bivpoisson: A STATA COMMAND ESTIMATING SEEMINGLY UNRELATED COUNT DATA

James Fisher, Joseph V. Terza, and Abbie Zhang

2022 Stata Conference
August 4-5, 2022
Washington DC

ssc install bivpoisson
https://github.com/zhangyl334/bivpoisson

Why this command?

Suppose you:

-- have a dataset that involves more than one count-valued outcome variables, and they are potentially correlated.

-- assume a fully parametrically specification [e.g. the joint probability mass function of the outcome variables] conditional on regressors.

-- want to make causal inference in terms of the average treatment effects (ATEs).

We offer a Stata package command on estimating the deep parameters under the context of bivariate count data.

(a post-estimation command on average treatment effects is forthcoming)

Why this command? (Cont’d)

Widely application in empirical research:

Example 1:

Prediction of traffic crash counts of: (1) fatal crashes, (2) property damage-only crashes using multiple inter-dependent sources of risk.

Example 2:

Investigating the association of Medicaid expansion with: (1) number of Ambulatory Care Sensitive Condition ED admissions, (2) number of Non-ED Outpatient Visits.

Example 3:

Estimating the causal effects of private insurance status on: (1) Physician office visits, (2) Non-physician health professional office visits.

Our Contribution

--`bivpoisson` estimates the deep parameters for 2-dimensional correlated count data

--`bivpoisson` achieves higher precision in terms of deep parameter estimates (compared to fitting a count valued system-of-equations using linear seemingly unrelated regression model via command “sureg”).

Our Contribution (Cont'd)

-- `bivpoisson` adds additional new functionality to Stata's "`gsem`" class command:

-- "`gsem`" (Stata's Structural Equation Modeling command class) offers many options in family and link functions[1] for system-of-equation estimation.

-- however, "`gsem`" does not allow Gaussian + Poisson combination:

If you specify both **family()** and **link()**, not all combinations make sense. You may choose from the following combinations:

	identity	log	logit	probit	cloglog
Gaussian	D	x			
Bernoulli			D	x	x
beta			D	x	x
binomial			D	x	x
ordinal			D	x	x
multinomial			D		
Poisson		D			
negative binomial		D			
exponential		D			
Weibull		D			
gamma		D			
loglogistic		D			
lognormal		D			
pointmass	D				

D denotes the default.

[1] See gsem family-and-link options (in Stata, type: help gsem family and link options)

Outline

In the rest of this presentation, we will:

-- Detail the fully parametric specification

-- Describe bivpoisson command

-- Provide a real-world application

-- Discussion of future works

Specification

--A structural model on correlated-count outcomes:

$$\text{pmf}(Y_{1X^*}, Y_{2X^*} \mid X_o) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{poi}_1(Y_{1X^*}; \lambda_1^*) \times \text{poi}_2(Y_{2X^*}; \lambda_2^*)$$

$$\times\, g(\eta_1, \eta_2; \rho_{12})]\, d\eta_1\, d\eta_2 \qquad (1)$$

$[Y_{1X^*} \quad Y_{2X^*}] \equiv$ **bivariate vector of count-valued potential outcomes.**

$\text{poi}_r(Y_{rX^*}; \lambda_r^*) \equiv$ **the pmf of the Poisson distributed r.v.** Y_{rX^*} **with parameters** λ_r^*,

with $\lambda_r^* \equiv \exp(X_o\beta_{ro} + X^*\beta_{rX} + \eta_r)$, $r = 1, 2$.

(η_1, η_2) **are the "structural cross-equation heterogeneity terms"**

$$g(\eta_1, \eta_2) \sim N(0, \Sigma) \qquad (2)$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12} \\ \rho_{12} & \sigma_2^2 \end{bmatrix} \qquad (3)$$

Specification (cont'd)

and

X_o ≡ the vector of observable control variables,

X^* ≡ counterfactually mandated version of the causal variable (any type)

and the parameters to be estimated are β_{ro}, β_{rX}, and ρ_{12}.

This model is designed to exploit possible statistical efficiency in estimation by taking explicit (parametric) account of cross-equation correlation through the bivariate normal mixture component (essentially ρ_{12}).

