ROCKWOOL Foundation Berlin
Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES 126/26

Fitting the bivariate mixed Poisson regression model by maximum simulated likelihood

Stephen Jenkins, Fernando Rios-Avila

www.rfberlin.com

APRIL 2026

Fitting the bivariate mixed Poisson regression model by maximum simulated likelihood

Authors

Stephen Jenkins, Fernando Rios-Avila

Reference

JEL Codes: C31, C35, C15

Keywords: bivariate mixed Poisson regression, maximum simulated likelihood, sampling function, count data

Recommended Citation: Stephen Jenkins, Fernando Rios-Avila (2026): Fitting the bivariate mixed Poisson regression model by maximum simulated likelihood. RFBerlin Discussion Paper No. 126/26

Access

Disclaimer

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com

Fitting the bivariate mixed Poisson regression model by maximum simulated likelihood

Stephen P. Jenkins
London School of Economics and Political Science
London, U.K.
s.jenkins@lse.ac.uk
Fernando Rios-Avila
London School of Economics and Political Science
London, U.K.
f.rios.a@gmail.com

Abstract. We introduce `bimpoisson`, a program to fit the bivariate mixed Poisson regression model by maximum simulated likelihood using the two approaches proposed by Munkin and Trivedi ('Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application', *The Econometrics Journal* 2, 1999, 29–48). By default, `bimpoisson` uses their sampling function approach; optionally their standard MSL approach is available. `bimpoisson` allows either pseudo-random uniform draws or Halton draws for simulation. Additional options allow use of antithetic acceleration and a first-order bias correction. Like Jumamyradov and Munkin ('Biases in maximum simulated likelihood estimation of bivariate models', *Journal of Econometric Methods* 11, 2022, 55–70), we use a modified version of Munkin and Trivedi's sampling function to provide better coverage. We also provide post-estimation tools to predict conditional count probabilities and expected counts. We examine `bimpoisson`'s performance using Monte Carlo simulation analysis and our empirical illustrations fit models using the same bivariate count data as used by Xu and Hardin ('Regression models for bivariate count outcomes', *The Stata Journal*, 16, 2016, 301–315) and Munkin and Trivedi (1999). We provide practical advice about which MSL estimator and types of draws and number to use.

Keywords: st0001, `bimpoisson`, bivariate mixed Poisson regression, maximum simulated likelihood, sampling function

1 Introduction

Poisson models, and variations on them such as negative binomial models, are popular ways to model an outcome that is a discrete (integer) count. For example, you might model the number of hospital visits per year made by individuals using Stata's `poisson` or `nbreg`. But what if this dataset also includes data about the number of visits per year made to a doctor? You might expect there to be a non-zero correlation between the outcomes, i.e., the types of people who are more likely than others to visit a doctor are also more likely to visit a hospital. It is of interest to model the two outcomes jointly accounting for the correlation, also investigating whether that correlation remains if one adjusts for differences between individuals in observed characteristics like age and sex. Cameron and Trivedi (2013, Chapter 8) review concepts of dependence in count data models but do not discuss software. Although official Stata does not currently include programs for fitting bivariate count data regression models, Xu and Hardin's (2016) community-contributed program `bivcnto` provides several models for this purpose, including specifications that accommodate dependence using a copula. In this paper, we extend the portfolio of estimators for bivariate count data regression models further by providing a program, `bimpoisson`, which models dependence using a bivariate normal mixture, as proposed by Munkin and Trivedi (1999).

The mixture model of Munkin and Trivedi (1999), hereafter 'MT', is an example of a model with latent factors, otherwise known as random effects, and commonly used in statistics. Random effects enter the models for the marginal outcomes in the same way as regressors do – each effect contributes an additional term to a linear index – and is assumed to be uncorrelated with the regressors. You may be familiar with the bivariate probit model (fit with `biprobit` in Stata), in which each of the two latent outcomes depends on a linear index that includes an error term (a random effect), and the two errors are distributed bivariate standard normal. In the bivariate mixed Poisson model, the count probability for each outcome depends on a linear index including an error, and the two errors are bivariate normally distributed with zero means (see section 2). The introduction of the random effects introduces overdispersion into the standard Poisson model. Unlike the log-likelihood expression for the bivariate probit model, the bivariate mixed Poisson model's log-likelihood does not have a closed form expression. Hence, MT's estimators for the model are based on maximum simulated likelihood (MSL) methods.

With MSL, each observation's likelihood contribution is derived by repeatedly taking (quasi) random draws from the distribution of random effects, calculating the contribution each replication, and then taking the average across the simulations. As MT explain, if S, $N \to \infty$ with $\sqrt{N}/S \to 0$, the MSL estimator is asymptotically equivalent to the exact maximum likelihood (ML) estimator with a limiting normal distribution (Gouriéroux and Monfort 1991). The covariance matrix is consistently estimated by the robust sandwich estimator. If S is fixed, the MSL estimator is inconsistent, but there is a first-order bias correction to the likelihood that can be applied to mitigate this problem, which is implemented by MT. A further variant on standard MSL methods that MT employ is a simulator that uses an importance sampling function such that the revised MSL estimator has the same properties as those cited above. The motivation is that maximization of the standard simulated log-likelihood using gradient-based methods is infeasible because its derivative is unbounded if the correlation between the random effects is close to -1 or $+1$.

Our program `bimpoisson` implements a sampling function approach by default (with the standard MSL estimator as an option), and the first-order bias correction as an option. We modify MT's sampling function, however. MT use the bivariate standard normal distribution with correlation set equal to zero. Instead, like Jumamyradov and Munkin (2022), hereafter 'JM', we use a sampling distribution that is bivariate normal with means zero, but with standard deviations unrestricted. We also allow a non-zero correlation (set by an option). With this sampling function approach, drawn values provide more effective coverage and improve MSL's estimation of the double integral that specifies the density of the bivariate mixed Poisson model. By default, and following MT, `bimpoisson`'s two MSL estimators use pseudorandom numbers (draws from the standard uniform distribution) when simulating probabilities. Optionally, `bimpoisson` allows Halton draws to be used instead. Users can also choose to use antithetic acceleration.

For further discussion of simulation estimation techniques and their applications, see Gouriéroux and Monfort (1997) and Train (2003). For textbook discussions, see, e.g., Greene (2003, Chapter 17) and Cameron and Trivedi (2005, Chapter 12). The first publicly available Stata implementation of MSL methods we are aware of was by Cappellari and Jenkins (2003) (`mvprobit` for fitting multivariate probit models), extended by Cappellari and Jenkins (2006). Other articles in the 2006 *Stata Journal* special issue on MSL include those by Deb and Trivedi (2006)(`mtreatnb` for fitting negative binomial models with endogenous multinomial treatment), Haan and Uhlendorff (2006) (fitting mixed multinomial logit models), and Stewart (2006) (`redpace`, for fitting random-effects dynamic probit models with autocorrelated errors). See also Hole (2007) about `mixlogit` for fitting mixed Logit models, and Plum (2016) about `bireprob` for fitting bivariate random-effects probit models.

There is no previous Stata implementation of MT's MSL estimators for the bivariate mixed Poisson regression model we know of. (JM used Matlab.) MT also discuss and fit the Marshall and Olkin (1985) bivariate negative binomial count data model, which is a 'one factor' version of MT's 'two factor' model. There is a single random effect entering both conditional mean functions, assumed to be Gamma distributed which in turn implies that there is a closed form expression for the likelihood. Xu and Hardin (2016) provide a Stata implementation of the Marshall-Olkin model, `bivcnto`, and we compare estimates derived with it with those derived using `bimpoisson` in Section 5.

In section 2, we outline the bivariate mixed Poisson model and its estimation using MSL, drawing heavily on MT's exposition. Section 3 introduces the `bimpoisson` command, including estimation and post-estimation options. Section 4 reports on the statistical performance of our estimators using Monte-Carlo simulation analysis, also reviewing previous work. Section 5 contains two empirical applications and our conclusions are in section 6. We provide guidance to users about empirical implementation choices such as type of MSL estimator, and type and number of draws.

2 The bivariate mixed Poisson regression model

`bimpoisson` fits bivariate mixed Poisson regression models by MSL following the approaches proposed by Munkin and Trivedi (1999). To quote MT, p. 35,

> [w]e wish to to specify a general bivariate count model in which overdispersion comes through bivariate unobserved heterogeneity which affects the conditional mean functions. Assume that the two count variables y_1 and y_2 have a Poisson distribution conditional on the σ-fields generated by the ran-

dom variables of unobserved heterogeneity ε_1, ε_2 and the set of independent explanatory variables x_1 and x_2:

$$y_j|x_j, \varepsilon_j \sim P(\mu_j), \mu_j > 0, j = 1, 2,$$

where we parameterize the mean parameters μ_1 and μ_2 as $\exp(x_1'\beta_1 + \varepsilon_1)$ and $\exp(x_2'\beta_2 + \varepsilon_2)$, respectively. The unobservable variables ε_1 and ε_2 are assumed jointly normally distributed, i.e.

$$(\varepsilon_1, \varepsilon_2) \sim N[(0, 0), (\sigma_1^2, \rho\sigma_1\sigma_2, \sigma_2^2)], \quad 0 \le |\rho| \le 1.$$

These assumptions imply that the expression for the joint distribution of y_1 and y_2, conditional on the given set of x_1 and x_2, $\Pr(y_1, y_2|x_1, x_2)$, does not have a closed form. However the model parameters can be estimated by MSL, in which case it is more convenient to exploit the Cholesky decomposition and sample ν_1 and ν_2 from the standard normal distribution and thence derive ε_1 and ε_2 as

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}.$$

The MSL sample log-likelihood function is then (MT, eqn. (11)):

$$\sum_{i=1}^{N} \log \left\{ \frac{1}{S} \sum_{s=1}^{S} \Pr(y_{1i}|x_{1i}, \sigma_1\nu_{1i}^s) \Pr(y_{2i}|x_{2i}, \sigma_2\rho\nu_{1i}^s + \sigma_2\sqrt{1-\rho^2}\nu_{2i}^s) \right\}$$

Maximization of this expression produces estimates asymptotically equivalent to exact ML estimates if $S, N \to \infty$ with $\sqrt{N}/S \to 0$, and can be implemented using `bimpoisson`'s `alt` option. Usage is subject to an important caveat, however, i.e., that $|\rho|$ is not too close to one. As MT explain, the derivative of the sum in the log-likelihood function with respect to ρ involves a term that is not bounded when $|\rho| = 1$, and the second derivative with respect to ρ is also unbounded. Hence, if using standard gradient-based methods to maximize the simulated log-likelihood, as with Stata's `maximize`, and $|\rho|$ is close to one, estimates are unreliable or infeasible. Symptoms may include `r(430) convergence not achieved` or repeated consecutive `(backed up)` messages in the iteration log.

