The Stata Journal (2016)
16, Number 2, pp. 301–315

Regression models for bivariate count outcomes

Xinling Xu
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
xxu@email.sc.edu

James W. Hardin
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
jhardin@sc.edu

Abstract. We present a new command, `bivcnto`, for fitting regression models suitable for analyzing correlated count outcomes. `bivcnto` allows specification of two correlated count outcomes with either two outcome-specific covariate lists or one common covariate list and fits models using a copula function approach in the general case or using specific parameterizations by Marshall and Olkin (1985, *Journal of the American Statistical Association* 80: 332–338) or Famoye (2010a, *Journal of Applied Statistics* 37: 969–981; 2010b, *Statistica Neerlandica* 64: 112–124). `bivcnto` also calculates a likelihood-ratio test comparing the joint model with estimation of two independent outcome-specific models.

Keywords: st0433, bivcnto, copula function, correlated count data, Poisson, negative binomial, Famoye bivariate Poisson regression, Marshall–Olkin bivariate negative binomial regression, Famoye bivariate negative binomial regression, Famoye bivariate generalized Poisson regression, general bivariate count regression

1 Introduction

While there are several official commands for analyzing a single-count outcome (`nbreg`, `poisson`, etc.) as well as user-written additions (for example, `nbregf` and `nbregw` from Harris, Hilbe, and Hardin [2014] and `nbregp` from Hardin and Hilbe [2014]), regression modeling of correlated count outcomes is not currently supported in Stata.

We present a new estimation command, `bivcnto`, to evaluate the Famoye bivariate Poisson regression model, the Marshall–Olkin bivariate negative binomial regression model, the Famoye bivariate negative binomial regression model, the Famoye bivariate generalized Poisson regression model, and a general bivariate count regression model (computed from copula functions) that allows the user to specify the marginal distribution of each outcome.

This article is organized as follows. In section 2, we introduce the general concepts of the copula approach and the various supported regression models. In section 3, syntax is presented for `bivcnto`, followed by examples in section 4.

 st0433

2 Bivariate count-data models

2.1 Copula approach for bivariate data

While the distribution of univariate discrete data has been well developed, it is sometimes necessary to introduce a bivariate distribution for correlated discrete data. In many cases, when the marginal distributions are not independent, there is no explicit form for the bivariate distribution. Copula functions were originally introduced in Sklar (1959) and then revisited in Sklar (1973).

In general, if we define q random uniform variables, $U_1, \ldots, U_q$ on the $[0,1]$ interval, we may define a function

$$C(u_1, \ldots, u_q) = P(U_1 \le u_1, \ldots, U_q \le u_q)$$

where the function $C(\cdot, \cdots, \cdot)$ is a copula and u_j is a particular realization of U_j for $j = 1, 2, \ldots, q$ for $q \ge 2$. To be considered a copula, the function C must have a domain on the q-dimensional unit hypercube, must be grounded, and must be increasing over its domain. If so, then we may write

$$\begin{aligned} C\{F_1(x_1), \ldots, F_q(x_q)\} &= P\left\{F_1^{-1}(U_1) \le x_1, \ldots, F_q^{-1}(U_q) \le x_q\right\} \\ &= F(x_1, \ldots, x_q) \end{aligned}$$

That is, the copula as a function of the marginal distributions can be used to calculate the joint distribution; see Frees and Valdez (1998) for a complete review.

Using this approach, we can define

$$C(u, v; \theta) = F(y_1, y_2)$$

where $u = F_1(y_1)$, $v = F_2(y_2)$ are the cumulative distribution functions of the two marginal distributions, respectively, and the θ parameter measures the dependence between the two outcomes y_1 and y_2. In the related software, we have built-in support for calculations based on the following copula functions, where $\eta = 1 - \exp(-\theta)$. $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution, and $\Phi_2(\cdot, \cdot, \cdot)$ is the cumulative distribution function of the bivariate normal distribution:

$$\begin{aligned} C_1(u, v; \theta) &= -\frac{1}{\theta} \log\left[\left\{\eta - (1 - e^{-\theta u})(1 - e^{-\theta v})\right\}/\eta\right] \\ C_2(u, v; \theta) &= \Phi_2\left\{\Phi^{-1}(u), \Phi^{-1}(v), \theta\right\} \\ C_3(u, v; \theta) &= \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta} \end{aligned}$$