Specification (cont'd)

Suppose that the requisite conditions establishing the legitimacy of following aspect of the data generating process specification are satisfied:

$$\text{pmf}(Y_1, Y_2 \mid X_o, X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{pois}_1[(Y_1; \lambda_1) \times \text{pois}_2(Y_2; \lambda_2) \times \varphi_2(\eta_1, \eta_2; \rho_{12})]\, d\eta_1\, d\eta_2 \qquad (4)$$

$[Y_1 \quad Y_2] \equiv$ **the observable version of the outcome vector**

$X \equiv$ **the observable version of the causal variable**

$\lambda_r \equiv \exp(X_o\beta_{ro} + X\beta_{rX} + \eta_r)$ **for** $r = 1, 2$

The New Command bivpoisson

--Syntax

```
bivpoisson (depvar1 = indepvar1) (depvar2 = indepvar2) [if]
```

where:

```
depvar1 = count-valued dependent variable for equation 1

depvar2 = count-valued dependent variable for equation 2

indepvar1 = vector of independent variables for equation 1

indepvar2 = vector of independent variables for equation 2
```

(indepvar1 and indepvar2 can be different or the same)

--Options

```
[if] allows computing results by subpopulations
```

The New Command bivpoisson

--Warning Message:

```
depvar1 is zero-inflated

or

depvar2 is zero-inflated

r(2000);
```

will show up when a dependent variable has mean less than 1 (indicating there are too many zero values in the dependent variable).

--In current version, optimization is unlikely to converge when data is zero-inflated.

--A two-part model is needed for each zero-inflated equation (future work).

Example Output

```
. use "https://github.com/zhangyl334/bivpoisson/raw/main/Health_Data.dta"

.
. bivpoisson (ofp = privins black numchron) (ofnp = privins black numchron age)
initial:       f(p) = -898.14156
rescale:       f(p) = -898.14156
rescale eq:    f(p) = -889.97635
Iteration 0:   f(p) = -889.97635  (not concave)
Iteration 1:   f(p) = -878.49262  (not concave)
Iteration 2:   f(p) = -845.96974  (not concave)
Iteration 3:   f(p) = -840.21573
Iteration 4:   f(p) = -832.94616
Iteration 5:   f(p) = -832.69668
Iteration 6:   f(p) = -832.69538
Iteration 7:   f(p) = -832.69538

                                                   Number of obs = 207
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ofp						
privins	.3997619	.1830324	2.18	0.029	.0410251	.7584988
black	-.1335776	.1905022	-0.70	0.483	-.506955	.2397997
numchron	.2380122	.053071	4.48	0.000	.133995	.3420294
_cons	.6682984	.1939622	3.45	0.001	.2881394	1.048457
ofnp						
privins	1.305625	.4458126	2.93	0.003	.4318483	2.179402
black	-2.151162	.9190452	-2.34	0.019	-3.952457	-.3498661
numchron	.2358258	.1392374	1.69	0.090	-.0370744	.508726
age	-.0809187	.3125795	-0.26	0.796	-.6935632	.5317257
_cons	-2.271814	2.292566	-0.99	0.322	-6.76516	2.221532
sigmasq1						
_cons	.8514478	.130599	6.52	0.000	.5954785	1.107417
sigmasq2						
_cons	3.478548	.6043013	5.76	0.000	2.294139	4.662956
sigma12						
_cons	.4178385	.2111368	1.98	0.048	.004018	.831659

Example Return List

```
. ereturn list

scalars:
                  e(rank) =  12
                     e(N) =  207
                    e(ic) =  7
                     e(k) =  12
                  e(k_eq) =  5
                  e(k_dv) =  2
             e(converged) =  1
                    e(rc) =  0

macros:
                e(indep2) : "privins black numchron age"
               e(depvar2) : "ofnp"
                e(indep1) : "privins black numchron"
               e(depvar1) : "ofp"
                 e(title) : "Bivariate Count Seemingly Unrelated Regression Estimation"
                   e(cmd) : "bivpoisson"
                   e(opt) : "moptimize"
               e(predict) : "ml_p"
                  e(user) : "BivPoissNormLF()"
             e(ml_method) : "lf0"
             e(technique) : "nr"
                 e(which) : "max"
                e(depvar) : "Y1 Y2"
            e(properties) : "b V"

matrices:
                     e(b) :  1 x 12
                     e(V) :  12 x 12
                  e(ilog) :  1 x 20
              e(gradient) :  1 x 12
```

Application

-- Use a health survey dataset "the 1987 National Medical Expenditure Survey Data"

--This data is used by many previous works such as Deb and Trivedi (1997), Chib and Winkelmann (2001), and Famoye (2015).

--Policy Relevancy: Causal effects of insurance coverage on use of health services.

Application (cont'd)

Variables we use:

--depvar1 = the number of physician office visits, denoted *ofp*

--depvar2 = the number of non-physician office visits, denoted *ofnp*

--indepvar1 = [Private Insurance Status, Black, Number of Chronic Conditions, Constant Term], denoted by: *[privins, black, numchron,1]*

--indepvar2 = [Private Insurance Status, Black, Number of Chronic Conditions, Age, Constant Term], denoted by *[privins, black, numchron, age,1]*

Example (cont'd)

Access the dataset:

--In Stata, type:

```
use https://github.com/zhangyl334/bivpoisson/raw/main/Health Data.dta
```

--Then type:

```
bivpoisson (ofp = privins black numchron) (ofnp = privins black numchron age)
```

Example (cont'd)

