m; \varphi)\right]$$

Where $d_i = 1$, if *y* falls in the uncensored region, and $d_i = 0$ otherwise.

3. SAS® MACRO %BICOUNT CAPABILITIES

We develop a versatile SAS® macro program "%bicount" that is capable handling ten different bivariate count data distributions. Below is the calling program for the %bicount SAS® macro:

```
*-------------------------------------------------------------------------------;
%macro bicount (indata=,summary=,dist=,type=,depend1=,depend2=,indep1=,indep2=,order=,
                zip=,hurdle=,zindep=,trunc1=,ltrunc1=,rtrunc1=,trunc2=,ltrunc2=,
                rtrunc2=,censor1=,lcensor1=,rcensor1=,censor2=,lcensor2=,rcensor2=,
                gfit=,vuong=,vdata1=,vdata2=);
*-------------------------------------------------------------------------------*
|  Macro bicount performs bivariate count regression modeling using the Proc NLMIXED|
|  procedure. The inputs into the macro include:                                  |
|                                                                                 |
|  indata  =  The name of the SAS data set that contains all the independent and  |
|             dependent variables this also contains the library name.            |
|  summary =  Identify if you want to run summary statistics on the dependent and |
|              independent variables(0=No,1=Yes)                                 |
|  dist    =  dist is the type of distribution that is used to estimate the       |
|             count regression model. Choices of distribution include:           |
|             1  Bivariate Poisson                                               |
|             2  Bivariate Poisson-Normal                                        |
|             3  Bivariate Generalized Poisson                                   |
|             4  Bivariate Negative Binomial                                     |
|             5  Bivariate Poisson Inverse Gaussian                              |
|             6  Bivariate Borel                                                 |
|             7  Bivariate Neyman Type A                                         |
|             8  Bivariate Generalized Waring                                    |
|             9  Bivariate Poisson With Series Expansion                         |
|             10 Bivariate Poisson Laguerre Polynomials                          |
|  type    =  How the likelihood function is derived. Choices include:           |
|             1  assumes independence between the two equations                  |
|              2  trivariate/reduction or convolution                            |
|              3  copula (Frank--Copula)                                         |
|  depend1 =  Name of the dependent variable for equation1                       |
|  depend2 =  Name of the dependent variable for equation2                       |
|  indep1  =  Name of the independent variables for equation1                    |
|  indep2  =  Name of the independent variables for equation2                    |
|  order   =  Order of the series expansion models (1, 2, or 3) dist=9           |
|  zip     =  If you want to estimate a zero-inflated model(0=No,1=Yes)         |
|  hurdle  =  If you want to estimate a hurdle model(0=No,1=Yes)                 |
|  zindep  =  Name of the independent variables that are used in the zero-inflated|
|             or hurdle models.                                                  |
|  trunc1  =  Dependent variable 1 is truncated (0=No, 1=Yes)                    |
|  ltrunc1 =  What value the dependent variable 1 is left truncated at           |
|  rtrunc1 =  What value the dependent variable 1 is right truncated at          |
|  trunc2  =  Dependent variable 2 is truncated (0=No, 1=Yes)                    |
|  ltrunc2 =  What value the dependent variable 2 is left truncated at           |
|  rtrunc2 =  What value the dependent variable 2 is right truncated at          |
|  censor1 =  Dependent variable 1 is censored (0=No, 1=Yes)                     |
|  lcensor1=  What value the dependent variable 1 is left censored at            |
|  rcensor1=  What value the dependent variable 1 is right censored at           |
|  censor2 =  Dependent variable 2 is censored (0=No, 1=Yes)                     |
|  lcensor2=  What value the dependent variable 2 is left censored at            |
|  rcensor2 = What value the dependent variable 2 is right censored at           |
|  gfit    =  Identify if you want to run the goodness of fit summary joint and  |
|              marginal prediction vs. actual (0=No, 1=Yes)                      |
|  vuong   =  If you want to test two non-nested models with the Voung test      |
|             (0=No, 1=Yes)                                                      |
|  vdata1 =  Name of the first data set that contains the log-likelihood values  |
|            for the first model you want to use with the Vuong test             |
|  vdata2 =  Name of the 2nd data set that contains the log-likelihood values    |
|            for the 2nd model you want to use with the Vuong test               |
*-------------------------------------------------------------------------------*;
```