We note that in each copula function, the dependence parameter θ has a different range and so is parameterized in the software in a function-specific manner:

$$\begin{aligned} C_1(u, v, \theta) &: -\infty < \theta < \infty \\ C_2(u, v, \theta) &: -1 \le \theta \le 1 \\ C_3(u, v, \theta) &: -1 < \theta < \infty, \theta \ne 0 \end{aligned}$$

In each case, an unrestricted parameter θ_u is used in the software, where θ is then calculated. The copula function C_1, known as Frank's copula and specified in the software as `copula(frank)`, is parameterized as $\theta = \theta_u \in \Re$ (without restrictions). The copula function C_2, known as the normal copula and specified in the software as `copula(normal)`, is parameterized using the inverse hyperbolic tangent function $\theta = \{\exp(2\theta_u) - 1\}/\{\exp(2\theta_u) + 1\} \in (-1, 1)$. The copula function C_3, known as the Kimeldorf and Sampson (KS) copula and specified in the software as `copula(kimeldorf)`, is parameterized as $\theta = \exp(\theta_u) - 1 \geq -1$; the software ensures the value of θ_u is never equal to 0. See Joe (1997) for additional properties of these copula functions.

In general, the normal copula function is popular in financial modeling. However, because it imposes a linear correlation structure on the two variables, it may not be suitable for every situation. As such, two well-known copula functions are introduced as alternatives.

Both Frank's copula and the KS copula belong to the family of Archimedean copulas. Frank's copula is a symmetric copula function, while the KS copula is asymmetric. One advantage of Archimedean copulas is they have an explicit form and are easily generated. Depending on whether the correlation is believed to be more positive or negative, the user can select an asymmetric copula.

Using the copula approach to calculating the joint distribution of the count outcomes, we can specify each marginal distribution, and the two distributions are not required to be the same. The `bivcnto` command will allow specification of each marginal distribution as Poisson, negative binomial, or generalized Poisson distribution; see Harris, Yang, and Hardin (2012), and download associated software for individual regression models based on this last distribution.

The Fréchet–Hoeffding theorem states that the following bounds hold for any copula:

$$\max\{F_1(y_1) + F_2(y_2) - 1, 0\} \leq F(y_1, y_2) \leq \min\{F_1(y_1), F_2(y_2)\}$$

The software imposes these bounds on the marginal probabilities to ensure that bivariate probabilities stay within the range.

There are many articles about the applications of copula functions for bivariate count data. Lee (1999) studied the application of the Frank copula in the Australian Rugby League dataset. The two marginals followed the negative binomial distributions. Even though the term "copula" was not explicitly mentioned, van Ophem (1999) used the normal copula with specified Poisson marginal distributions. McHale and Scarf (2007) described how to apply Archimedian copulas for Poisson–Poisson and negative-binomial–negative-binomial scenarios and presented detailed examples using the Frank copula and KS copula. The command we develop and describe, `bivcnto`, can be used for carrying out all the analyses described in these articles. The main inspiration for the work described here is Cameron et al. (2004). However, they were actually more focused on the induced distribution of the difference of two correlated count distributions.

2.2 The Famoye bivariate Poisson regression model

The Famoye bivariate Poisson regression model allows negative, zero, or positive correlation.

Famoye (2010a) presents a bivariate Poisson distribution given by

$$P(y_1, y_2) = \mu_1^{y_1}\mu_2^{y_2}e^{-\mu_1-\mu_2}\left\{1+\lambda(e^{-y_1}-e^{-d\mu_1})(e^{-y_2}-e^{-d\mu_2})\right\}/y_1!y_2!$$

where μ_1 and μ_2 are the mean parameters for the two marginal Poisson distributions, respectively, and $d = 1 - \exp(-1)$. In this presentation, the two outcomes are independent when $\lambda = 0$. Also, note that the sign of λ indicates the sign of the correlation between the outcomes. The correlations can be calculated per observation as

$$\rho_i = \lambda d^2 \sqrt{\mu_{1i}\mu_{2i}}\exp\{-d(\mu_{1i}+\mu_{2i})\}$$

Subsequently, the average of the ρ_i can be used to summarize the correlation of the outcome variables.