```
. use "https://github.com/zhangyl334/bivpoisson/raw/main/Health_Data.dta"

.
. bivpoisson (ofp = privins black numchron) (ofnp = privins black numchron age)
initial:       f(p) = -898.14156
rescale:       f(p) = -898.14156
rescale eq:    f(p) = -889.97635
Iteration 0:   f(p) = -889.97635  (not concave)
Iteration 1:   f(p) = -878.49262  (not concave)
Iteration 2:   f(p) = -845.96974  (not concave)
Iteration 3:   f(p) = -840.21573
Iteration 4:   f(p) = -832.94616
Iteration 5:   f(p) = -832.69668
Iteration 6:   f(p) = -832.69538
Iteration 7:   f(p) = -832.69538

                                                      Number of obs = 207
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ofp						
privins	.3997619	.1830324	2.18	0.029	.0410251	.7584988
black	-.1335776	.1905022	-0.70	0.483	-.506955	.2397997
numchron	.2380122	.053071	4.48	0.000	.133995	.3420294
_cons	.6682984	.1939622	3.45	0.001	.2881394	1.048457
ofnp						
privins	1.305625	.4458126	2.93	0.003	.4318483	2.179402
black	-2.151162	.9190452	-2.34	0.019	-3.952457	-.3498661
numchron	.2358258	.1392374	1.69	0.090	-.0370744	.508726
age	-.0809187	.3125795	-0.26	0.796	-.6935632	.5317257
_cons	-2.271814	2.292566	-0.99	0.322	-6.76516	2.221532
sigmasq1						
_cons	.8514478	.130599	6.52	0.000	.5954785	1.107417
sigmasq2						
_cons	3.478548	.6043013	5.76	0.000	2.294139	4.662956
sigma12						
_cons	.4178385	.2111368	1.98	0.048	.004018	.831659

Equation1's coefficient estimates

Equation2's coefficient estimates

Ancillary parameter estimates

Exploring Estimator's Statistical Property

Simulation study shows seemingly unrelated count regression (`bivpoisson`) achieves better precision than linear seemingly unrelated regression (`sureg`) in ATE.

			Poisson SUR		Linear SUR	
Design	ρ_{12}	**True ATE**	**Average ATE**	**AAPB**	**Average ATE**	**AAPB**
1 (Over-Dispersed Correlated Counts) Omega = -0.1	0.75	4.765	4.035	34.19%	2.185	53.06%
	0.5	4.767	4.265	36.93%	2.191	52.89%
	0.25	4.767	4.147	35.53%	2.213	52.57%
	0	4.767	4.224	36.27%	2.209	52.64%

100 replications with 10,000 observations for each replication

-- AAPB's formula:

$$\text{AAPB } \widehat{\text{ATE}(\Delta)} = \frac{1}{R} \times \sum_{r=1}^{R} \left| \frac{\widehat{\text{ATE}(\Delta)}_r - \text{ATE}(\Delta)}{\text{ATE}(\Delta)} \right| \quad (5)$$

Exploring Estimator's Statistical Property (Cont'd)

Estimating the effects of private insurance status on 2 correlated health utilization counts (sureg versus bivpoisson)

Average Treatment Effects:
Private Insurance Status on Two Correlated Health Care Utilization Counts

	Linear Seemingly Unrelated Regression (SUR) Model				**Count-Outcome SUR Model (Poisson case)**			
	ATE	**S.E.**	**T-Stat**	**P-Value**	**ATE**	**S.E.**	**T-Stat**	**P-Value**
Count of Physician Office Visits (Y_1)	1.6302	0.2784	5.8536	0.0000	1.8830	0.5036	3.7400	0.0002
Count of Non-Physician Office Visits (Y_2)	0.5958	0.2288	2.6034	0.0092	4.0088	0.4783	3.7470	0.0002

Future works

--post estimation command:

bivpoisson_predi

bivpoisson_ate

--Include a plug-and-play feature for more choices of marginal distributions: Conway-Maxwell-Poisson (CMP), Negative Binomial (NB), Zero-inflated negative binomial (ZINB) regression.

--Increase the dimensionality of the correlated outcome to 3+.

Discussion and Conclusion

--Introduced new community contributed package "bivpoisson" to estimate 2-dimensional correlated count-valued data.

--Applied to a health care survey dataset.

--Compared to Linear SUR (by Stata command: sureg) and show precision gain in policy effect estimation.

Thank you!

Contact: Abbie Zhang

Email: zhangyl334@gmail.com

GitHub Repository: https://github.com/zhangyl334/bivpoisson

Twitter: @abbiezhang_econ

Website: yileizhang.com

References

Aitchison, J., & Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, *76*(4), 643–653. https://doi.org/10.1093/biomet/76.4.643

Baum, C. F. (2015). Ado-file programming. NCER, Queensland University of Technology.