To address this maximization feasibility issue, MT propose a sampling function approach. (This is an application of 'importance sampling', about which see Cameron and Trivedi (2005, Section 12.7.2) and Train (2003, Section 9.2.7).) Denote the sampling function by $g(\varepsilon_1, \varepsilon_2)$. It implies a revised MSL log-likelihood function given by:

$$\sum_{i=1}^{N} \log \left\{ \frac{1}{S} \sum_{s=1}^{S} \Pr(y_{1i}|x_{1i}, \varepsilon_{1i}^s) \Pr(y_{2i}|x_{2i}, \varepsilon_{2i}^s) \frac{f(\varepsilon_{1i}^s, \varepsilon_{2i}^s)}{g(\varepsilon_{1i}^s, \varepsilon_{2i}^s)} \right\}$$

where ε_{1i}^s and ε_{2i}^s are drawn from $g(\varepsilon_{1i}^s, \varepsilon_{2i}^s)$ for each observation i. The neat thing about this approach is that the relevant sum in the MSL log-likelihood is now bounded even if $|\rho| = 1$ and the first and second derivatives of the sum with respect to ρ are also bounded, so maximization of this MSL function converges to the maximum likelihood solution when the relevant conditions on S and N are satisfied (see above), and the covariance matrix of the estimates is consistently estimated by the robust sandwich variance estimator.

How should the sampling function be chosen? MT explain (p. 34) that: (i) it has to be easy to take draws from; (ii) have the same support as density $f(\varepsilon_{1i}^s, \varepsilon_{2i}^s)$ and integrate to one; and (iii) it has to be straightforward to evaluate the inner sum (over S) in the modified log-likelihood expression, and be bounded and smooth in the parameters. MT (p. 36) propose the standard bivariate normal density as a sampling function as a "natural choice" and use it in their empirical application; it satisfied all three conditions. However, a problem with their choice arises if the true values of σ_1 and σ_2 are larger than one. As JM, p. 63, explain, "MSL estimation becomes a challenge" in this situation because a standard bivariate normal distribution under-samples tail values of the true distribution. Hence, like JM, we use a sampling distribution that is bivariate normal with means equal to zero, but with standard deviations unrestricted. JM argue that, with their modification, drawn values provide more effective coverage and improve MSL's estimation of the double integral that specifies the bivariate density of the bivariate mixed Poisson model.

Taking the same idea further, we allow users to set the sampling function's correlation, using `bimpoisson`'s `rho_sf(#)` option (# defaults to zero). If the true ρ is large in magnitude, it is more effective in terms of coverage to concentrate draws in the more important areas of the bivariate normal space. Because this sampling function modification has not been used before, as far as we know, we examine its properties using Monte Carlo analysis.

Like JM, we reparameterize the model, assuming $\varepsilon_{1i} = \sigma_1 \upsilon_{1i}$ and $\varepsilon_{2i} = \sigma_2 \upsilon_{2i}$, hence also rewriting μ_{1i}^s and μ_{2i}^s as $\exp(x_{1i}'\beta_1 + \sigma_1 \upsilon_{1i}^s)$ and $\exp(x_{2i}'\beta_2 + \sigma_2 \upsilon_{2i}^s)$, respectively. For notational convenience, define $F = \log f(\varepsilon_{1i}^s, \varepsilon_{2i}^s)$ and $G = \log g(\varepsilon_{1i}^s, \varepsilon_{2i}^s)$ and observe that the final term in the MSL log-likelihood function with sampling function can be written $\exp(F - G)$. Then, in our parameterization, we have

$$F = -\log[2\pi\sigma_1\sigma_2(\sqrt{1-\rho^2})] - 0.5[(\upsilon_1^s)^2 - 2\rho\upsilon_{1i}^s\upsilon_{2i}^s + (\upsilon_{2i}^s)^2]/(1-\rho^2)$$

and

$$G = -\log[2\pi\sigma_1\sigma_2(\sqrt{1-\rho_{\mathrm{SF}}^2})] - 0.5[(\upsilon_1^s)^2 - 2\rho_{\mathrm{SF}}\upsilon_{1i}^s\upsilon_{2i}^s + (\upsilon_{2i}^s)^2]/(1-\rho_{\mathrm{SF}}^2)$$

where ρ_{SF} is the sampling function correlation. The overall sample log-likelihood and first and second derivatives are bounded in this case, for the reasons MT and JM set out. If one assumes $\rho_{\mathrm{SF}} = 0$, the reparameterized MSL log-likelihood is the same as that set out in JM's equation (5). Parameters σ_1, σ_2 and ρ are estimated jointly with the regression parameters.

MT (section 4.2, pp. 33–34 and 37–38) also propose a first-order method for correcting asymptotic bias in the MSL estimator, adopting the approach proposed by Gouriéroux and Monfort (1991). If $\hat{f}$ is an unbiased simulator of the true density f, the MSL estimator is inconsistent because $\log(\hat{f})$ is a biased estimator of $\log(f)$. With a second-order Taylor series expansion of $\log(\hat{f})$ around $\log(f)$ and an integration out, one derives an approximation that (partially) corrects for the bias in the parameter estimates by adding an adjustment term to the MSL function. To implement this method with `bimpoisson`, use the `bias` option.

A final and important issue is derivation of the bivariate distribution of draws to implement the MSL estimators. By default and following MT, `bimpoisson`'s two MSL estimators use pseudorandom numbers (draws from the standard uniform distribution) when simulating probabilities. We also optionally allow antithetic acceleration. This is a variance reduction technique according to which negatively correlated draws are used instead of independently distributed random numbers. See Train (2003, Section 9.3.1). We follow MT (section 4.3, pp. 38–39) and create one antithetic draw for each draw by reversing the sign of both of the two elements. To implement antithetic acceleration, use the `antithetic` option. The total number of draws in this case is $2S$.

Optionally, Halton draws can be used instead. (JM use Halton draws.) According to Train (2003, p. 231), "[t]he superior coverage and the negative correlation over observations that are obtained with Halton draws combine to make Halton draws far more effective than random [uniform] draws for simulation". Train (2003, p. 230) also states that "[d]iscarding the initial part of the sequence eliminates the correlation [between the initial terms of two Halton sequences]. The number of initial elements to discard needs to be at least as large as the largest prime that is used in creating the sequences". By default, `bimpoisson`'s `burn(#)` option uses $\# = 5$, noting that our sequences are created using the first two primes (2, 3). We allow users to combine Halton draws with antithetic acceleration but we have found little research that investigates the properties of such combination. (One exception is Sándor and András (2004).) Indeed, Train (2003, p. 222) states that

> [t]he randomness of simulation draws is a necessary component in the derivation of the asymptotic properties of the simulation-based estimators ... The asymptotic distributions become relatively accurate only when enough random draws have been taken. Therefore, for a given total number of draws, the goal of better coverage ... needs to be traded off against the goal of having enough randomness for the asymptotic formulas to apply .... The same issue applies to the antithetics discussed earlier.

We read this as warning that combination of antithetic acceleration with Halton draws (which are not random in the way uniform draws are) may compromise the goal of having enough randomness. In preliminary Monte Carlo analyses not reported here, we found there were no advantages for the combination – run times were longer but there was no general reduction in parameter bias.

In `bimpoisson`, the default number of simulations $S = 10$. This is a very low value and should only be used for checking syntax. For serious modelling, the number of simulations should be very much larger, as we illustrate in later sections. We recommend that users, having fitted a model with some specific number of simulations, increase the number systematically to check that estimates of the mixture distribution parameters in particular are robust.

On the basis of their Monte Carlo analysis of the MSL estimator using the sampling function approach, MT (p. 40) report that "[t]he first-order correction of the asymptotic bias significantly reduces the number of simulations needed for the estimation up to a given degree of bias. First-order corrected SML together with antithetic sampling produces most efficient estimates." JM provide further analysis of potential bias and how it varies with the true values of the mixture distribution parameters. We report some new Monte Carlo analysis of estimator performance in Section 4 to complement the analysis of MT and JM, which we also review.

2.1 Post-estimation calculations

Our discussion of post-estimation calculations is based on MT, pp. 43, 45–46. For outcome $j = 1, 2$, the expected count for observation i in the bivariate mixed Poisson model is given, after integrating out the random effect, by

$$\mu_{ji} = \exp(x'_{ji}\beta_j + \sigma_j^2/2)$$

`bimpoisson postestimation` calculates these objects by replacing β_j and σ_j^2 by their estimated counterparts. Expected counts conditional on the random effect are given by $\exp(x'_{ji}\beta_j)$.

Letting $k_j = 0, 1, 2, ..., max\_count$ be a cell frequency for outcome j, the cell frequencies are calculated, by simulation, from

$$\Pr(k_j|x_{ji}) = \left(\frac{1}{S}\right)\sum_{s=1}^{S}\Pr(k_j|x_{ji},\ \varepsilon_{ji}^s)$$

where

$$\Pr(k_j|x_{ji},\ \varepsilon_{ji}^s) = \exp(-\hat{\mu}_{ji})\hat{\mu}_{ji}^{k_j}/k_j!$$

To calculate the probability, the ε_{ji}^s are drawn from $N(0, \hat{\sigma}_j^2)$, $j = 1, 2$. We use `rnormal()` and set $S = 100$ by default (allowing users to choose different values).