2.3 Two parameterizations of the bivariate negative binomial regression model

Many approaches to bivariate count-data modeling have been suggested and researched. The new `bivcnto` command includes support for two parameterizations of bivariate negative binomial regression.

The Marshall–Olkin model

The Marshall–Olkin model is a "shared frailty" model for which the dispersion parameters of the marginal distributions are set to be equal, and there is no additional parameter directly measuring correlation between the two outcomes; see Marshall and Olkin (1985). The probability mass function if given by

$$f(y_1, y_2|\lambda_1, \lambda_2, \alpha) = \frac{\Gamma(y_1+y_2+\alpha)}{y_1!y_2!\Gamma(\alpha)}\left(\frac{\lambda_1}{\lambda_1+\lambda_2+1}\right)^{y_1}\left(\frac{\lambda_2}{\lambda_1+\lambda_2+1}\right)^{y_2}\left(\frac{1}{\lambda_1+\lambda_2+1}\right)^{\alpha}$$

where λ_1 and λ_2 are the two marginal means and α is the (common) overdispersion parameter.

While this approach defines a specific distribution for which estimates can easily be computed, the marginals are required to be negative binomial, the correlation between the outcomes is positive, and the heterogeneity in each marginal is assumed to be equal. The correlation between the outcomes for this model is given by

$$\mathrm{Corr}(y_1, y_2) = \frac{\lambda_1\lambda_2}{\sqrt{(\lambda_1^2+\alpha\lambda_1)(\lambda_2^2+\alpha\lambda_2)}}$$

The Famoye model

In comparison with the Marshall–Olkin bivariate negative binomial regression model, Famoye's presentation is more general. The Famoye model allows negative, zero, or positive correlation, and allows separate dispersion parameters for the marginal distributions.

Famoye (2010a) presents a bivariate negative binomial distribution given by

$$\begin{aligned} P(y_1, y_2) &= \left\{ \prod_{k=1}^{2} \binom{m_k^{-1} + y_k - 1}{y_k} \left(\frac{\mu_k}{m_k^{-1} + \mu_k} \right)^{y_k} \left(\frac{m_k^{-1}}{m_k^{-1} + \mu_k} \right)^{m_k^{-1}} \right\} \\ &\quad \times \left\{ 1 + \lambda \left(e^{-y_1} - c_1 \right) \left(e^{-y_2} - c_2 \right) \right\} \end{aligned}$$

where μ_k is the mean of the marginal negative binomial distribution with $k = 1, 2$, and $c_k = \{(1-\theta_k)/(1-\theta_k e^{-1})\}^{m_k^{-1}}$ with $\theta_k = \mu_k/(m_k^{-1} + \mu_k)$. In this presentation, the two outcomes are independent when $\lambda = 0$. Also, note that the sign of λ indicates the sign of the correlation between the outcomes. The parameter m_k is the dispersion parameter for the negative binomial distribution for $k = 1, 2$. As $m_k \to 0$, the negative binomial marginal model reduces to that of a Poisson marginal model. The correlations can be calculated per observation as

$$\rho_i = \lambda d^2 \sqrt{\mu_{1i}\mu_{2i}(1 + m_1\mu_{1i})(1 + m_2\mu_{2i})} \; (1 + dm_1\mu_{1i})^{-1-1/m_1} (1 + dm_2\mu_{2i})^{-1-1/m_2}$$

Subsequently, the average of the ρ_i can be used to summarize the correlation of the outcome variables.

2.4 The Famoye bivariate generalized Poisson regression model

Famoye (2010b) presents the following bivariate generalized Poisson distribution, which allows negative, zero, or positive correlation.

$$\begin{aligned} P(y_1, y_2) &= \left(\prod_{k=1}^{2} \left[\frac{\theta_k^{y_k} (1 + \alpha_k y_k)^{y_k - 1}}{y_k!} \exp\{-\theta_k (1 + \alpha_k y_k)\} \right] \right. \\ &\qquad \left. \left\{ 1 + \lambda (e^{-y_1} - c_1)(e^{-y_2} - c_2) \right\} \right) \\ c_k &= \exp\{\theta - t(s_k - 1)\} \\ 0 &= \ln(s_k) - \alpha_k \theta_k (s_k - 1) + 1 \end{aligned}$$

The mean parameters of the marginal generalized Poisson distributions are given by $\mu_k = \theta_k/(1 - \alpha_k\theta_k)$. In this presentation, the two outcomes are independent when $\lambda = 0$. Also, note that the sign of λ indicates the sign of the correlation between the outcomes.