4. APPLICATIONS AND MODEL EVALUATION

We use our SAS® macro program to analyze healthcare utilization data from the 1977-1978 Australian Health Survey. The data contains 5190 single-person households' healthcare service utilization information. Cameron et al. (1988) was the first using this data to analyze health service utilization. However, they applied two independent univariate negative binomial regression models to analyze doctor visits, days in hospital, and number of medicines taken. Gurmu and Elder (2000) analyzed the same data using a bivariate Poisson-Laguerre polynomial regression model to jointly estimate the factors affecting the number of doctor visits and the number of non-doctor professional visits. Wang (2003) proposes a bivariate zero-inflated negative binomial regression model to analyze the same data. Gurmu and Elder (2008) further analyzes the same data using a bivariate zero-inflated Poisson-Laguerre polynomial regression model. We analyze two possibly jointly dependent variables of health service utilization measures: (1) the number of consultations with doctors during the 2-week period prior to the survey (Doctorco); and (2) the number of prescribed medicines used in the past two weeks (Prescrib).

As in previous research, the explanatory variables consist of three groups: four socio-economic variables, four insurance status variables, and five health status variables. The socio-economic variables are: (sex) a dummy variable for gender; (Age) age in years divided by 100; (Agesq) squared of (Age); and (Income) annual income in ten-thousands of dollars. The four insurance status variables are: (Levy), default government insurance converges paid by income levy; (Levyplus), private insurance coverage; (Freepoor), free government insurance due to low income; (Freerepa) free government coverage due to old age, disability or veteran status. The five health status variables are: (Illness), number of illnesses in the past two weeks; (Actdays), number of days of reduced activity in the past two weeks due to illness or injury; (Hscore), general health questionnaire score using Golberg's method with high score indicating bad health; (Chcond1), indicator variable for chronic condition not limiting activity; and (Chcond2), indicator variable for chronic condition limiting activity. Cameron et al. (1988) provides summary statistics of dependent and explanatory variables.

Table 1 reports the results of parameter estimates for the zero-inflated bivariate Poisson Laguerre polynomial (Frank-copula) model. This model has the best AIC (Akaike Information Criterion) among all models. The estimates show that recent health status measures (illness, actdays) and two of the long-term health measures (hscore, chcond1) are important determinants of both doctor visits and number of prescribed medicines. The positive coefficient on the health insurance status (levyplus) indicates that private insurance is associated with higher use of doctor's visits and prescription drugs. While age and gender are unimportant determinants in doctor's visits; they are positively related to use of prescription drugs.

Table 1

Estimates for the Zero Inflated Bivariate Poisson Laguerre Polynomial (Frank--copula)

	Doctor Visits		Prescription Medicines		Zero Inflation	
Parameter	Estimate	T-value	Estimate	T-value	Estimate	T-value
intercept	-1.649	-6.86	-2.244	-14.71	0.279	1.93
freepoor	-0.501	-2.15	0.019	0.11	0.418	0.82
freerepa	0.167	1.45	0.237	3.42	-0.614	-2.54
illness	0.155	6.25	0.166	11.74	-0.455	-2.82
actdays	0.122	16.22	0.029	5.79	-1.819	-1.8
hscore	0.038	2.89	0.017	2.14	-0.062	-1.1
chcond1	-0.174	-1.98	0.440	7.05	-1.487	-5.76
chcond2	-0.093	-0.86	0.654	9.05	-2.425	-2.48
sex	0.107	1.54	0.488	11.51		
age	-0.145	-0.11	2.747	3.81		
age_squ	0.587	0.43	-0.951	-1.25		
income	-0.158	-1.4	0.010	0.16		
levyplus	0.201	2.3	0.252	4.35		
α	1.534	7.15	2.541	3.04		
a11	0.490	8.43	-1.613	-3.45		
φ			1.532	8.18		

Table 2 summarizes all bivariate models incorporated in our SAS® macro. The three sets of model evaluation statistics are: maximum log likelihood function, the AIC (Akaike Information Criterion), and BIC (Bayesian information criterion). Table 2 indicates that estimating the two count events jointly is better than estimating the two count events independently. Among all bivariate models, the bivariate generalized Waring (Frank-copula) is the best model based on log likelihood function and AIC; and the Poisson-lognormal (copula) model is the best based on the BIC. Formulas for the AIC (Akaike information criterion) and BIC (Bayesian information criterion) include:

$$AIC = -2lnL + 2k$$ k= number of parameters estimated

$$BIC = -2lnL + (ln(n))\mathrm{k}$$ n=number of observations in the data set

Table 2 Model Evaluation Statistics for Bivariate Models

Bivariate Model	Log Likelihood Function	AIC	BIC
Independent Poisson	-8,886.31	17,825	17,995
Bivariate Poisson	-8,825.74	17,706	17,882
Bivariate Poisson (Frank--copula)	-8,773.32	17,601	17,778
Independent Poisson-LogNormal	-8,632.73	17,321	17,505
Bivariate Poisson-LogNormal	-8,551.51	17,161	17,351
Bivariate Poisson-LogNormal (Frank--copula)	-8,506.03	17,071	17,261
Independent Generalized Poisson	-8,648.78	17,354	17,537
Bivariate Generalized Poisson	-8,576.88	17,214	17,410
Bivariate Generalized Poisson (Frank--copula	-8,534.65	17,127	17,317
Independent Negative Binomial	-8,640.23	17,336	17,520
Bivariate Negative Binomial	-8,600.72	17,255	17,432
Bivariate Negative Binomial (Frank--copula)	-8,522.70	17,103	17,293
Independent Poisson Inverse Gaussian	-8,635.23	17,326	17,510
Bivariate Poisson Inverse Gaussian	-8,603.09	17,260	17,437
Bivariate Poisson Inverse Gaussian (Frank-- copula)	-8,517.98	17,094	17,284
Independent Borel-Tanner	-9,238.44	18,529	18,699
Bivariate Borel-Tanner	-8,570.49	17,199	17,386
Bivariate Borel-Tanner (Frank--copula)	-9,091.58	18,237	18,414
Independent Neyman Type A	-8,668.85	17,394	17,577
Bivariate Neyman Type A	-8,602.56	17,261	17,445
Bivariate Neyman Type A (Frank--copula)	-8,555.46	17,169	17,359
Independent Generalized Waring	-8,618.52	17,297	17,494
Bivariate Generalized Waring	-8,576.90	17,210	17,393
Bivariate Generalized Waring (Frank--copula)	-8,503.76	17,070	17,273
Independent Poisson Series Expansion	-8,680.29	17,421	17,617
Bivariate Poisson Series Expansion	-8,576.22	17,214	17,418
Bivariate Poisson Series Expansion (Frank--copula)	-8,563.27	17,189	17,392
Independent Poisson-Laguerre Polynomial	-8,629.01	17,318	17,515
Bivariate Poisson-Laguerre Polynomial	-8,559.04	17,176	17,366
Bivariate Poisson-Laguerre Polynomial (Frank--copula)	-8,511.64	17,085	17,288

The model evaluation statistics for zero-inflated bivariate count regression models are reported in Table 3. Since almost 54% of all observations occur when doctor visits and the number of prescribed medicines are zero a zero-inflated model would seem more appropriate. Comparing the model evaluation criterion between Table 2 and Table

3, one can conclude that the zero-inflated bivariate models perform better than their counter parts in table 2. Among zero-inflated bivariate models, the bivariate Poisson Laguerre polynomial (Frank copula) is the best model.

Table 3. Model Evaluation Statistics of Zero-inflated Bivariate Models

Bivariate Model	Log Likelihood Function	AIC	BIC
Bivariate Poisson	-8,592.32	17,255	17,484
Bivariate Poisson (Frank--copula)	-8,576.22	17,222	17,452
Bivariate Poisson-LogNormal	-8,463.06	17,001	17,243
Bivariate Generalized Poisson	-8,487.80	17,052	17,301
Bivariate Generalized Poisson (Frank--copula	-8,476.88	17,028	17,270
Bivariate Negative Binomial	-8,507.00	17,084	17,313
Bivariate Negative Binomial (Frank--copula)	-8,449.05	16,972	17,215
Bivariate Poisson Inverse Gaussian	-8,507.82	17,086	17,315
Bivariate Poisson Inverse Gaussian (Frank-- copula)	-8,444.76	16,964	17,206
Bivariate Borel-Tanner	-8,470.60	17,015	17,258
Bivariate Borel-Tanner (Frank--copula)	-9,081.63	18,163	18,463
Bivariate Neyman Type A	-8,507.67	17,087	17,323
Bivariate Neyman Type A (Frank--copula)	-8,495.70	17,065	17,308
Bivariate Generalized Waring	-8,506.79	17,086	17,322
Bivariate Generalized Waring (Frank--copula)	-8,455.72	16,989	17,245
Bivariate Poisson Series Expansion	-8,474.73	17,027	17,283
Bivariate Poisson Series Expansion (Frank--copula)	-8,456.39	16,995	17,264
Bivariate Poisson-Laguerre Polynomial	-8,458.65	16,991	17,234
Bivariate Poisson-Laguerre Polynomial (Frank--copula)	-8,440.71	16,959	17,215

