E C O N O M I C S B U L L E T I N

A simple bivariate count data regression model

Shiferaw Gurmu
Georgia State University

John Elder
North Dakota State University

Abstract

This paper develops a simple bivariate count data regression model in which dependence between count variables is introduced by means of stochastically related unobserved heterogeneity components. Unlike existing commonly used bivariate models, we obtain a computationally simple closed form of the model with an unrestricted correlation pattern. An application to Medicaid utilization is provided.

Citation: Gurmu, Shiferaw and John Elder, (2007) "A simple bivariate count data regression model." *Economics Bulletin*, Vol. 3, No. 11 pp. 1-10
Submitted: February 19, 2007. **Accepted:** April 1, 2007.
URL: http://economicsbulletin.vanderbilt.edu/2007/volume3/EB-07C30029A.pdf

1 Introduction

Bivariate count data regressions arise in situations where two dependent counts are correlated and joint estimation is required mainly due to efficiency considerations. For example, common measures of health-care utilization, such as the number of doctor consultations and the number of other ambulatory visits, are likely to be jointly dependent. Other leading examples include the number of voluntary and involuntary job changes, the number of firms which enter and exit an industry, and the number of patents granted to and papers published by scientists.

Existing commonly used count models accommodate only non-negative correlation between the counts (Mayer and Chappell 1992, Gurmu and Elder 2000, and Wang 2003). The statistics literature gives examples and general techniques on constructing negatively correlated multivariate Poisson distributions having Poisson marginals. In particular, Aitchison and Ho (1989) consider a log-normal mixture of independent Poisson distributions. Since the resulting mixture, the Poisson-log normal distribution, does not have a closed form solution, estimation of the model requires numerical integration (Munkin and Trivedi 1999 and Hellstrom 2006).

In this paper, we develop a simple bivariate count regression model in which dependence between count variables is introduced by means of stochastically related unobserved heterogeneity components. The proposed bivariate Poisson mixture model is based on the first-order series expansion for the unknown joint density of the unobserved heterogeneity components. Unlike existing commonly used bivariate models, we obtain a computationally simple closed form of the model with an unrestricted correlation pattern. We also provide an extension to truncated models. An application to Medicaid utilization is provided.

2 The framework

This section provides the basic framework for two-factor mixture models in which dependence between count variables is introduced through correlated unobserved heterogeneity components. Consider two jointly distributed random variables, Y_1 and Y_2, each denoting event counts. For observation i $(i = 1, 2, ..., N)$, we observe $\{y_{ji}, x_{ji}\}_{j=1}^{2}$, where x_{ji} is a $(k_j \times 1)$ vector of covariates. Without loss of generality, the mean parameter associated with

y_{ji} can be parameterized as

$$\theta_{ji} = \exp(x'_{ji}\beta_j), \qquad j = 1, 2 \tag{1}$$

where β_j is a $(k_j \times 1)$ vector of unknown parameters. We model the dependence between y_1 and y_2 by means of correlated unobserved heterogeneity components ν_1 and ν_2. Each of the components is associated with only one of the event counts. Accordingly, for $j = 1, 2$, suppose $(y_{ji} \mid x_{ji}, \nu_{ji}) \sim$ Poisson$(\theta_{ji}\nu_{ji})$ with (ν_{1i}, ν_{2i}) having a bivariate distribution $g(\nu_{1i}, \nu_{2i})$ in $\Re^2_+$. Then the ensuing mixture density can be expressed as

$$f(y_{1i}, y_{2i} \mid x_i) = \int\int \left[\prod_{j=1}^{2} \frac{\exp(-\theta_{ji}\nu_{ji})\,(\theta_{ji}\nu_{ji})^{y_{ji}}}{\Gamma(y_{ji}+1)}\right] g(\nu_{1i}, \nu_{2i})\mathrm{d}\nu_{1i}\mathrm{d}\nu_{2i}. \tag{2}$$