3 Syntax

Software accompanying this article includes the command files as well as supporting files for prediction and help. In the following syntax diagrams, unspecified options include the usual collection of maximization and display options available to all estimation commands.

Equivalent in syntax to the `biprobit` command, the basic syntax for bivariate count regression models is presented as a bivariate syntax (common covariates for the two outcomes),

`bivcnto` *depvar*$_1$ *depvar*$_2$ [*indepvars*] [*if*] [*in*] [*weight*] [, `pfamoye` `gfamoye` `famoye` `molkin` `offset1(`*varname*`)` `offset2(`*varname*`)` `dist1(`*distribution*`)` `dist2(`*distribution*`)` `copula(`*function*`)`]

and as an outcome-specific covariate list syntax,

`bivcnto` *equation*$_1$ *equation*$_2$ [*if*] [*in*] [*weight*] [, `pfamoye` `gfamoye` `famoye` `molkin` `dist1(`*distribution*`)` `dist2(`*distribution*`)` `copula(`*function*`)`]

where *equation*$_1$ and *equation*$_2$ are specified as

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

In either case, the user specifies one of 1) `pfamoye` to fit Famoye's parameterization of a bivariate Poisson model, 2) `famoye` to fit Famoye's parameterization of a bivariate negative binomial model, 3) `molkin` to fit Marshall–Olkin's parameterization of a bivariate negative binomial model, 4) `gfamoye` to fit Famoye's parameterization of a bivariate generalized Poisson model, or 5) a combination of `dist1()`, `dist2()`, and `copula()` to fit a bivariate count model with given marginal distributions. *distribution* can be specified as `poisson`, `nbinomial`, or `gpoisson`; `copula()` can be specified as `frank`, `normal`, or `kimeldorf`.

Help files are included for the estimation and postestimation specifications of these models. The help files include example specifications of many possible models.

4 Example

In this example, we use data from the German health registry for the year 1984; see Hardin and Hilbe (2012). The data include responses by persons as to the number of doctor visits and number of hospital visits occurring in the previous year, as well as independent variables, including marital status, sex, whether the person has children, etc. Initially, we load the data and generate indicator variables of interest to assess the interaction of sex and marital status.

Because we believe that the number of doctor visits and hospital visits should be correlated for each person, we fit bivariate count-data models. Model 1 is a bivariate generalized Poisson model for which joint probabilities are estimated using the normal copula function. Relative to Poisson marginal distributions, this model will accommodate underdispersion or overdispersion.

```
. use rwm1984
(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

. generate byte postHS = edlevel1==0
. generate byte MarM = (married==1 & female==0) // married males
. generate byte MarF = (married==1 & female==1) // married females (reference
> group)
. generate byte SinM = (married==0 & female==0) // single males
. generate byte SinF = (married==0 & female==1) // single females
```

```
. bivcnto docvis hospvis MarM SinM SinF kids outwork postHS, copula(normal) irr
> dist1(gp) dist2(gp) nolog
Bivariate count regression                          Number of obs   =       3874
Distribution 1 : gpoisson                           LR chi2(12)     =     279.76
Distribution 2 : gpoisson                           Prob > chi2     =     0.0000
Copula function: Normal
Log likelihood = -9528.074                          Pseudo R2       =     0.0145

                    IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]

docvis
        MarM   .7794394    .040415    -4.81   0.000     .7041195    .8628163
        SinM   .5928822   .0482416    -6.42   0.000     .5054845    .6953909
        SinF   1.012853   .0635227     0.20   0.839     .8956983     1.14533
        kids   .7019465   .0292074    -8.51   0.000      .646973     .761591
     outwork   1.364784    .063108     6.73   0.000     1.246534    1.494251
      postHS   .9114119   .0486906    -1.74   0.083     .8208064    1.012019
       _cons   3.686083   .1952554    24.63   0.000     3.322585    4.089348

hospvis
        MarM   1.332045   .1976339     1.93   0.053     .9959257    1.781601
        SinM   .8377243   .2073132    -0.72   0.474      .515766     1.36066
        SinF   1.283144   .2433393     1.31   0.189     .8848118    1.860801
        kids    .820675   .0997965    -1.63   0.104       .64664    1.041549
     outwork    1.32767   .1787069     2.11   0.035     1.019805    1.728477
      postHS   .7786416   .1272684    -1.53   0.126     .5652093     1.07267
       _cons   .0929525   .0143513   -15.39   0.000     .0686816    .1258003

/atanhdelta1   .8420648   .0159916    52.66   0.000     .8107218    .8734077
/atanhdelta2    .262389     .02783     9.43   0.000     .2078433    .3169348
 /atanhtheta   .3643823   .0280477    12.99   0.000     .3094099    .4193548

      delta1   .6869011   .0084462                      .6699882    .7031014
      delta2   .2565287   .0259986                      .2049013    .3067327
       theta   .3490683   .0246301                      .2999001    .3963867

LR test of independence: chi2(1) = 166.729               Prob > chi2 = 0.0000
```