5. CONCLUSIONS

This paper develops a SAS® macro program using PROC NLMIXED to model a variety of bivariate count data regression models. The SAS® macro is capable of estimating ten count data distributions with the option of fitting zero-inflated, hurdle, truncated and censored models for each of these distributions. We demonstrate the power and versatility of the proposed SAS® macro by applying it to a healthcare utilization data. By specifying the parameters of the proposed SAS® macro, we can estimate a variety of bivariate count data regression models. Table 2 shows a number of bivariate count data regression models that are incorporated into our SAS® macro program. In addition to providing coefficient estimates, it also provides the model evaluation statistics such as log-likelihood function, AIC and BIC to help model selection. The proposed SAS® macro also provides comparisons between the predicted and actual values (goodness -of-fit test) that can also be used in the model evaluation.

REFERENCES

Atella, V. and P. Deb (2008) Are Primary Care Physicians, Public and Private Sector Specialists Substitutes or Complements? Evidence from a Simultaneous Equations Model for Count Data, *Journal of Health Economics*, Vol. 27 770–785.

Brown, S., S. Hillegeist, K. Lo (2006), The Effect of Meeting Or Missing Earnings Expectations on Information Asymmetry, working paper, The University of British Columbia.

Cameron, C., Trivedi, P.K., Milne, F., Piggott, J. (1988). A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia. Review of Economic Studies, Vol. 55, 85–106.

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

Cameron, A, T. Li, P. Trivedi, D. Zimmer (2004) Modeling the Differences in Counted Outcomes Using Bivariate Copula Models with Application to Mismeasured Counts, *Econometrics Journal*, Vol. 7, 566-584.

Chou, N. and D. Steenhard (2009), A Flexible Count Data Regression Model Using SAS® PROC NLMIXED, *Proceedings SAS Global Forum 2009*, paper 250-2009.

Gurmu, S. and J. Elder (2000) Generalized Bivariate Count Data Regression Models, *Economics Letters*, Vol. 68, 31-36.

_____, (2008) A Bivariate Zero-Inflated Count Data Regression Model with Unrestricted Correlation, *Economics Letters*, Vol. 100, 245-248.

Ho, W, P. Wang, J. Alba (2009) Merger and Acquisition FDI, Relative Wealth and Relative Access to Bank Credit: Evidence from a Bivariate Zero-inflated Count Model, *International Review of Economics and Finance*, Vol. 18, 26-30.

Jorgensen, B. (1987), Exponential Dispersion Models, *Journal of the Royal Statistical Society B*, 49, 127-162.

Kocherladota, S. and K. Kocherlakota (1992), *Bivariate Discrete Distributions*, Marcel Dekker: New York.

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

Lee A., K. Wang, K. Yau, P. Carrivick, and M. Stevenson (2005) Modeling Bivariate Count Series with Excess Zeros, *Mathematical Biosciences*, vol. 196, 226-237.

Mayer, W. and W. Chappell (1992), Determinants of Entry and Exit: An Application of the Compounded Bivariate Poisson Distribution to U.S. industries, 1972-1977, *Southern Economic Journal*, Vol. 53, No.3, 770-778.

Morat, L. (2009) A Priori Ratemaking Using Bivariate Poisson Regression Models, *Insurance Mathematics and Economics*, Vol. 44, 135-141.

Mullahy, J. (1986), Specification and Testing in Some Modified Count Data Models, *Journal of Econometrics*, 33: 341-365.

Riphahn, R, A. Wambach, A. Million (2003) Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation, *Journal of Applied Econometrics*, Vol. 18, 387-405.

Trivedi, P. and D. Zimmer (2007), Copula Modeling: An Introduction for Practitioners, *Foundations and Trends in Econometrics*, Vol. 1, No. 2, 1-111.

Vernic, R. (1997), On the Bivariate Generalized Poisson Distribution, *Astin Bulletin*, Vol. 27, No. 1, 23-31.

Wang, K., A. Lee, K. Yau, P. Carrivick (2003) A Bivariate Zero-Inflated Poisson Regression Model to Analyze Occupational Injuries, Accident Analysis and Prevention, Vol. 35, 625-629.

Wang, P. (2003) A Bivariate Zero-Inflated Negative Binomial Regression Model for Count Data with Excess Zeros, *Economics Letters*, Vol. 78, 373-378.

CONTACT INFORMATION

Nan-Ting Chou
University of Louisville,
Economics Department, College of Business
2301 South Third Street
Louisville, KY 40208
ntchou01@louisville.edu

David Steenhard, Senior Statistician
LexisNexis
9443 Springboro Pike
Dayton, OH 45342
Email: david.steenhard@lexisnexis.com