Chib, S., & Winkelmann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business & Economic Statistics*, *19*(4), 428–435. https://doi.org/10.1198/07350010152596673

Mander, A. (2018). INTEGRATE_AQ: Stata module to do adaptive quadrature for integrals. Statistical Software Components from Boston College Department of Economics. Retrieved from https://econpapers.repec.org/software/bocbocode/s458502.htm

Zhang, Y. (2021). Exploring the Importance of Accounting for Nonlinearity in Correlated Count Regression Systems from the Perspective of Causal Estimation and Inference. https://doi.org/10.7912/C2/2873

Terza, J.V. (2020): "Regression-Based Causal Analysis from the Potential Outcomes Perspective," Journal of Econometric Methods, DOI: https://doi.org/10.1515/jem-2018-0030

Terza, J. V., & Zhang, A. (2020). Two-Dimensional Gauss-Legendre Quadrature: Seemingly Unrelated Dispersion-Flexible Count Regressions. 2020 Stata Conference. Retrieved from https://www.stata.com/meeting/us20/slides/us20_Terza.pdf

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, *57*(298), 348–368. https://doi.org/10.1080/01621459.1962.10480664

Appendix

Simulation Study Design

For each of these designs, 100 sample of size 10,000 were generated. In each replication, we

-- calculate true ATE

-- estimate deep parameters using both Zellner's Linear SUR model and Count-Outcome CMP SUR model with our simulated bivariate CMP data

-- calculate the averaged estimated AIE and the averaged absolute percent bias (AAPB) of ATE using both models.

-- compare estimated ATEs and AAPBs of both models, and to true ATE

Appendix

More Simulation Results

(Conway Maxwell Poisson SUR versus Linear SUR)

			CMP SUR		Linear SUR	
Design	ρ_{12}	**True ATE**	**Average ATE()**	**AAPB ATE()**	**Average ATE()**	**AAPB ATE()**
1 (Over-Dispersed Correlated Counts) Omega = -0.1	0.75	4.765	4.502	10.48%	2.185	53.06%
	0.5	4.767	4.865	25.48%	2.191	52.89%
	0.25	4.767	4.502	19.00%	2.213	52.57%
	0	4.767	4.7079	26.87%	2.209	52.64%

100 replications with 10,000 observations for each replication

Takeaways:

--The accuracy and efficiency gains with CMP SUR estimator persist across all correlation structures

--The efficiency gains tend to increase with the correlations among the count outcomes

Appendix

ATE Estimation under the General Potential Outcome Framework

Observation Data

Patient ID	Xo (Policy Variable) Private Insurance Status	Y1 Number of Physician Office Visits in the past 2	Y2 Number of Non-Physician Health Professional Office Visits in the past 2 weeks
1	0	2	2
2	0	1	2
3	1	3	4
4	1	1	0
5	1	3	5

$\widehat{\beta_1}, \widehat{\sigma_{12}}, \widehat{\beta_2},$

obtained via MLE using observational data

Counterfactural Prediction of Conditional Means (Counterfactural Scenario 1)

Patient ID	Xo (Policy Variable) Private Insurance Status	Counterfactural Mandated Private Insurance Status	Y1 Number of Physician Office Visits in the past 2 weeks	Y2 Number of Non-Physician Health Professional Office Visits in the past 2 weeks
1	0	0	E0(Y1|Xo=0, Covariates)	E0(Y2|Xo=0, Covariates)
2	0	0		
3	1	0		
4	1	0		
5	1	0		

Predict $E_o(Y1|Xo = 0, Covariates)$ using conditional mean functions:

$m1(Xo = 0, Covaraites; \widehat{\beta_1}, \widehat{\sigma_{12}}, \widehat{\beta_2},)$

Counterfactural Prediction of Conditional Means (Counterfactural Scenario 2)

Patient ID	Xo (Policy Variable) Actual Insurance Status	Counterfactural Mandated Private Insurance Status	Y1 Number of Physician Office Visits in the past 2 weeks	Y2 Number of Non-Physician Health Professional Office Visits in the past 2 weeks
1	0	1	E1(Y1|Xo=1, Covariates)	E1(Y2|Xo=1, Covariates)
2	0	1		
3	1	1		
4	1	1		
5	1	1		

Predict $E_1(Y1|Xo = 1, Covariates)$ using conditional mean functions:

$m1(Xo = 1, Covaraites; \widehat{\beta_1}, \widehat{\sigma_{12}}, \widehat{\beta_2},)$

- Estimated ATE formula:

$$\widehat{\mathbf{ATE}(\Delta)} = \sum_{i=1}^{n} \frac{1}{n} \{m(X_i^{pre} + \Delta_i, X_{oi}; \hat{\pi})] - E[m(X_i^{pre}, X_{oi}; \hat{\pi})$$