Model 2 is a bivariate negative binomial model for which joint probabilities are estimated using the normal copula function. Because the negative binomial marginal distributions allow for overdispersion relative to the Poisson model, the results should be comparable to model 1.

```
. bivcnto docvis hospvis MarM SinM SinF kids outwork postHS, copula(normal) irr
> dist1(nbinomial) dist2(nbinomial) nolog
Bivariate count regression                      Number of obs   =       3874
Distribution 1 : nbinomial                      LR chi2(12)     =     204.00
Distribution 2 : nbinomial                      Prob > chi2     =     0.0000
Copula function: Normal
Log likelihood = -9563.0986                     Pseudo R2       =     0.0106

------------------------------------------------------------------------------
             |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
        MarM |   .7553233   .0482379    -4.39   0.000     .6664565    .8560398
        SinM |   .5428058   .0539187    -6.15   0.000      .446778    .6594733
        SinF |   1.094595   .0964304     1.03   0.305     .9210115    1.300893
        kids |   .6699012   .0361331    -7.43   0.000     .6026964    .7445997
     outwork |   1.367538    .080724     5.30   0.000     1.218131     1.53527
      postHS |   .7767982   .0525775    -3.73   0.000      .680291    .8869961
       _cons |   3.815242   .2435281    20.98   0.000     3.366585     4.32369
-------------+----------------------------------------------------------------
hospvis      |
        MarM |   1.339777   .2106125     1.86   0.063     .9845185    1.823229
        SinM |   .8486239   .2191493    -0.64   0.525     .5115643    1.407765
        SinF |   1.267091    .265061     1.13   0.258     .8409039     1.90928
        kids |   .8669514   .1132898    -1.09   0.275     .6710628    1.120022
     outwork |   1.318504   .1887649     1.93   0.053     .9959056    1.745599
      postHS |   .7299941   .1267115    -1.81   0.070     .5194799    1.025817
       _cons |   .0821457   .0128806   -15.94   0.000     .0604109    .1117003
-------------+----------------------------------------------------------------
    /lnalpha1 |    .801881   .0312111    25.69   0.000     .7407083    .8630536
    /lnalpha2 |   1.517031   .1581257     9.59   0.000      1.20711    1.826951
 /atanhtheta |   .3524641    .027713    12.72   0.000     .2981476    .4067807
-------------+----------------------------------------------------------------
      alpha1 |   2.229731   .0695924                      2.097421    2.370388
      alpha2 |   4.558669   .7208429                      3.343807    6.214911
       theta |   .3385591   .0245365                      .2896165    .3857358
------------------------------------------------------------------------------
LR test of independence: chi2(1) = 194.597             Prob > chi2 = 0.0000
```