The joint frequency $(y_1, y_2) = (k_1, k_2)$ is calculated from

$$\Pr(k_1, k_2|x_{ji}) = \left(\frac{1}{S}\right)\sum_{s=1}^{S}\Pr(k_1|x_{ji},\ \varepsilon_{1i}^s)\Pr(k_2|x_{2i},\ \varepsilon_{2i}^s)$$

To calculate the joint probability, $\varepsilon_{1i}^s, \varepsilon_{2i}^s$ are drawn from a bivariate normal distribution with zero means, standard deviations $\hat{\sigma}_1, \hat{\sigma}_2$, and correlation $\hat{\rho}$. We use `drawnorm()` for this, again setting $S = 100$ as the default for the simulation.

3 The `bimpoisson` command

The `bimpoisson` command fits the bivariate mixed Poisson regression model described in Section 2. `bimpoisson postestimation` provides post-estimation estimates of predicted counts and probabilities. This section describes command syntax, options, and saved results.

3.1 Syntax

There are two syntaxes for `bimpoisson`. If the predictors are the same in each equation:

`bimpoisson` *depvar1 depvar2* [*varlist*] [*if*] [*in*] [*weight*] [`,` *options*]

If the predictors differ:

`bimpoisson` *equation1 equation2* [*if*] [*in*] [*weight*] [`,` *options*]

where *equation1* and *equation2* are specified as

([*eqname:*] *depvar* [=] [*varlist*] [, *noconstant offset(varname)*])

varlist may contain factor variables; see `help fvvarlist`. *depvar1*, *depvar2*, *varlist*, and *depvar* may contain time-series operators; see `help tsvarlist`. `by` and `collect` are allowed; see `help prefix`. `pweights`, `fweights`, `aweights`, and `iweights` are allowed; see `help weight`.

The second syntax is likely rarely used. If one is interested in the conditional means of the outcomes, the conditioning variables should be the same unless extra restrictions are being imposed.

3.2 Options

Main

`nsim(`*#*`)` specifies the number of simulations, S. The default value (10) should only be used for syntax checking, and a much larger number should be used for estimation, as discussed elsewhere in this article.

`antithetic` specifies that antithetic draws be used, in which case the total number of simulations is $2S$.

`halton` specifies use of Halton draws (rather than the default which is standard uniform pseudorandom draws).

`burn(`*#*`)` specifies the number of elements dropped when creating sequences of Halton draws (default is 5).

`bias` specifies that the first-order bias correction be used.

`seed(`#`)` sets the seed for reproducibility; relevant if using standard uniform pseudorandom draws.

`rho_sf(`#`)` sets the sampling function correlation (# must be between −1 and 1). See text discussion about using values other than the default value (0).

`alt` specifies use of the standard MSL estimator as alternative to the default estimator which uses a sampling function.

`rng0` sets the random number generator ('rng') state back to its pre-estimation value; relevant if seed not set and uniform draws are used.

Model

`nonconstant` suppresses the constant term.

`offset1(`*varname*`)` specifies an offset variable for the first equation.

`offset2(`*varname*`)` specifies an offset variable for the second equation.

`constraints(`*constraints*`)` applies specified linear constraints.

SE/Robust

`vce(`*vcetype*`)` sets the vcetype, which may be `oim`, `robust`, `cluster` *clustvar*, `opg`, `bootstrap`, or `jackknife`. `robust` is recommended.

Reporting

`irr` reports incidence rate ratios.

`eform` reports exponentiated coefficients.

`level(`#`)` sets the confidence level; default is `level(95)`.

`nocnsreport` ensures that constraints are not displayed.

`display_options` are the standard options for controlling columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling.

Maximization

`maximize_options` control the maximization process: see `help maximize`. These are seldom used, with the main exception concerning `from()`. Suppose you fit a bivariate mixed Poisson regression model with number of draws S_1 and then wish to fit the same model with a larger number of draws $S_2 > S_1$. After fitting the first model, type something like `matrix(b1) = e(b)`, and then use the option `from(b1)` when fitting the second model.

`collinear` keeps collinear variables.

`coeflegend` displays legend instead of statistics.

3.3 Saved results

In addition to the usual results saved after `ml`, bimpoisson also saves the following in `e()`:

Scalars	
`e(nsims)`	number of simulations, S
`e(sigma1)`	estimate of sigma1
`e(sigma1_se)`	estimate of the standard error of sigma1
`e(sigma2)`	estimate of sigma2
`e(sigma2_se)`	estimate of the standard error of sigma2
`e(rho)`	estimate of rho
`e(rho_se)`	estimate of the standard error of rho
`e(antithetic)`	1 if `antithetic` option used, 0 otherwise
`e(halton)`	1 if `halton` option used, 0 otherwise
`e(burn)`	value of # set by `burn(#)`
`e(bias)`	1 if `bias` option used, 0 otherwise
`e(rho_sf)`	value of # set by `rho_sf(#)` option
`e(rng0)`	1 if `rng0` option used, 0 otherwise
Macros	
`e(seed)`	contains the seed if set; “seed not set” otherwise

3.4 `bimpoisson` postestimation

`predict` is available after fitting a model with `bimpoisson` for deriving multiple statistics. The calculations available are based on the expressions shown in section 2.1.

Syntax

`predict` [*type*] *newvar* [*if*] [*in*] [, *statistic nooffset*] [*options*]

Statistics

`xb1` linear prediction for equation 1.

`xb2` linear prediction for equation 2.

`stdp1` standard error of the linear prediction for equation 1.

`stdp2` standard error of the linear prediction for equation 2.

`n1` count prediction for equation 1 (with random effect integrated out).

`n2` count prediction for equation 2 (with random effect integrated out).

`pr1(#)` predicted prob(outcome 1 = #); # is integer-valued. Calculated using simulation (see Options below).

`pr2(#)` predicted prob(outcome 2 = #); # is integer-valued. Calculated using simulation (see Options below).

`pr12(#1 #2)` predicted prob(outcome 1 = #1, outcome 2 = #2); #1, #2 are integer-valued. Calculated using simulation (see Options below).

Options

`nsim(#)` sets number of simulations used to calculate predicted probabilities (default = 100).

`seed(#)` sets the random seed used when calculating predicted probabilities. (If not set, defaults to current rng state.)

`nooffset` is relevant only if you specified `offset()` when you fit the model. It modifies the calculations made by predict so that they ignore the offset variable; the linear prediction is treated as $x'b$ rather than $x'b + offset$.

4 Monte Carlo analyses

This section presents our Monte Carlo analyses, intended to complement MT's while also providing a bridge to the extensive recent analysis by JM, which we review. MT focused on three issues: how the statistical performance of the estimates varies with S, the use of the first-order bias correction, and antithetic acceleration. MT demonstrate the important bias-reducing effect of both the first-order correction and use of antithetic draws in conjunction with uniform draws. They also found that model parameters other than ρ were generally unbiased regardless of the choice of S. In addition, results depended on the magnitude of ρ rather than its sign: MT report similar performance statistics for $\rho = 0.6$ and $\rho = -0.6$. The bias-reducing impact of the first-bias similarities and 'symmetry' results are also apparent in JM's simulation results.

Our own Monte Carlo analysis complements MT's. For example, we explicitly consider estimator performance for a case of ρ close to 1 (0.9), not only a mid-range value (0.6), though we retain their assumption that $\sigma_1 = \sigma_2 = 0.5$ throughout. We contrast use of the standard MSL estimator and the estimator with sampling function for the $\rho = 0.6$ case (both are valid in principle), and uniform draws combined with antithetic draws on the one hand and Halton draws on the other hand. These features also distinguish our analysis from JM's.

JM's article is entirely devoted to Monte Carlo analysis of estimator performance for the bivariate mixed Poisson regression model (and also for a random parameter bivariate normal regression model). Distinctive features include an almost exclusive focus on Halton draws ($S = 250, 500, 1000$), a modified sampling function (as discussed earlier), and consideration of bias for many more combinations of mixture parameters than considered by MT and us. (For example, they consider values of σ_1, σ_2 ranging from 0.05 to 1.25 and ρ ranging from -1 to 1.) The sheer number of combinations considered, and reliance on graphical summaries of results, means that it is hard to digest and summarize their findings, especially since bias magnitudes are contingent on specific combinations of mixture parameters. That said, JM's headline finding is that serious biases can arise. The worst biases appear to be when the mixture standard deviations are small. When σ_1 is distinctly less than 0.5, ρ is overestimated (bias greater the smaller ρ is): see JM's Figures 3, A5, and A6. Also, second, σ_1, σ_2 are overestimated if true $\sigma_1 < 0.3$ and $\rho \leq 0.5$ (bias greater, the smaller σ_1, σ_2 are): see JM's Figures 5, A9, A10. Third, these same figures suggest that $\hat{\sigma}_1, \hat{\sigma}_2$ are upwardly biased if the true standard deviations are sigificantly greater than one. As it happens, we do not find small standard deviations in our two empirical applications; instead, we find $\hat{\sigma_1}, \hat{\sigma_2} > 1.3$ in both cases. We return to the bias issue below. Another conclusion we draw from JM's research, which they do not state explicitly, is that, although increasing S reduces bias (as expected), the number of draws may need to be very large to reduce bias to acceptable levels. We return to this issue below also.

MT, JM, and we all use the same basic data generating process for the non-mixture components, as described by MT, p. 40. There are 1000 observations on two explanatory variables x_1 and x_2 drawn independently from $N(0, 1/16)$, two unobserved heterogeneity terms ε_1 and ε_2, normally distributed as $N[(0,0), (\sigma_1^2, \rho\sigma_1\sigma_2, \sigma_2^2)]$, and two Poisson distributed count outcome variables y_1 and y_2 that are distributed $y_j \sim P(\mu_j)$, where $\mu_j = \exp(\alpha_j + x_j\beta_j + \varepsilon_j), j = 1, 2$. The true values of the parameters are taken to be $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$ throughout. JM also briefly consider a data-generating process in which explanatory variables were generated from asymmetric rather than symmetric distributions. This led to smaller biases.