Let $M(-\theta_{1i}, -\theta_{2i}) = E_\nu\left[\exp\left(-\theta_{1i}\nu_{1i} - \theta_{2i}\nu_{2i}\right)\right]$ denote the bivariate moment generating function (MGF) of (ν_{1i}, ν_{2i}) evaluated at $(-\theta_{1i}, -\theta_{2i})$. It can readily be seen that (2) takes the form

$$f(y_{1i}, y_{2i} \mid x_i) = \left[\prod_{j=1}^{2} \frac{(\theta_{ji})^{y_{ji}}}{\Gamma(y_{ji}+1)}\right] M^{(y_1, y_2)}\left(-\theta_{1i}, -\theta_{2i}\right), \tag{3}$$

where, suppressing i, $M^{(y_1,y_2)}(-\theta_1, -\theta_2) = \partial^{y_\cdot} M(-\theta_1, -\theta_2) \,/\, (\partial(-\theta_1)^{y_1}\partial(-\theta_2)^{y_2})$ is the derivative of $M(-\theta_1, -\theta_2)$ of order $y_\cdot = y_1 + y_2$.

The sign of the correlation coefficient between y_1 and y_2 is determined by the sign of the covariance between the two unobserved variables, $\mathrm{cov}(\nu_1, \nu_2)$. In the case of univariate mixing, the correlation between the counts is affected only by the variance of the common unobserved heterogeneity term. Hence, correlation is non-negative. In the bivariate mixing, the variance of each unobserved component as well as the correlation between the components affect $\mathrm{corr}(y_{1i}, y_{2i} \mid x_i)$. Hence, the sign of this correlation is unrestricted.

The form of the density (3) depends upon the choice of the distribution of the unobservables, $g(\nu_{1i}, \nu_{2i})$. If $g(.)$ follows a bivariate (or generally multivariate) log-normal distribution, we get the bivariate (or multivariate) Poisson log-normal distribution proposed by Aitchison and Ho (1989). The computational difficulty with the Poisson log-normal mixture arises from the unavailability of the MGF of the log-normal distribution. Hence, evaluation of $M^{(y_1,y_2)}(-\theta_{1i}, -\theta_{2i})$ and estimation of the model require numerical integration. For example, Munkin and Trivedi (1999) study the Poisson log-normal correlated model using the simulated maximum likelihood estimation method, while Hellstrom (2006) uses Markov chain Monte Carlo Methods.

3 A general bivariate model

We obtain a closed form for a mixture model of the type given in (3), while at the same time allowing for both positive and negative correlations. The proposed simple mixture model is based on first-degree Laguerre polynomial expansion of the bivariate distribution of unobserved heterogeneity, where the leading term is the product of gamma densities. The proposed density for (ν_{1i}, ν_{2i}) is

$$g(\nu_{1i}, \nu_{2i}) = \frac{w(\nu_{1i})w(\nu_{2i})}{(1+\rho_{11}^2)} \left[1 + \rho_{11} P_1(\nu_{1i}) P_1(\nu_{2i})\right]^2, \tag{4}$$

where, for $j = 1, 2$,

$$w(\nu_{ji}) = \frac{\nu_{ji}^{\alpha_j - 1} \lambda_j^{\alpha_j}}{\Gamma(\alpha_j)} e^{-\lambda_j \nu_{ji}} \tag{5}$$

are the baseline gamma weights,

$$P_1(\nu_{ji}) = \left(\sqrt{\alpha_j} - \frac{\lambda_j}{\sqrt{\alpha_j}} \nu_{ji}\right) \tag{6}$$

are the first-order polynomials each with unit variance, and ρ_{11} is an unknown correlation parameter; $\rho_{11} = \text{corr}(P_1(\nu_1), P_1(\nu_2))$. The polynomials in (4) are squared to ensure non-negativity of the density.