Model 3 is a simplification wherein we illustrate whether the marginal distributions could be modeled as Poisson. This is for illustration more than anything because the first two models already established significant overdispersion. This model is a specific parameterization of the joint distribution that is not estimated via copula functions.

```
. bivcnto docvis hospvis MarM SinM SinF kids outwork postHS, irr pfamoye nolog

Famoye bivariate Poisson regression             Number of obs   =       3874
Distribution 1 : poisson                        LR chi2(12)     =    1525.70
Distribution 2 : poisson                        Prob > chi2     =     0.0000
Parameteriz.   : Famoye
Log likelihood = -17292.442                     Pseudo R2       =     0.0423

------------------------------------------------------------------------------
             |       IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
        MarM |   .8256116   .0194582    -8.13   0.000     .7883417    .8646434
        SinM |   .7184228   .0265246    -8.96   0.000     .6682719    .7723373
        SinF |    1.09806   .0306338     3.35   0.001     1.039631    1.159773
        kids |   .7111962   .0137636   -17.61   0.000     .6847252    .7386906
     outwork |   1.496067   .0309472    19.47   0.000     1.436625    1.557969
      postHS |   .7867019   .0205356    -9.19   0.000     .7474652    .8279982
       _cons |   3.475433   .0773372    55.98   0.000     3.327113    3.630365
-------------+----------------------------------------------------------------
hospvis      |
        MarM |   1.506575   .1606378     3.84   0.000     1.222452    1.856734
        SinM |   .8145506   .1690816    -0.99   0.323     .5422863     1.22351
        SinF |   .9681015   .1582902    -0.20   0.843     .7026589     1.33382
        kids |   .9758035   .0906394    -0.26   0.792     .8133863    1.170652
     outwork |   1.685831   .1670805     5.27   0.000     1.388201    2.047271
      postHS |   .7455801   .0964673    -2.27   0.023     .5785769    .9607879
       _cons |   .1135148    .013544   -18.24   0.000     .0898446    .1434211
-------------+----------------------------------------------------------------
     /lambda |   1.239064   .0311917    39.72   0.000      1.17793    1.300199
------------------------------------------------------------------------------
LR test of independence: chi2(1) = 150.989              Prob > chi2 = 0.0000
```

Model 4 is a specific parameterization of a bivariate generalized Poisson regression model. It is similar to model 1, except that joint probabilities are not estimated using a copula function approach.

```
. bivcnto docvis hospvis MarM SinM SinF kids outwork postHS, irr gfamoye nolog
Famoye bivariate gen. Poisson regression         Number of obs   =       3874
Distribution 1 : gpoisson                        LR chi2(12)     =     274.35
Distribution 2 : gpoisson                        Prob > chi2     =     0.0000
Parameteriz.   : Famoye
Log likelihood = -9565.9571                      Pseudo R2       =     0.0141

------------------------------------------------------------------------------
             |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
        MarM |   .7763308    .039299    -5.00   0.000      .703004    .857306
        SinM |   .6034113   .0488873    -6.24   0.000     .5148142   .7072555
        SinF |   1.013709   .0634444     0.22   0.828     .8966841   1.146006
        kids |    .711594   .0296011    -8.18   0.000     .6558789   .7720418
     outwork |   1.363803   .0616085     6.87   0.000     1.248244   1.490061
      postHS |   .9097265   .0485966    -1.77   0.077     .8192956   1.010139
       _cons |   3.716485   .1944046    25.10   0.000      3.35434   4.117728
-------------+----------------------------------------------------------------
hospvis      |
        MarM |   1.329661   .1971473     1.92   0.055     .9943383   1.778065
        SinM |   .8904528   .2160404    -0.48   0.632     .5534694   1.432611
        SinF |   1.251321   .2385659     1.18   0.240     .8611644   1.818241
        kids |   .8117869   .0988215    -1.71   0.087     .6394734   1.030532
     outwork |   1.430081   .1917152     2.67   0.008     1.099637   1.859824
      postHS |   .7621674   .1252311    -1.65   0.098     .5523192   1.051745
       _cons |   .1052071    .016458   -14.39   0.000     .0774263   .1429559
-------------+----------------------------------------------------------------
/atanhdelta1 |    .850666   .0161206    52.77   0.000     .8190702   .8822619
/atanhdelta2 |   .3611645   .0316393    11.42   0.000     .2991526   .4231764
      /theta |    1.61282   .1117518    14.43   0.000     1.393791    1.83185
-------------+----------------------------------------------------------------
      delta1 |   .6914173    .008414                      .6745635   .7075506
      delta2 |   .3462394   .0278463                      .2905369    .399603
------------------------------------------------------------------------------
LR test of independence: chi2(1) = 90.9624               Prob > chi2 = 0.0000
```