We use 250 Monte Carlo replications, so there is less Monte Carlo variability than in MT's and JM's experiments (based on 100 and 200 replications respectively). We calculated performance statistics for all parameters using `simsum` (White 2010), but report only those for ρ because all other parameters were unbiased across our experiments (as MT also found). We derived all estimates using the first-order bias correction and the robust variance estimator. To save computing time, true parameter values were used as starting values.

4.1 Estimator performance, by estimator, ρ, and S

Table 1 shows results for various combinations of MSL estimator (with the sampling function (SF) or the alternative (Alt)), different types of draws, and S ranging from 50 through to 400. There are no Alt estimates for $\rho = 0.9$ given the problems cited earlier. Within each block of estimates, mean_rho refers to the mean of the estimated ρ across the 250 replications; bias arises if this differs from the true ρ. The other rows in each block show the standard deviation of the estimated ρ (SE_empirical) and the average of the model-estimated standard errors (SE_model). If corresponding estimates are approximately the same, that is evidence that model standard errors perform well. Reassuringly, we can see this is the case throughout the table, underlining the relevance of using the robust variance estimator as recommended. Throughout the table, the numbers in parentheses are Monte Carlo standard errors calculated by `simsum` using the formulae shown in White (2010).

The estimates in the first two columns of Table 1 are derived using the standard MSL estimator but with different draw types (and number). The table shows that the Alt estimator yields approximately unbiased estimates with $S = 200$. Remember that for the UA case, this means 200 uniform draws plus 200 antithetic draws, i.e., 400 in total, and so computing time is correspondingly longer (more on this later). Increasing S from 200 to 400 when $\rho = 0.6$ makes little difference to bias indicating that the asymptotic properties of MSL are kicking in.

The third and fourth columns of Table 1 provide estimates derived using MT's preferred set of choices, i.e., those shown in the final column of their Table 1 ($\rho = 0.6$, SF estimator with bias correction, UA draws). MT report mean_rho = 0.607 for $S = 50$. We are unable to match that performance with those choices unless S is much larger (at least 200) and we use Halton draws. We cannot explain the differences in results – they are greater than one might expect from using different seeds – though we note that our findings are more consistent with JM's. A potential source of non-comparability is the difference between MT's and JM's and our sampling function definitions. However, in initial code development, and before we became aware of JM's article, we used MT's sampling function definition, and also could not match the performance MT report using their preferred estimation choices.

All in all, these simulations suggest that for a mid-range value of ρ such as 0.6, and for 1000 observations, there is minimal bias if $S \geq 200$ with either the Alt estimator with UA or Halton draws or the SF estimator with Halton draws. Since Halton draws are more economical in terms of run time, the Monte Carlo analysis suggests they should be used.

The case of $\rho = 0.9$ is quite different: look at the fifth and sixth columns of estimates in Table 1 based only on the MSL SF estimator for the reasons discussed earlier. Estimates using UA draws are more biased than those using Halton draws for all values of S shown (though the gap narrows as S increases) – so the recommendation to use Halton draws continues to stand. Observe that even when 400 Halton draws are used, there remains a small downward bias in $\hat{\rho}$: mean_rho = 0.880. We also ran simulations for the $\rho = 0.9$ case using 1,000 and 2,000 Halton draws (not reported in Table 1), resulting in mean_rho estimates of 0.888 (MCse = 0.004) and 0.889 (MCse = 0.003) respectively. The take-away is that, when ρ is close to 1, the number of draws required to yield unbiased estimates may be very large. (This is also suggested by JM's Figures 4, A7, A8.) Another option might be to modify the sampling function correlation, and we explore this next.

Table 1: Monte Carlo analysis of ρ parameter, by ρ, MSL method, draws type, and number of draws

	$\rho = 0.6$				$\rho = 0.9$	
	Alt		SF		SF	
	UA	H	UA	H	UA	H
			$S = 50$			
mean_rho	0.623	0.588	0.518	0.572	0.748	0.805
	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)
SE_empirical	0.076	0.073	0.071	0.073	0.058	0.059
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
SE_model	0.088	0.075	0.071	0.080	0.057	0.063
	(0.004)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
			$S = 100$			
mean_rho	0.610	0.591	0.546	0.591	0.823	0.829
	(0.005)	(0.004)	(0.004)	(0.005)	(0.003)	(0.003)
SE_empirical	0.072	0.065	0.065	0.072	0.048	0.055
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
SE_model	0.075	0.071	0.070	0.074	0.054	0.064
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.007)
			$S = 200$			
mean_rho	0.605	0.596	0.568	0.599	0.823	0.853
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
SE_empirical	0.068	0.067	0.070	0.069	0.048	0.050
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
SE_model	0.071	0.070	0.069	0.073	0.054	0.060
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)
			$S = 400$			
mean_rho	0.603	0.596	0.581	0.601	0.850	0.880
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)
SE_empirical	0.065	0.066	0.069	0.067	0.053	0.056
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)
SE_model	0.070	0.067	0.069	0.071	0.053	0.057
	(0.001)	(0.000)	(0.001)	(0.001)	(0.002)	(0.002)

NOTES: rho_sf = 0. UA = "uniform draws + antithetic draws" (total number of draws = $2S$). H = Halton draws. Alt: MSL. SF: MSL with sampling function. First-order bias correction applied. Robust standard errors. 250 Monte Carlo repetitions. Numbers in parentheses are Monte Carlo standard errors. Mean_rho: mean of estimated ρ. SE_empirical: standard deviation of estimated rho. SE_model: average of model-estimated SEs.

4.2 Estimator performance when ρ_{SF} is varied

Table 2 reports performance statistics for different values of ρ_{SF} ('rho_sf'), with all calculations based on the MSL SF estimator with Halton draws, first-order bias correction, and robust standard errors. As in Table 1, $\rho = 0.6$ or 0.9 and S ranges from 50 to 400. The two rho_sf = 0 columns were already shown in Table 1 and are repeated to ease comparisons within Table 2.

Table 2 shows that for both values of ρ and all S values, increasing ρ_{SF} above zero leads to an increase in mean_rho. This means that less biased results may be possible with fewer draws than required when $\rho_{\mathrm{SF}} = 0$, For example, for $\rho = 0.6$, Table 1 indicated that 200 Halton draws were required; Table 2 suggests that less biased estimates may be possible when $S = 50$ with $\rho_{\mathrm{SF}} = 0.1$, (Estimates of other parameters remain unbiased.) With $\rho = 0.9$ there are similar pay-offs but, as before, the required S is greater than for the $\rho = 0.6$ case. Table 1 indicated that more than 400 Halton draws were required when ρ was large, but Table 2 shows that less biased estimates may be possible with only $S = 100$ if $\rho_{\mathrm{SF}} = 0.3$ or with $S = 200$ if $\rho_{\mathrm{SF}} = 0.2$. Thus, using small values of ρ_{SF} may be a fruitful way to derive less biased estimates using a smaller number of draws that one might otherwise have used, hence also saving computing time.

The problem is that it is not obvious ex ante which ρ_{SF} values will deliver unbiased estimates of ρ for any given S, and there can be downsides to making the wrong choice. Table 2 shows that if too large a ρ_{SF} value is used, one may derive upwardly-biased estimates of ρ. For example, with $\rho = 0.9$, $S = 400$, and $\rho_{\mathrm{SF}} = 0.3$, mean_rho = 0.924, an overshoot.

Our findings lead us to recommend that, when ρ is relatively large in magnitude, users can experiment with a small number of values of $\rho_{\mathrm{SF}} \neq 0$. Check whether these lead to marked changes in $\hat{\rho}$ relative to the default $\rho_{\mathrm{SF}} = 0$ case. If so, this is a signal that the number of draws does need to be increased substantially. We illustrate this idea in our application using MT's data that follows.

5 Two applications

We provide two applications employing data used in previously published articles, Xu and Hardin (2016) and Munkin and Trivedi (1999), respectively. The first application is a case when ρ is mid-range, nowhere near its upper bound, and compares estimates from standard MSL and MSL with a sampling function, and use of Halton draws versus uniform draws with antithetic acceleration. The second application is a case when ρ is close to one, so we only apply MSL with a sampling function. In both applications, we show the consequences for estimates of choosing different numbers of draws.

5.1 Application using the Xu-Hardin data

We reanalyze the data used by Xu and Hardin (2016) and available in the package accompanying their *Stata Journal* article (`search bivcnto`). There are 3,874 observa-

Table 2: Monte Carlo analysis of ρ parameter, by rho_sf(#), ρ, and S

	$\rho = 0.6$				$\rho = 0.9$			
rho_sf	0	0.1	0.2	0.3	0	0.1	0.2	0.3
$S = 50$								
mean_rho	0.572	0.599	0.620	0.633	0.805	0.837	0.864	0.890
	(0.005)	(0.005)	(0.005)	(0.006)	(0.004)	(0.003)	(0.003)	(0.003)
SE_empirical	0.073	0.077	0.085	0.095	0.059	0.054	0.048	0.040
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)
SE_model	0.080	0.086	0.122	0.116	0.063	0.062	0.058	0.067
	(0.002)	(0.002)	(0.022)	(0.007)	(0.002)	(0.003)	(0.003)	(0.012)
$S = 100$								
mean_rho	0.591	0.607	0.616	0.620	0.829	0.864	0.885	0.896
	(0.005)	(0.005)	(0.005)	(0.005)	(0.003)	(0.004)	(0.004)	(0.003)
SE_empirical	0.072	0.073	0.075	0.078	0.055	0.061	0.060	0.048
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
SE_model	0.074	0.078	0.086	0.086	0.064	0.071	0.060	0.067
	(0.001)	(0.002)	(0.004)	(0.002)	(0.007)	(0.010)	(0.003)	(0.011)
$S = 200$								
mean_rho	0.599	0.608	0.611	0.611	0.853	0.878	0.900	0.912
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)
SE_empirical	0.069	0.070	0.069	0.071	0.050	0.056	0.059	0.054
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
SE_model	0.073	0.076	0.078	0.079	0.060	0.058	0.056	0.056
	(0.001)	(0.002)	(0.001)	(0.001)	(0.004)	(0.003)	(0.003)	(0.004)
$S = 400$								
mean_rho	0.601	0.606	0.608	0.607	0.880	0.889	0.908	0.924
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)
SE_empirical	0.067	0.068	0.068	0.069	0.056	0.053	0.055	0.056
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
SE_model	0.071	0.072	0.073	0.073	0.057	0.054	0.089	0.056
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.023)	(0.004)

NOTES: MSL with sampling function, Halton draws, first-order bias correction, and robust standard errors. 250 Monte Carlo repetitions. Numbers in parentheses are Monte Carlo standard errors. Mean_rho: mean of estimated ρ. SE_empirical: standard deviation of estimated rho. SE_model: average of model-estimated SEs.

tions from German health registry data for 1984, and no missing values. The focus of the article was on fitting and comparing various bivariate regression models including some using a copula approach to model dependence and also the Marshall-Olkin one-factor model. Here we fit bivariate mixed Poisson regression models using the same data and make brief comparisons with the results from fitting other model specifications.