The mixture density in (2) can now be derived using specification (4). After some algebra, we obtain the following bivariate density for the counts:

$$f(y_{1i}, y_{2i} \mid x_i) = \left[\prod_{j=1}^{2} \frac{\Gamma(y_{ji} + \alpha_j)}{\Gamma(\alpha_j)\Gamma(y_{ji}+1)} \left(\frac{\theta_{ji}}{\lambda_j}\right)^{y_{ji}} \left(1 + \frac{\theta_{ji}}{\lambda_j}\right)^{-(\alpha_j + y_{ji})}\right] \Psi_i \tag{7}$$

where

$$\lambda_j = \tfrac{1}{1+\rho_{11}^2}\left[\alpha_j + \rho_{11}^2(\alpha_j + 2)\right], \qquad j = 1, 2 \tag{8}$$

and

$$\begin{aligned} \Psi_i = \tfrac{1}{1+\rho_{11}^2} \big[&1 + 2\rho_{11}\sqrt{\alpha_1\alpha_2}\,(1-\eta_{1i})\,(1-\eta_{2i}) + \\ &\rho_{11}^2 \alpha_1\alpha_2\,(1 - 2\eta_{1i} + \eta_{1i}\zeta_{1i})\,(1 - 2\eta_{2i} + \eta_{2i}\zeta_{2i})\big], \end{aligned} \tag{9}$$

with $\eta_{ji} = \frac{y_{ji}+\alpha_j}{\alpha_j}\left(1 + \frac{\theta_{ji}}{\lambda_j}\right)^{-1}$ and $\zeta_{ji} = \frac{y_{ji}+1+\alpha_j}{\alpha_j}\left(1 + \frac{\theta_{ji}}{\lambda_j}\right)^{-1}$ for $j = 1, 2$. The bivariate density in (7) can also be expressed in the general form (3).

This is achieved by replacing $M^{(y_1,\,y_2)} - \theta_{1i}, -\theta_{2i})$ in (3) with

$$M_a^{(y_1,\,y_2)}\left(-\theta_{1i}, -\theta_{2i}\right) = \left[\prod_{j=1}^{2} \frac{\Gamma(y_{ji}+\alpha_j)}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j} \left(\lambda_j + \theta_{ji}\right)^{-(\alpha_j + y_{ji})}\right] \Psi_i. \qquad (10)$$

The alternative representation of the approximated density is useful in obtaining the moments of the model. As in the univariate Poisson mixture model, we have set the mean of each unobserved heterogeneity to unity. This imposes restriction on λ_j given in (8). The unknown parameters, $\varphi = (\beta_1, \beta_2, \alpha_1, \alpha_2, \rho_{11})$, can then be obtained by maximizing the log-likelihood function, $\sum_{i=1}^{N} \log f(y_{1i}, y_{2i} \mid x_i)$. The mixture model based on (7) is called the bivariate Poisson-Laguerre polynomial (BIVARPL) model. It can be thought of as a mixture of Poisson and a variant of a bivariate gamma distribution.[1]

Interest lies in lower order conditional moments of the BIVARPL, including the conditional correlation between y_1 and y_2. For the BIVARPL model, since $\text{Mean}(y_{ji} \mid x_i) = \theta_{ji}$, the marginal effects of a certain explanatory variable, say u_i, on the expected number of counts (*e.g.*, trips) is $\text{ME}_u = \theta_{ji} \times \beta_{ju}, \quad j = 1, 2$. Finite difference method can be used for discrete regressors. The correlation coefficient for the BIVARPL model is:

$$\text{corr}\left(y_{1i}, y_{2i} \mid x_i\right) = \frac{\theta_{1i}\theta_{2i}\left[M_a^{(1,1)}(0,0) - 1\right]}{\sqrt{\left[\theta_{1i} + \theta_{1i}^2\left(M_a^{(2,0)}(0,0) - 1\right)\right]\left[\theta_{2i} + \theta_{2i}^2\left(M_a^{(0,2)}(0,0) - 1\right)\right]}}, \qquad (11)$$

where

$$M_a^{(1,1)}(0,0) = \left[\alpha_1\alpha_2 + 2\rho_{11}\sqrt{\alpha_1\alpha_2} + \rho_{11}^2(\alpha_1+2)(\alpha_2+2)\right]/\lambda_1\lambda_2, \qquad (12)$$

$$M_a^{(2,0)}(0,0) = \frac{(\alpha_1+1)\left[\alpha_1 + \rho_{11}^2(\alpha_1+6)\right]}{\lambda_1^2(1+\rho_{11}^2)}, \qquad (13)$$

and for $M_a^{(0,2)}(0,0)$, we replace α_1 and λ_1 in the preceding equation by α_2 and λ_2, respectively. Note that, for example, $\text{Var}(y_{1i} \mid x_i) = \theta_{1i} + \theta_{1i}^2\left[M_a^{(2,0)}(0,0) - 1\right]$ in (11). The conditional correlation can take on zero,