Model 5 is a specific parameterization of a bivariate negative binomial regression model. It is similar to model 2 (and model 6), except that joint probabilities are not estimated using a copula function approach.

```
. bivcnto docvis hospvis MarM SinM SinF kids outwork postHS, irr famoye nolog
Famoye bivariate neg bin regression              Number of obs   =       3874
Distribution 1 : nbinomial                       LR chi2(12)     =     196.14
Distribution 2 : nbinomial                       Prob > chi2     =     0.0000
Parameteriz.   : Famoye
Log likelihood = -9613.7926                      Pseudo R2       =     0.0101

------------------------------------------------------------------------------
             |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
        MarM |   .7850615   .0504826    -3.76   0.000     .6920987     .890511
        SinM |   .7066418   .0680447    -3.61   0.000      .585106    .8534226
        SinF |   1.114674   .0996621     1.21   0.225     .9354971    1.328168
        kids |   .7344114   .0395042    -5.74   0.000     .6609262     .816067
     outwork |   1.499644   .0882375     6.89   0.000     1.336301    1.682953
      postHS |   .7512271   .0515644    -4.17   0.000     .6566661     .859405
       _cons |   3.513729   .2202173    20.05   0.000     3.107567    3.972976
-------------+----------------------------------------------------------------
hospvis      |
        MarM |    1.50087   .2589676     2.35   0.019     1.070218    2.104813
        SinM |   .8452774   .2431846    -0.58   0.559     .4809615    1.485553
        SinF |   .9809291   .2345484    -0.08   0.936     .6139149    1.567354
        kids |   1.042998   .1494766     0.29   0.769     .7875786    1.381252
     outwork |   1.630445   .2583429     3.09   0.002     1.195181    2.224223
      postHS |   .7499424    .138282    -1.56   0.119     .5224861    1.076418
       _cons |   .0858934   .0148254   -14.22   0.000     .0612407      .12047
-------------+----------------------------------------------------------------
   /lnalpha1 |   .8320476   .0308378    26.98   0.000     .7716065    .8924886
   /lnalpha2 |   2.259601   .1077669    20.97   0.000     2.048381     2.47082
     /lambda |   1.690835   .1250408    13.52   0.000      1.44576    1.935911
-------------+----------------------------------------------------------------
      alpha1 |   2.298019   .0708659                      2.163239    2.441197
      alpha2 |   9.579261   1.032327                      7.755337    11.83214
------------------------------------------------------------------------------
LR test of independence: chi2(1) = 93.209              Prob > chi2 = 0.0000
```

Model 6 is a specific parameterization of a bivariate negative binomial regression model. It is similar to model 2 (and model 5), except that joint probabilities are not estimated using a copula function approach. Also, the Marshall–Olkin model (model 6) assumes that dependency of the two marginal distributions is captured through the constraint that the two marginal distributions have equal dispersion parameters; this is the reason there is only one dispersion parameter in the model output.

```
. bivcnto docvis hospvis MarM SinM SinF kids outwork postHS, irr molkin nolog

Marshall-Olkin bivariate neg bin regression      Number of obs   =       3874
Distribution 1 : nbinomial                       LR chi2(12)     =     231.33
Distribution 2 : nbinomial                       Prob > chi2     =     0.0000
Parameteriz.   : Marshall-Olkin
Log likelihood = -9814.5153                      Pseudo R2       =     0.0116

------------------------------------------------------------------------------
             |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
        MarM |    .790513   .0505189    -3.68   0.000     .6974478    .8959964
        SinM |   .7052426   .0671038    -3.67   0.000     .5852572    .8498267
        SinF |   1.127196   .0998536     1.35   0.177     .9475349    1.340923
        kids |   .7256723   .0385847    -6.03   0.000     .6538549    .8053781
     outwork |    1.49835   .0876521     6.91   0.000     1.336038    1.680381
      postHS |   .7449296   .0504891    -4.34   0.000     .6522639    .8507601
       _cons |    3.51977   .2189786    20.23   0.000     3.115715    3.976224
-------------+----------------------------------------------------------------
hospvis      |
        MarM |   1.369553   .1743942     2.47   0.014     1.067062    1.757794
        SinM |   .6829585   .1530451    -1.70   0.089     .4401974    1.059598
        SinF |   .9624444   .1805015    -0.20   0.838     .6664035    1.389998
        kids |   1.019477   .1084942     0.18   0.856     .8275446    1.255925
     outwork |   1.662235   .1954399     4.32   0.000     1.320113    2.093022
      postHS |   .7681872   .1119307    -1.81   0.070     .5773519      1.0221
       _cons |   .0915673   .0120894   -18.11   0.000     .0706901    .1186102
-------------+----------------------------------------------------------------
    /lnalpha |   .8042573   .0304566    26.41   0.000     .7445636    .8639511
-------------+----------------------------------------------------------------
       alpha |   2.235036   .0680715                      2.105522    2.372516
------------------------------------------------------------------------------
LR test of independence: chi2(1) = 308.236               Prob > chi2 = 0.0000
```