The two outcome variables are counts of numbers of doctor visits per year (`docvis`) and days spent in hospital per year (`hospvis`) and the explanatory variables describe marital status and sex, and whether has children, is working, and has education beyond high school level (all explanatory variables are binary indicators). Table 3 lists the outcomes, explanatory variables, and their means.

Table 3: Variables in the Xu-Hardin data

Variable name	Description	Mean
`docvis`	MD visits/year	3.163
`hospvis`	Days in hospital/year	0.121
`MarF`	1=married woman (reference category)	0.368
`MarM`	1=married man	0.421
`SinM`	1=single man	0.100
`SinF`	1=single woman	0.111
`kids`	1=have children; 0=no children	0.449
`outwork`	1=not working; 0=working	0.367
`postHS`	1=High school graduate or higher	0.186

First we read in the data, provide some descriptive statistics about the outcomes, and put the explanatory variable list into a global macro for later convenience. The Stata output that follows shows that the correlation between outcomes is relatively small (around 0.15). Both outcome variables are skewed. The median number of doctor visits is 1 (mean = 3) though 5% of the sample visit at least 12 times. Only around 7.7% of individuals spend any days in hospital but those that do spend up to 17 days (mean number is 0.12).

```
. use rwm1984.dta, clear
(German health data for 1984; Hardin & Hilbe, GLM and Extensions, 3rd ed)

. correlate docvis hospvis
(obs=3,874)

             |   docvis  hospvis
-------------+------------------
      docvis |   1.0000
     hospvis |   0.1458   1.0000


. su docvis hospvis, de

                        MD visits/year
-------------------------------------------------------------
      Percentiles      Smallest
```

```
 1%            0              0
 5%            0              0
10%            0              0       Obs               3,874
25%            0              0       Sum of wgt.       3,874

50%            1                      Mean           3.162881
                        Largest       Std. dev.      6.275955
75%            4             71
90%            8             72       Variance       39.38761
95%           12             80       Skewness       6.032647
99%           31            121       Kurtosis       65.70611

                    Days in hospital/year
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs               3,874
25%            0              0       Sum of wgt.       3,874

50%            0                      Mean           .1213216
                        Largest       Std. dev.       .695784
75%            0             14
90%            0             14       Variance       .4841154
95%            1             16       Skewness       14.62932
99%            2             17       Kurtosis       285.3169

. global xvars MarM SinM SinF kids outwork postHS
```

Next we fit bivariate mixed Poisson models to these data, so we have to make decisions about which estimator to use and how many draws. Noting the relatively low correlation between outcomes and some preliminary analyses – fitting models without explanatory variables – we ascertained that ρ was likely mid-range. Our Monte Carlo analysis (Table 1) indicates that in this case a key choice is the number of draws: unbiased estimates arose for samples with $N = 1,000$ of either 200 UA draws (400 draws in total) or 200 Halton draws with the alternative (standard) MSL estimator, or 200 Halton draws for MSL with the sampling function estimator. To choose the number of draws, MT suggest a practical heuristic: for a sample with N^* observations, the appropriate number of draws is $S^*\sqrt{(N^*/1000)}$, where S^* is the number of draws that the Monte Carlo analysis suggests yields unbiased parameter estimates. We have $N^* = 3,874$ and $S^* = 200$. The calculation implies $S = 394$ which we round up to 400. We also show estimates based on larger S.

The Stata output below shows the results from fitting a model using standard MSL, 400 UA draws, and (as in all specifications reported) with the first-order bias correction. A seed is specified to ensure reproducibility. The model took around 14 minutes to fit on a Dell Precision 3591 laptop with an Intel Ultra 9 185H 2.3Mhz chip. Specification choices are echoed in the output: at the top is the type of draws and at the bottom their number and whether antithetic draws and bias correction options are chosen. The estimation table shows the estimator type, which is "MSL" here (and would be "MSL with sampling function" if the `alt` option had not been specified). The mixture parameters are fitted in metrics that respect the constraints on their values and are then transformed back to their natural metric and reported at the bottom of the table. It is common to see repeated `(not concave)` messages when fitting these models but they are not a problem: we have found Stata's modified Newton-Raphson maximization algorthim to lead to convergence without problems. Also, we found no noticeable pay-off in computing time to using estimates from two independent Poisson models to provide starting values.

```
. bimpoisson docvis hospvis $xvars, robust nsim(400) bias alt ///
>         antithetic seed(13159)

Generating pseudorandom uniform draws ...

Iteration 0:  Log pseudolikelihood = -13419.598  (not concave)
Iteration 1:  Log pseudolikelihood =  -12022.94  (not concave)
Iteration 2:  Log pseudolikelihood = -11450.799  (not concave)
Iteration 3:  Log pseudolikelihood = -11071.132  (not concave)
Iteration 4:  Log pseudolikelihood = -10353.703  (not concave)
Iteration 5:  Log pseudolikelihood = -9825.9706  (not concave)
Iteration 6:  Log pseudolikelihood = -9726.7421  (not concave)
Iteration 7:  Log pseudolikelihood = -9725.1224  (not concave)
Iteration 8:  Log pseudolikelihood = -9647.9827  (not concave)
Iteration 9:  Log pseudolikelihood = -9614.5337
Iteration 10: Log pseudolikelihood = -9592.4443
Iteration 11: Log pseudolikelihood = -9584.8933
Iteration 12: Log pseudolikelihood = -9583.4684
Iteration 13: Log pseudolikelihood = -9583.3006
Iteration 14: Log pseudolikelihood = -9583.3003

Bivariate mixed Poisson regression (MSL)              Number of obs =  3,874
                                                      Wald chi2(6)  = 140.91
Log pseudolikelihood = -9583.3003                     Prob > chi2   = 0.0000

------------------------------------------------------------------------------
             |               Robust
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
docvis       |
        MarM |  -.2648561   .0778074    -3.40   0.001    -.4173559   -.1123564
        SinM |  -.4691794   .0950494    -4.94   0.000    -.6554728    -.282886
        SinF |   .0742453   .1320539     0.56   0.574    -.1845756    .3330661
        kids |  -.4425165    .062366    -7.10   0.000    -.5647515   -.3202815
     outwork |   .5105739   .0926181     5.51   0.000     .3290458     .692102
      postHS |  -.2619842   .0769608    -3.40   0.001    -.4128246   -.1111438
       _cons |   .3326537   .0733339     4.54   0.000      .188922    .4763855
-------------+----------------------------------------------------------------
hospvis      |
        MarM |   .4414582   .1752062     2.52   0.012     .0980604     .784856
        SinM |  -.2212477   .2987271    -0.74   0.459     -.806742    .3642467
        SinF |   .1689398   .2508584     0.67   0.501    -.3227337    .6606133
        kids |  -.1766938   .1586167    -1.11   0.265    -.4875769    .1341893
     outwork |   .4563236   .1616185     2.82   0.005     .1395573      .77309
      postHS |  -.4952979   .1829073    -2.71   0.007    -.8537896   -.1368062
       _cons |  -4.371475   .2342582   -18.66   0.000    -4.830613   -3.912338
-------------+----------------------------------------------------------------
    /ln_sig1 |   .4447017   .0193587    22.97   0.000     .4067592    .4826441
    /ln_sig2 |   .8123536   .0531459    15.29   0.000     .7081896    .9165177
       /arho |   .7033847   .0560584    12.55   0.000     .5935122    .8132572
-------------+----------------------------------------------------------------
```

```
      sigma1 |   1.560025   .0302001    51.66   0.000     1.501942    1.620353
      sigma2 |   2.253205   .1197486    18.82   0.000     2.030312    2.500568
         rho |   .6065118    .035437    17.12   0.000     .5324169    .6713832
------------------------------------------------------------------------------
Note: 400 draws plus 400 antithetic draws, with first-order bias correction
```

The estimates indicate that single and married men make fewer doctor visits than married women (the reference category) but married men make more hospital visits. Perhaps surprisingly, having children is associated with fewer doctor visits. Non-working individuals make more doctor and hospital visits than working individuals, whereas individuals with post-high school education make fewer visits of both types. The bottom rows of the table point to statistically significant unobserved heterogeneity with the random effect standard deviation greater for the hospital visit count than for the doctor visit count. There is a positive correlation between the random effects of 0.61.

Table 4 contrasts these estimates (shown in column (1)) with those of four other model specifications. Model 2 is the same as Model 1 except that Halton draws are used. This reduced run time substantially, to around 5 minutes, and the fit in terms of log pseudolikelihood and AIC is slightly better. Associations between explanatory variables and outcomes are broadly the same, especially if one focuses on the statistically significant estimates. The mixture parameters are noticeably smaller in magnitude in Model 2, especially $\hat{\rho}$ (0.46 rather than 0.61).