[1] For computational simplicity, this paper focuses on bivariate models with only first-order expansion. Higher order polynomial expansions, say of order K, can be considered (Gurmu and Elder 2006).

positive or negative values. When the correlation parameter $\rho_{11} = 0$ in (7), we get a density that is a product of two independent negative binomial distributions.

The above analysis can be extended to the estimation of truncated and censored models. We focus on the empirically relevant case, the zero- truncated model, where the zero class is missing for both dependent variables so that $y_{ji} = 1, 2, 3, ...$ for $j = 1, 2$. The zero-truncated bivariate distribution takes the form

$$\frac{f(y_1, y_2; \delta)}{\phi}, \quad y \in S^*$$

where δ is a parameter vector, $\phi = \sum\sum_{y \in S^*} f(y_1, y_2; \delta)$, and S^* is a set of positive integers in $\Re^2$. The normalization constant can be derived as

$$\phi_i = 1 - f(y_1 = 0) - f(y_2 = 0) + f(y_1 = 0, y_2 = 0), \tag{14}$$

where, for example, $f(y_1 = 0) = f(y_1 = 0,\ y_2 \geq 0)$ and

$$f(y_1 = 0, y_2 = 0) = \left[\prod_{j=1}^{2} \left(1 + \frac{\theta_{ji}}{\lambda_j}\right)^{-\alpha_j}\right] \Psi_i(y_{1i} = 0, y_{2i} = 0)$$

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

4 An application

Using bivariate regressions, we model the number of doctor and other ambulatory visits during a period of four months based on data from the 1986 Medicaid Consumer Survey. The survey was part of the data collection activity of the Nationwide Evaluation of Medicaid Competition Demonstrations. This paper focuses on data obtained from two sites in California, and originally analyzed by Gurmu (1997) using univariate models. The California survey was conducted in personal interviews with samples of demonstration enrollees in Santa Barbara county and a fee-for-service comparison group of nonenrollees from nearby Ventura county. A stratified random sample of individuals qualifying for Aid to Families with Dependent Children was obtained in 1986. The sample size is 243 for enrollees, and 242 for nonenrollees.

An important feature of the data set is that enrollment in the programs was mandatory for all Medicaid beneficiaries.

The dependent variables are (1) the number of doctor office and clinic visits (*Doctor*) and (2) the number of other ambulatory visits, including hospital clinic, outpatient, health center, and home visits (*Ambulatory*), both observed over a period of four months. The explanatory variables include the number of children in the household, age of the respondent in years, annual household income, dummy variables for race and marital status, years of schooling, access to health services, and measures of health status. Three of the health related variables, functional limitations, chronic conditions, and acute conditions, are highly correlated. Accordingly, the first two of the principal components (called *PC1* and *PC2*) are used as explanatory variables. The first principal component accounts for 68.5% of the variation, and is positively correlated with each of the health related variables. Thus, one would expect the first principal component to have a positive impact on health care utilization.

The two count variables are negatively correlated, with the sample correlation of -0.044. Most of the observed joint frequencies for (*Doctor*, *Ambulatory*) visits are at cells : (0, 0), (0, y_2), and (y_1, 0). The counts are characterized by relatively high proportion of nonusers; 61.9% for doctor visits and 73.8% for other ambulatory visits. In each case, about 10% of the respondents have 4 or more visits during the reporting period. As compared to nonenrollees, the means of both utilization variables are lower for enrollees.