In a bivariate model of doctor visits and hospital visits (across all the highlighted examples), it can be seen that married men have a lower rate of doctor visits than married women. Single men have an even lower rate, but single women do not differ from married women in terms of their rates of doctor visits.

The correlation from the bivariate generalized Poisson model indicates that the outcomes are positively correlated at about 35%. This model has a better fit than two independent generalized Poisson regression models, with a χ^2 value of 166.7 and associated p-value less than 0.0001, thus rejecting the assumption of independence. The `delta1` and `delta2` from the output are the dispersion parameters for the marginal Poisson distribution. Relative to Poisson distribution, a positive value of this dispersion parameter indicates overdispersion and a negative value indicates underdispersion. From the output of the first model, we conclude that (relative to marginal Poisson distributions) the overdispersion of the number of doctor visits is about 2.5 times the overdispersion of the number of hospital visits (`delta1`/`delta2` = 2.7). The Famoye bivariate generalized poisson model result is similar, as shown in the fourth model output.

For the second model output using negative binomial distributions with the normal copula, the `alpha1` and `alpha2` output lines are for the dispersion parameters of the marginal negative binomial distributions, with larger values indicative of greater dispersion (relative to a marginal Poisson distribution). When $\alpha = 0$, the dispersion of the distribution is 0, and it reduces to a Poisson distribution. In both outputs using copula functions, the two estimated θ correlation parameters are close. The likelihood-ratio test of independence again rejects the independence assumption.

When we compare the last two models with the copula method, assuming the marginal distributions are negative binomial distributions, the significance of predictors for doctor visits do not change. The significant dependence parameters indicate a positive correlation between doctor visits and hospital visits.

5 References

Cameron, A. C., T. Li, P. K. Trivedi, and D. M. Zimmer. 2004. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *Econometrics Journal* 7: 566–584.

Famoye, F. 2010a. On the bivariate negative binomial regression model. *Journal of Applied Statistics* 37: 969–981.

———. 2010b. A new bivariate generalized Poisson distribution. *Statistica Neerlandica* 64: 112–124.

Frees, E. W., and E. A. Valdez. 1998. Understanding relationships using copulas. *North American Actuarial Journal* 2: 1–25.

Hardin, J. W., and J. M. Hilbe. 2012. *Generalized Linear Models and Extensions*. 3rd ed. College Station, TX: Stata Press.

———. 2014. Regression models for count data based on the negative binomial(p) distribution. *Stata Journal* 14: 280–291.

Harris, T., J. M. Hilbe, and J. W. Hardin. 2014. Modeling count data with generalized distributions. *Stata Journal* 14: 562–579.

Harris, T., Z. Yang, and J. W. Hardin. 2012. Modeling underdispersed count data with generalized Poisson regression. *Stata Journal* 12: 736–747.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Lee, A. 1999. Applications: Modelling rugby league data via bivariate negative binomial regression. *Australian & New Zealand Journal of Statistics* 41: 141–152.

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: 332–338.

McHale, I., and P. Scarf. 2007. Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica* 61: 432–445.

Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 8: 229–231.

———. 1973. Random variables, distribution functions, and copulas. *Kybernetica* 9: 449–460.

van Ophem, H. 1999. A general method to estimate correlated discrete random variables. *Econometric Theory* 15: 228–237.

About the authors

Xinling Xu is a PhD candidate in the Department of Epidemiology and Biostatistics at the University of South Carolina in Columbia, SC.

James W. Hardin is an associate professor in the Department of Epidemiology and Biostatistics and an affiliated faculty in the South Carolina Rural Health Research Center at the University of South Carolina in Columbia, SC.