The remaining three columns of Table 4 are for models with $S = 2,000$, a five-fold increase intended to check robustness. Starting values were taken from models with $S = 400$ and deployed using the `from()` option. Model 3 took around 37 minutes to run, Model 4 around 16 minutes, and Model 5 around 19 minutes. Reassuringly, all three models provide broadly similar estimates of coefficients and mixture parameters, despite the differences in MSL estimator and draws type and number (remember Model 3 uses 4,000 draws in total). Model 4 fits a bit better than the other two (the log pseudolikelihood is slightly larger and the Akaike Information Criterion (AIC) value is slightly smaller).

Another issue to consider is potential bias. We reported earlier that JM's Monte Carlo analysis suggested that estimates of the random effect standard deviations were upwardly biased if the true values were significantly greater than one, which could be the case here (e.g, the estimates for Model 4 are 1.5 and 2.1). If there is upward bias, we would expect increasing S to lead to smaller standard deviation estimates, and this is the case with UA draws (Model 3 versus Model 1) but not for Halton draws (Model 4 versus Model 1), and so clear cut conclusions are hard to draw.

The estimates with $S = 2,000$ appear somewhat different from the estimates based on with $S = 400$. To control for differences in estimator, and draw type, it is best to constrast Models 1 and 3, or Models 2 and 4. Take the latter case, for example: $\hat{\rho}$ is around 0.46 for Model 0.46 but 0.54 for Model 4. It is up to users to decide how precise they require estimates to be – are differences such as these substantively relevant in the context of interest? Some checking of estimate sensitivity to choice of S is warranted but may not need to use as many as 2,000 draws. For example, refitting Model 4 with

Table 4: Bivariate mixed Poisson regression model estimates, Xu-Hardin data

	(1)	(2)	(3)	(4)	(5)
	Alt UA	Alt H	Alt UA	Alt H	SF H
S	400	400	2,000	2,000	2,000
docvis					
MarM	−0.265	−0.149	−0.304	−0.338	−0.376
	(0.078)	(0.089)	(0.077)	(0.066)	(0.060)
SinM	−0.469	−0.588	−0.434	−0.718	−0.740
	(0.095)	(0.090)	(0.103)	(0.079)	(0.069)
SinF	0.074	0.152	0.136	0.044	0.044
	(0.132)	(0.110)	(0.087)	(0.090)	(0.078)
kids	−0.443	−0.494	−0.377	−0.477	−0.500
	(0.062)	(0.078)	(0.057)	(0.050)	(0.050)
outwork	0.511	0.574	0.537	0.430	0.424
	(0.093)	(0.062)	(0.067)	(0.058)	(0.049)
postHS	−0.262	−0.144	−0.209	−0.189	−0.153
	(0.077)	(0.081)	(0.071)	(0.068)	(0.072)
_cons	0.333	0.278	0.282	0.421	0.455
	(0.073)	(0.068)	(0.073)	(0.059)	(0.051)
hospvis					
MarM	0.441	0.563	0.342	0.461	0.350
	(0.175)	(0.213)	(0.206)	(0.199)	(0.173)
SinM	−0.221	−0.254	−0.301	−0.295	−0.439
	(0.299)	(0.311)	(0.276)	(0.275)	(0.292)
SinF	0.169	0.234	0.174	0.231	0.151
	(0.251)	(0.248)	(0.225)	(0.227)	(0.241)
kids	−0.177	−0.068	−0.185	−0.116	−0.108
	(0.159)	(0.212)	(0.145)	(0.131)	(0.164)
outwork	0.456	0.714	0.611	0.533	0.497
	(0.162)	(0.180)	(0.181)	(0.176)	(0.156)
postHS	−0.495	−0.096	−0.283	−0.254	−0.237
	(0.183)	(0.351)	(0.192)	(0.195)	(0.267)
_cons	−4.371	−4.522	−4.400	−4.415	−4.486
	(0.234)	(0.305)	(0.253)	(0.252)	(0.291)
mixture					
sigma1	1.560	1.401	1.497	1.456	1.432
	(0.031)	(0.030)	(0.031)	(0.020)	(0.020)
sigma2	2.253	2.048	2.130	2.081	2.233
	(0.120)	(0.115)	(0.092)	(0.075)	(0.136)
rho	0.607	0.461	0.532	0.538	0.549
	(0.035)	(0.050)	(0.040)	(0.039)	(0.043)
LogPL	$-9,583$	$-9,575$	$-9,564$	$-9,551$	$-9,559$
AIC	19,201	19,185	19,162	19,136	19,153

NOTES: Alt: MSL. SF: MSL with sampling function. First-order bias correction applied. UA: S uniform draws + S antithetic draws. H: S Halton draws. Robust standard errors in parentheses. N = 3,874.

800 Halton draws produced estimates (not reported) quite similar to those reported for the original Model 4 in Table 4.

How do these bivariate mixed Poisson regression model estimates compare with those derived using alternative bivariate count data models? Xu and Hardin (2016) report estimates for six models derived using their `bivcnto` command applied to the same data and using the same sets of explanatory variables. Table 4's Models 4–6 have AIC values that are similar to the AIC for a model with the outcomes each specified as negative binomial and dependence summarized using a Gaussian copula (AIC = 19,160). Conclusions should be drawn cautiously because the comparison involves AIC values derived by different methods (MSL and exact ML). That said, the congruence is perhaps unsurprising since both models allow for over-dispersion. In contrast, a Marshall-Olkin one-factor negative binomial model has AIC = 19,659. The much poorer fit is unsurprising too, because the bivariate mixed Poisson model estimates show a correlation between random effects that is well below one and standard deviations of different sizes.

Given model estimates, one can also derive predicted counts and relative frequencies, as explained in Section 2. The Stata output below shows the post-estimation commands used to derive the predictions shown in Table 5. The predictions are based on the estimates of Model 4 in Table 4.

```
quietly {
 predict n1, n1
 predict n2, n2
 predict pr10, pr1(0)
 predict pr11, pr1(1)
 predict pr12, pr1(2)
 predict pr13, pr1(3)
 predict pr14, pr1(4)
 predict pr15, pr1(5)
 predict pr20, pr2(0)
 predict pr21, pr2(1)
 predict pr22, pr2(2)
 predict pr23, pr2(3)
 predict pr24, pr2(4)
 predict pr25, pr2(5)
 predict pr12_00, pr12(0 0)
 predict pr12_10, pr12(1 0)
 predict pr12_01, pr12(0 1)
 predict pr12_11, pr12(1 1)
}
su docvis hospvis
su n1 n2
ta docvis
su pr1?  // analogue to MT Table 5
ta hospvis
su pr2?  // analogue to MT Table 5
ta docvis hospvis, nokey freq cell
su pr12_??  // analogue to MT Table 7
```

The top panel of Table 5 shows the predictions for each outcome taken separately; the bottom panel shows the bivariate predictions. In each case, the 'fitted' (predicted) frequencies are compared with their 'empirical' (observed) counterparts. The table shows that the model is better at fitting the counts for `hospvis` than for `docvis`, but this is an easier task given the very large fraction of zero counts for the latter. The model overpredicts counts of one doctor visit, and this also affects the joint count predictions: observe the over-prediction of the relative frequency of the (1,0) counts, for example. The differences between fitted and empirical frequencies could be used as the basis for a chi-square goodness of fit test: see Cameron and Trivedi (2013, Section 5.3.4).

Table 5: Postestimation predictions from Xu-Hardin data (column %)

Counts	*docvis*		*hospvis*	
	empirical	fitted	empirical	fitted
0	41.6	37.5	92.3	92.3
1	11.6	20.0	6.1	5.4
2	11.4	11.3	1.1	1.2
3	9.1	7.1	0.2	0.4
4	5.5	4.8	0.1	0.2
5	4.3	3.4	0.1	0.2
other	16.5		0.1	
mean	3.16	3.60	0.12	0.05
(*docvis*,*hospvis*)				
(0,0)	40.1	37.0		
(1,0)	11.2	19.4		
(0,1)	1.3	1.0		
(1,1)	0.3	1.0		
other	47.1			

NOTES: Predictions from Model 4 in Table 4. Table entries show column percentages, except for ‘mean’ (count).

5.2 Application using the Munkin-Trivedi data

We now reanalyze the data used by Munkin and Trivedi (1999), previously used by Deb and Trivedi (1997). We downloaded the data from the *Journal of Applied Econometrics*’s archive, read them into Stata using `import delimited`, added labels, and saved the file in Stata format. The data refer to a sample of 4,406 elderly people who were respondents to the U.S. National Medical Expenditure Survey conducted in 1987 and 1988. There are no missing values.

Table 6 lists the variables in the dataset and their means. The two count outcomes are the number of emergency room visits (`emr`) and the number of hospital stays (`hosp`) during the survey reference period. There is a more extensive set of explanatory variables than in the Xu and Hardin (2016) data. Variables with descriptions including “1=” are binary indicator variables.

Table 6: Variables in the Munkin-Trivedi data

Variable name	Description	Mean
`emr`	# emergency room visits	0.264
`hosp`	# hospital stays	0.296
`exclhlth`	1= self-perceived health is excellent	0.078
`poorhlth`	1= self-perceived health is poor	0.126
`numchron`	# chronic health conditions	1.542
`adldiff`	1=has a condition that limits ADLs	0.204
`noreast`	1=lives in northeastern US	0.190
`midwest`	1=lives in the midwestern US	0.263
`west`	1=lives in the western US	0.181
`age`	age (years/10)	7.402
`black`	1=African American	0.117
`male`	1=male	0.404
`married`	1=married	0.546
`school`	# years of education	10.290
`faminc`	family income ($10,000)	2.527
`employed`	1=employed	0.103
`privins`	1=covered by private health insurance	0.776
`medicaid`	1=covered by Medicaid	0.091

We first read in the data, provide some descriptive statistics, and put all the explanatory variables into a global macro. The Stata output that follows shows that the bivariate correlation between count outcomes, 0.48, which is around three times the magnitude of the corresponding correlation in the Xu-Hardin data. This is an early signal that model fitting will be trickier with the Munkin-Trivedi data.