Table 1 presents parameter estimates from two bivariate models. All t-ratios are based on heteroscedasticity-robust standard errors. Fore comparison, the estimates from the bivariate negative binomial model, which restrict correlations to be positive, are also included. The BIVARPL model dominates the bivariate negative binomial model in terms of the Akaike information criterion (AIC). The main health status variable PC1 has a highly significant positive impact on the number of doctor and other ambulatory visits. Both doctor and ambulatory visits decrease with the number of children, and tend to have a concave relationship with age. The enrollment coefficient for doctor visits is negative and significant. This suggests that enrollment in the managed care program leads to a decrease in the number of doctor office visits. On the other hand, the enrollment coefficient is insignificant in the ambulatory equation.

We have also computed the predicted marginal effects of changes in the

Table 1: Coefficient Estimates and t-ratios for Bivariate Negative Binomial and Bivariate Poisson-Laguerre Polynomial Models

Variable	Bivariate Negative Binomial				BIVARPL			
	Doctor		*Ambulatory*		*Doctor*		*Ambulatory*	
	Est.	$\mid t \mid$	Est.	$\mid t \mid$	Est.	$\mid t \mid$	Est.	$\mid t \mid$
Constant	-1.147	.62	-1.413	.87	-1.103	.33	-.635	.02
Children	-.234	2.33	-.150	1.65	-.149	.50	-.223	.64
Age	.085	.91	.069	.78	.065	.32	.011	.01
(Age)^2 $\times 10^2$	-.135	1.06	-.120	1.00	-.117	.40	-.033	.02
Income $\times 10^{-4}$	.297	.64	.548	1.38	.194	.20	.770	.22
PC1	.394	5.83	.296	3.47	.372	3.94	.365	4.27
PC2	-.060	.58	.034	.41	.009	.10	.046	.03
Access	.009	1.71	-.009	1.41	.010	1.13	-.010	1.34
Married	-.082	.27	-.634	1.79	-.014	.18	-.653	1.05
White	-.192	.76	.222	.90	-.008	.01	.222	.26
Schooling	.009	.25	.033	.73	.019	.23	.048	.13
Enroll	-.683	2.56	.008	.09	-.609	2.31	-.072	.24
$\log(\theta_{12}$)								
$\log(\alpha)$	-.370	2.79						
$\log(\alpha_j)$					-.839	6.18	-1.638	10.33
ρ_{11}					-.163	5.28		
Log-likelihood	-1507.0				-1166.8			
AIC	3064.0				2387.8			

explanatory variables on the mean number of doctor and ambulatory visits (not reported). Generally, the estimated marginal effects are smaller in BIVARPL than in the bivariate negative binomial model. The sample average of the correlation between *Doctor* and *Ambulatory* is 0.544 for the bivariate negative binomial and about 0.014 for BIVARPL model.

5 Conclusion

We have developed a general bivariate count regression model for which, unlike existing commonly used models, we obtain a computationally simple closed form of the model with an unrestricted correlation pattern. The model allows for truncation and censoring without further computational complexity. In the empirical illustration, the proposed model fits the data better than the bivariate negative binomial model.

References

[1] Aitchison, J. and C.H. Ho (1989) "The Multivariate Poisson-log Normal Distribution" *Biometrika* 76, 643-653.

[2] Gurmu, S. (1997) "Semi-parametric Estimation of Hurdle Regression Models with an Application to Medicaid Utilization" *Journal of Applied Econometrics* 12, 225-242.

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

[4] Gurmu, S. and J. Elder (2006) "Estimation of Multivariate Count Regression Models with Application" Working Paper, Georgia State University.

[5] Hellstrom, J. (2006) "A Bivariate Count Data Model for Household Tourism Demand" *Journal of Applied Econometrics* 21, 213-226.

[6] Mayer, W.J. and W.F. Chappell (1992) "Determinants of Entry and Exit : An Application of the Compounded Poisson Distribution to US Industries, 1972-1977" *Southern Economic Journal* 58, 770-778.

[7] Munkin, M. and P. K. Trivedi (1999) "Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models,With Application" *Econometric Journal* 1, 1-20.

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