The `summarize` output shows that only around one-tenth of the sample visited an emergency room or had had a hospital stay. The mean counts are 0.26 and 0.29, respectively.

```
. use debtrivedi, clear

. correlate emr hosp
(obs=4,406)

             |      emr     hosp
-------------+------------------
         emr |   1.0000
        hosp |   0.4761   1.0000


. su emr hosp, de

                   # emergency room visits
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
```

```
 5%            0              0
10%            0              0       Obs               4,406
25%            0              0       Sum of wgt.       4,406

50%            0                      Mean           .2635043
                        Largest       Std. dev.      .7036586
75%            0              8
90%            1              8       Variance       .4951354
95%            1             11       Skewness       5.066106
99%            3             12       Kurtosis       49.81375

                       # hospital stays
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs               4,406
25%            0              0       Sum of wgt.       4,406

50%            0                      Mean           .2959601
                        Largest       Std. dev.      .7463978
75%            0              8
90%            1              8       Variance       .5571097
95%            2              8       Skewness       3.964427
99%            3              8       Kurtosis       25.80247

. global xvars exclhlth poorhlth numchron adldiff ///
>         noreast midwest west age black male ///
>         married school faminc employed privins medicaid
```

Seven sets of `bimpoisson` estimates are shown in Table 7. Because MT report $\hat{\rho} = 0.92$, we only used MSL with sampling function when fitting models to their data, as they do. The first column of Table 7, headed 'MT', contains the estimates reported in MT's Table 3, derived using UA draws and $S = 150$ (300 draws in total). Estimates reported in columns labelled 1–6 were derived using `bimpoisson`. Model 1 uses the same choices as MT; Model 2 is the same except that it switches to Halton draws, and these are used for Models 3–6 as well. Models 3 and 4 both increase S to 850, with the latter also setting $\rho_{\mathrm{SF}} = 0.2$ rather than the default value of 0. The final two models increase the number of draws to 2,500 and 5,000, respectively. Models 1 and 2 took less than three minutes to run. Models 3 and 4 took around 16 and 13 minutes respectively. Model 5 took around 25 minutes (starting values taken from Model 3 estimates) and Model 6 took around 61 minutes (starting values from Model 5 estimates).

MT's choice of $S = 150$ was based on rule-of-thumb calculations of the kind described in the previous subsection and their Monte Carlo analysis of the $\rho = 0.6$ case. Because our own simulation analysis for the $\rho = 0.9$ case led us to recommend at least 400 draws for 1,000 observations, our heuristic calculation for the Munkin-Trivedi data implies using $S = 850$. ($400\sqrt{(4406/1000)} = 839.62$ which we rounded up.)

Table 7: Bivariate mixed Poisson regression model estimates, Munkin-Trivedi data

	MT	(1)	(2)	(3)	(4)	(5)	(6)
	UA	UA	H	H	H	H	H
S	150	150	150	850	850	2,500	5,000
rho_sf(#)	0	0	0	0	0.2	0	0
emr							
exclhlth	−0.643	−0.785	−0.624	−0.725	−0.710	−0.668	−0.702
	(0.21)	(0.243)	(0.220)	(0.189)	(0.212)	(0.215)	(0.217)
poorhlth	0.518	0.490	0.503	0.609	0.515	0.540	0.481
	(0.11)	(0.103)	(0.113)	(0.125)	(0.122)	(0.105)	(0.113)
numchron	0.256	0.241	0.218	0.240	0.238	0.249	0.249
	(0.03)	(0.029)	(0.031)	(0.033)	(0.031)	(0.030)	(0.029)
adldiff	0.360	0.380	0.461	0.386	0.441	0.433	0.415
	(0.11)	(0.099)	(0.113)	(0.110)	(0.106)	(0.096)	(0.095)
noreast	0.050	0.045	0.080	0.044	0.022	0.071	0.060
	(0.12)	(0.120)	(0.124)	(0.109)	(0.113)	(0.116)	(0.109)
midwest	0.027	0.029	−0.021	0.029	0.013	0.013	0.002
	(0.11)	(0.104)	(0.104)	(0.105)	(0.103)	(0.102)	(0.103)
west	0.170	0.173	0.249	0.218	0.144	0.215	0.171
	(0.12)	(0.119)	(0.138)	(0.137)	(0.137)	(0.123)	(0.127)
age	0.108	0.116	0.126	0.164	0.173	0.104	0.137
	(0.06)	(0.070)	(0.071)	(0.068)	(0.069)	(0.063)	(0.063)
black	0.226	0.193	0.148	0.186	0.191	0.129	0.148
	(0.14)	(0.123)	(0.143)	(0.125)	(0.129)	(0.136)	(0.120)
male	0.109	0.105	0.074	0.072	0.165	0.095	0.121
	(0.09)	(0.097)	(0.102)	(0.093)	(0.091)	(0.096)	(0.095)
married	−0.099	−0.177	−0.120	−0.138	−0.138	−0.124	−0.103
	(0.09)	(0.101)	(0.107)	(0.091)	(0.095)	(0.100)	(0.097)
school	−0.006	−0.014	−0.021	−0.014	−0.012	−0.017	−0.011
	(0.01)	(0.011)	(0.014)	(0.013)	(0.014)	(0.011)	(0.012)
faminc	0.007	0.005	−0.003	−0.002	−0.001	−0.001	0.000
	(0.01)	(0.016)	(0.018)	(0.019)	(0.023)	(0.016)	(0.015)
employed	0.115	0.193	0.151	0.245	0.209	0.213	0.184
	(0.14)	(0.143)	(0.134)	(0.132)	(0.128)	(0.133)	(0.135)
privins	0.115	0.038	0.065	0.064	0.090	0.024	0.019
	(0.12)	(0.116)	(0.127)	(0.112)	(0.119)	(0.108)	(0.103)
medicaid	0.145	0.135	0.130	0.128	0.240	0.165	0.201
	(0.17)	(0.144)	(0.154)	(0.155)	(0.163)	(0.142)	(0.137)
_cons	−3.535	−3.522	−3.604	−3.991	−4.118	−3.510	−3.795
	(0.54)	(0.579)	(0.573)	(0.553)	(0.583)	(0.519)	(0.528)
hosp							
exclhlth	−0.722	−0.812	−0.679	−0.740	−0.804	−0.744	−0.785
	(0.18)	(0.211)	(0.191)	(0.192)	(0.212)	(0.188)	(0.190)
poorhlth	0.561	0.504	0.514	0.661	0.526	0.557	0.497
	(0.12)	(0.110)	(0.107)	(0.109)	(0.111)	(0.105)	(0.123)
numchron	0.297	0.281	0.257	0.280	0.270	0.285	0.281
	(0.03)	(0.027)	(0.025)	(0.030)	(0.028)	(0.026)	(0.026)

	MT	(1)	(2)	(3)	(4)	(5)	(6)
	UA	UA	H	H	H	H	H
S	150	150	150	850	850	2,500	5000
rho_sf(#)	0	0	0	0	0.2	0	0
adldiff	0.350	0.341	0.352	0.295	0.368	0.365	0.345
	(0.11)	(0.102)	(0.105)	(0.106)	(0.104)	(0.095)	(0.094)
noreast	0.010	0.044	0.067	0.042	0.002	0.057	0.041
	(0.13)	(0.117)	(0.118)	(0.110)	(0.115)	(0.114)	(0.109)
midwest	0.132	0.174	0.125	0.129	0.150	0.150	0.137
	(0.11)	(0.108)	(0.097)	(0.101)	(0.103)	(0.099)	(0.101)
west	0.204	0.186	0.278	0.213	0.123	0.203	0.153
	(0.12)	(0.116)	(0.122)	(0.114)	(0.118)	(0.116)	(0.135)
age	0.180	0.192	0.216	0.240	0.253	0.182	0.220
	(0.06)	(0.068)	(0.064)	(0.064)	(0.068)	(0.060)	(0.062)
black	0.141	0.106	0.112	0.117	0.113	0.046	0.069
	(0.15)	(0.124)	(0.132)	(0.120)	(0.133)	(0.132)	(0.123)
male	0.189	0.198	0.139	0.156	0.262	0.188	0.213
	(0.09)	(0.090)	(0.091)	(0.090)	(0.088)	(0.090)	(0.089)
married	−0.013	−0.105	−0.008	−0.024	−0.041	−0.023	−0.001
	(0.09)	(0.098)	(0.097)	(0.088)	(0.096)	(0.097)	(0.096)
school	0.006	0.001	−0.003	0.006	0.008	0.002	0.008
	(0.01)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)
faminc	−0.008	−0.001	0.002	−0.004	−0.002	−0.003	−0.001
	(0.01)	(0.013)	(0.013)	(0.014)	(0.020)	(0.014)	(0.013)
employed	−0.051	0.082	−0.013	0.117	0.075	0.076	0.042
	(0.41)	(0.150)	(0.139)	(0.136)	(0.133)	(0.136)	(0.138)
privins	0.216	0.213	0.214	0.175	0.245	0.182	0.180
	(0.13)	(0.117)	(0.122)	(0.112)	(0.129)	(0.116)	(0.113)
medicaid	0.158	0.175	0.200	0.115	0.276	0.200	0.238
	(0.19)	(0.138)	(0.146)	(0.140)	(0.169)	(0.143)	(0.144)
_cons	−4.440	−4.417	−4.585	−4.862	−5.039	−4.440	−4.759
	(0.54)	(0.570)	(0.530)	(0.514)	(0.593)	(0.505)	(0.520)
mixture							
sigma1	1.39	1.324	1.419	1.332	1.349	1.354	1.326
	(0.06)	(0.053)	(0.066)	(0.057)	(0.054)	(0.059)	(0.052)
sigma2	1.36	1.323	1.369	1.301	1.338	1.363	1.331
	(0.06)	(0.058)	(0.053)	(0.046)	(0.058)	(0.057)	(0.062)
rho	0.92	0.891	0.929	0.935	0.993	0.986	0.996
	(0.02)	(0.021)	(0.018)	(0.013)	(0.004)	(0.010)	(0.003)
LogPL	−5,290	−5,305	−5,252	−5,241	−5,224	−5,251	−5,243
AIC	10,658	10,685	10,577	10,557	10,522	10,577	10,561

NOTES: MSL with sampling function. First-order bias correction applied. UA: S uniform draws + S antithetic draws. H: S Halton draws. Robust standard errors in parentheses. $N = 4{,}406$. 'MT' column estimates: Munkin and Trivedi (1999, Table 3).

Table 7 reveals that coefficient estimates and mixture variances, and their statistical significance are much the same for all models. The poorer an elderly individual's health, whether self-assessed or in terms of measured conditions, the more visits there are to the emergency room and hospital. Adjusting for health characteristics, most other characteristics have no statistically significant association with either outcome. (One exception is that men have more hospital visits than women.) The standard deviations of the random effects are approximately equal (c. 1.3). What differs most across the models shown is $\hat{\rho}$, which is around 0.9 for the models with relatively small S, increasing to just under 1 for Models 4–6. Observe that using a sampling function obviates any maximization problems even though $\hat{\rho}$ is extremely large.

On the basis of our Monte Carlo analysis, the relatively small values of $\hat{\rho}$ from MT's and our Models 1 and 2 are consistent with downward bias, and that bias is reduced when $S = 850$. But our Monte Carlo analyses also led us to ask whether this number is sufficiently large.

To help answer this question, and following our earlier recommendation, we try a specification that is identical to Model 3 except that we use the `rho_sf(0.2)` option: see Model 4. (With `rho_sf(0.1)`, we got very similar estimates.) The effect is dramatic: $\hat{\rho}$ increases to over 0.99. This led us to refit models with substantially larger S and `rho_sf(0)`. Compared to Model 3, and as expected, $\hat{\rho}$ increases markedly, up to 0.996 for Model 6 with $S = 5,000$.

Because `bimpoisson` estimation constrains $|\rho| < 1$, we did not try raising S even further. Instead, we observe that the very similar estimates for random effect standard deviations $\hat{\sigma_1}, \hat{\sigma_2}$ and correlation $\hat{\rho}$ near one suggest that a one-factor model provides a more appropriate fit to the Munkin-Trivedi data than MT's two-factor model does. When we fitted the Marshall-Olkin negative binomial model using `bivcnto`, its AIC $(10,479)$ was smaller than all those shown in Table 7 and much the same as a negative binomial model with Gaussian copula for dependence (AIC $= 10,481$).

Finally, we show postestimation predictions based on Model 6: see Table 8. The underlying code and calculations are the same as used for the application to the Xu-Hardin data. By constrast with that application, the predicted relative frequencies are closer to the observed ones. For example, the fitted and empirical fractions for zero emergency room visits are both 82%; for zero hospital visits, they are 80% and 81%. The main differences between fitted and empirical counts are for the bivariate count frequencies: (1,0) and (0,1) are over-estimated and (1,1) under-estimated. MT's predictions are shown in their Tables 5 and 7 and are similar to those shown in Table 8. Consistent with the discussion in the previous paragraph, MT show that the Marshall-Olkin model implies predicted frequencies that are closer to the observed ones than does their mixture Poisson model.

6 Conclusions

With the release of `bimpoisson`, we add to the portfolio of models for correlated bivariate count data available to Stata users. Our implementation uses MSL estimators

Table 8: Postestimation predictions from the Munkin-Trivedi data (column %)

Counts	*emr*		*hosp*	
	empirical	fitted	empirical	fitted
0	81.8	82.4	80.4	80.9
1	13.4	12.3	13.6	13.0
2	3.1	3.0	4.0	3.3
3	1.2	1.1	1.1	1.2
4	0.3	0.5	0.5	0.6
5	0.2	0.2	0.1	0.3
other	0.0		0.3	
mean	2.26	0.29	0.30	0.33
(*emr*,*hosp*)				
(0,0)	72.4	72.1		
(1,0)	6.4	8.2		
(0,1)	7.3	9.2		
(1,1)	5.0	2.8		
other	8.9			

NOTES: Predictions from Model 6 in Table 7. Table entries show column percentages, except for ‘mean’ (count).

with or without a sampling function, as MT proposed, but we have added functionality by allowing Halton draws as an alternative to uniform draws, and the possibility of modifying the sampling function correlation. Although MT emphasised their sampling function estimator, we have shown that alternative (standard) MSL estimator works well as long as the mixture correlation is not large. Like much previous MSL-based research, our work has shown the benefits of using Halton draws – for a given S maintaining accuracy but with shorter run times – but we have also shown that the number of draws required to derive unbiased estimates may be very large. We attribute this to the complicated non-linear nature of the bivariate mixture Poisson regression model by comparison with most other models using MSL (cf. those cited in the Introduction).

Whether a researcher should model dependence using a copula or a mixture of random effects is not clearcut. The nice feature of MT's mixture model approach is that there is a straightforward interpretation in terms of unobserved heterogeneity (which may include measurent error), albeit at the cost of assuming a specific functional form for the heterogenity distribution. And, if one is prepared to assume multivariate normality, MSL enables fitting of mixture models with more than two dimensions. Writing programs to fit trivariate or higher-dimension versions of MT's model is a task for the future.

In contrast with MT's bivariate mixture model, the principal advantages of copula approaches are computational. Dependence is modelled separately from the two marginal count distributions, and it is easier to allow for covariate-dependent random effect standard deviations. (See, e.g., Marra and Radice (2025).) There are closed form likelihood expressions, and shorter run-times (as `bivcnto` illustrates).

Future research could fruitfully investigate semi-parametric versions of the bivariate mixture Poisson regression model in which the mixture distribution is summarized using a bivariate mass point distribution in the sense of Heckman and Singer (1984). MSL would not be required.

A more general issue for future research concerns the benefits and costs of using Poisson models with cross-equation dependence relative to using Poisson models assuming independence. We understand that the models identify the same regression parameters. Assuming the mixture model is true, there is an efficiency gain, but departures from the maintained assumption of bivariate normality may lead to lack of robustness. A possible route to examine this would be to adapt the methods used by Dhaene and Santos Silva (2012) in the univariate context to the bivariate one.

7 Acknowledgements

We thank Arne Risa Hole, Pravin Trivedi, and Maksat Jumamyradov for useful discussions, and especially João Santos Silva for helpful comments on a first draft.

8 References

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics. Methods and Applications.* 13th ed. Cambridge: Cambridge University Press.

———. 2013. *Regression Analysis of Count Data.* 2nd ed. Cambridge: Cambridge University Press.

Cappellari, L., and S. P. Jenkins. 2003. Multivariate probit regression using simulated maximum likelihood. *The Stata Journal* 3(3): 278–294.

———. 2006. Calculation of multivariate normal probabilities by simulation, with applications to maximum simulated likelihood estimation. *The Stata Journal* 6(2): 156–189.

Deb, P., and P. K. Trivedi. 1997. Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* 12(3): 313–336.

———. 2006. Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment. *The Stata Journal* 6(2): 246–255.

Dhaene, G., and J. M. C. Santos Silva. 2012. Specification and testing of models estimated by quadrature. *Journal of Applied Econometrics* 27(2): 322–332.

Gouriéroux, C., and A. Monfort. 1991. Simulation based inference in models with heterogeneity. *Annales d'Économie et de Statistique* (20–21): 69–107.

———. 1997. *Simulation-based Econometric Methods*. Oxford: Oxford University Press.

Greene, W. H. 2003. *Econometric Analysis*. 5th ed. Harlow: Longman.

Haan, P., and A. Uhlendorff. 2006. Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood. *The Stata Journal* 6(2): 229–245.

Heckman, J. J., and B. Singer. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52(2): 271–320.

Hole, A. R. 2007. Fitting mixed logit models by using maximum simulated likelihood. *The Stata Journal* 7(3): 388–401.

Jumamyradov, M., and M. K. Munkin. 2022. Biases in maximum mimulated likelihood estimation of bivariate models. *Journal of Econometric Methods* 11(1): 55–70.

Marra, G., and R. Radice. 2025. Bivariate copula-based regression for joint modeling of healthcare visits. *Health Economics* 35(2): 332–345.

Marshall, A. W., and I. Olkin. 1985. A family of bivariate distributions generated by the bivariate Bernoulli distribution. *Journal of the American Statistical Association* 80(390): 332–338.

Munkin, M. K., and P. K. Trivedi. 1999. Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application. *The Econometrics Journal* 2(1): 29–48.

Plum, A. 2016. bireprob: An estimator for bivariate random-effects probit models. *The Stata Journal* 16(1): 96–111.

Sándor, Z., and P. András. 2004. Alternative sampling methods for estimating multivariate normal probabilities. *Journal of Econometrics* 120(2): 207–234.

Stewart, M. B. 2006. Maximum simulated likelihood estimation of random-effects dynamic probit models with autocorrelated errors. *The Stata Journal* 6(2): 256–272.

Train, K. E. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

White, I. R. 2010. simsum: Analyses of simulation studies including Monte Carlo error. *The Stata Journal* 10(3): 369–385.

Xu, X., and J. W. Hardin. 2016. Regression models for bivariate count outcomes. *The Stata Journal* 16(2): 301–315.

About the authors

Stephen P. Jenkins is Professor of Economic and Social Policy at the LSE, and Managing Editor of *The Stata Journal*. He has contributed many programs to the SSC and written articles for the *Stata Journal*.

Fernando Rios-Avila is a data analyst at the New York Department of Health's Office of Aging and Long Term Care, and a Visiting Senior Fellow at the LSE. He has contributed many programs to the SSC and written articles for *The Stata Journal*.