Multivariate Sarmanov Count Data Models

Eugenio J. Miravete

University of Texas at Austin
& Centre for Economic Policy Research

February 12, 2009

Disclaimer:

Despite what you are about to see, I do not consider myself an econometrician but rather an empirical IO economist who happen to run into obscure econometric problems.

Fact:

I first encountered the Sarmanov distribution when searching for arguments to convince an NSF reviewer of a proposal dealing with the estimation of a structural model of nonlinear pricing competition.

Refreshing Your Memory:

Andrews, Econometrica'02, pp. 119-162:

k-step bootstrap: Under very general conditions, we can approximate the parameter estimates of each bootstrap sample by taking k iterations of a Newton-Raphson procedure starting from the ML parameter estimate using the full sample (it also works with (Gauss-Newton).

More Memory Refreshing:

Andrews, Econometrica'00, pp. 399-405:

Rescaled bootstrap (a variation of subsampling) is appropriate to obtain consistent inference when parameters may be on the boundary defined by a set of linear or nonlinear inequality constraints.

A Request:

Questions:

- General: Do you see any concerns in combining k-step bootstrap with subsampling or rescaled bootstrap?
- Particular: Is there anything in my model that makes the combination of these two methods questionable?

Application: Competing with Menus of Tariff Options

- Do firms offer a similar number of tariff options? If they do beyond what the heterogeneity of consumers and cost of offering these options justifies, we should expect them to be positively correlated.
  $\Rightarrow$ The number of tariff options are strategic complements.

- Do firms, on the contrary, try to differentiate themselves by offering a very different menu of tariff options? In this case, the number of tariff options should be negatively correlated.
  $\Rightarrow$ The number of tariff options are strategic substitutes.

Empirical Challenge

- We need an econometric model that allows for the possibility that counts are negatively correlated.
- The model also needs to accommodate for the possibility of underdispersion of counts that, while less frequent, it appears to be present in the data.

Over and Underdispersion:

I am aware of only two univariate models can accommodate both features (ignore hurdle and zero-inflated models):

- Efron (1986): Double Poisson model.
- Winkelmann (1995): Count data model based on a gamma distributed renewal process.

Existing Multivariate Count Data Regression Models:

- There are very few: Kocherlakota-Kocherlakota (1993); Marshall-Olkin (1990); Gourieroux-Monfort-Trognon (1984).
- They are all restricted to the case where the correlation coefficient is positive and of limited range.
- Correlation is modeled as a consequence of the same unobserved heterogeneity that explains over or underdispersion.
  - Most of the times the over/underdispersion of all counts and the correlation among them is modeled as a function of a single parameter.

Features of the Present Model

- It can accommodate both over and underdispersion of the distribution of each count separately.
- Both positive and negative correlations are possible.
- Dispersion and correlation depends on different parameters of the model.
- The model can easily be extended beyond the bivariate case.
- The estimation is not particularly time consuming (bootstraping is a different matter): the likelihood function can always be written in closed form and thus, simulation is not needed to obtain the parameter estimates.
- The range of the correlation coefficient may be smaller than $[-1, 1]$ and is effectively bounded by the estimates of the rest of parameters of the model.

The Sarmanov Family of Distributions

Let y_k, $k=1,2$ denote two random variables with univariate probability density function $f_k(y_k)$ on $A_k \subseteq \mathbb{R}$ and with mean and variance:

$$\mu_k = \int_{-\infty}^{\infty} y_k f_k(y_k) dy_k \qquad \text{and} \qquad \sigma_k^2 = \int_{-\infty}^{\infty} (y_k - \mu_k)^2 f_k(y_k) dy_k \,.$$

This bivariate Sarmanov probability density function is written as:

$$f_{12}(y_1, y_2) = f_1(y_1) f_2(y_2) \times [1 + \omega_{12} \psi_1(y_1) \psi_2(y_2)] \,,$$

where $\psi_k(y_k)$, $k = 1, 2$ are bounded and nonconstant mixing functions such as:

$$\int_{-\infty}^{\infty} \psi_k(y_k) f_k(y_k) dy_k = 0 \,.$$

Sarmanov Distributions on Positive Orthants

Assume that marginal distributions have support on $\mathbb{R}_+$. Lee (1996) shows that the mixing functions are then given by:

$$\psi_k(y_k) = \exp(-y_k) - L_k(1), \qquad \forall y_k \geq 0\,,$$

where:

$$L_k(\zeta) = \int_0^\infty \exp(-\zeta y_k) f_k(y_k) dy_k\,.$$

is the Laplace transform of the assumed marginal distribution evaluated at $\zeta = 1$.

Constraints

For the Sarmanov distribution to be properly defined we need:

$$\omega_{12} \in \mathbb{R} : \; 1 + \omega_{12}\psi_1(y_1)\psi_2(y_2) \geq 0 \qquad \forall y_1, y_2 \,,$$

or equivalently:

$$\underline{\omega}_{12} \leq \omega_{12} \leq \overline{\omega}_{12} \,,$$

where:

$$\underline{\omega}_{12} = \frac{-1}{\max\{L_1(1)L_2(1), [1 - L_1(1)][1 - L_2(1)]\}} \,,$$

$$\overline{\omega}_{12} = \frac{1}{\max\{L_1(1)[1 - L_2(1)], [1 - L_1(1)]L_2(1)\}} \,.$$

Correlation

Additionally:

$$E\left[y_1 y_2\right] = \mu_1 \mu_2 + \omega_{12} \nu_1 \nu_2 \,,$$

$$\nu_k = \int_{-\infty}^{\infty} y_k \psi_k(y_k) f_k(y_k) dy_k = -L_k'(1) - L_k(1)\mu_k \,,$$

$$\rho_{12} = \frac{\omega_{12} \nu_1 \nu_2}{\sigma_1 \sigma_2} \,.$$

Double Poisson Distribution

Let $y_k = 0, 1, 2, \ldots$ be distributed according to a double Poisson distribution with parameters μ_k and θ_k, conditional on a set of regressors $\mathbf{x_k}$ in a sample with $i = 1, 2, \ldots, n$ observations.

The probability function of a double Poisson distribution is:

$$\tilde{f}_k(y_k|\mu_k, \theta_k) = c(\mu_k, \theta_k) f_k(y_k|\mu_k, \theta_k)\,,$$

$$f_k(y_k|\mu_k, \theta_k) = \sqrt{\theta_k}\exp(-\theta_k\mu_k)\exp(-y_k)\frac{{y_k}^{y_k}}{y_k!}\left(\frac{e\mu_k}{y_k}\right)^{\theta_k y_k}\,,$$

$$\frac{1}{c(\mu_k, \theta_k)} = \sum_{y_k=0}^{\infty} f_k(y_k|\mu_k, \theta_k) \simeq 1 + \frac{1-\theta_k}{12\theta_k\mu_k}\left(1 + \frac{1}{\theta_k\mu_k}\right)\,,$$

Over and Underdispersion

$$\mathrm{E}[y_{ki}|\mathbf{x_{ki}}] \simeq \mu_{ki} \,,$$

$$\sigma^2_{ki} = \mathrm{Var}[y_{ki}|\mathbf{x_{ki}}] \simeq \frac{\mu_{ki}}{\theta_k} \,,$$

$$\mu_{ki} = \exp\left(\mathbf{x'_{ki}}\beta_{\mathbf{k}}\right) \,.$$

Approximations

Use Stirling's formula $z! \simeq \sqrt{2\pi z} \cdot z^z \cdot \exp(-z)$ for $z = y_k$ and $z = \theta_k y_k$, respectively to approximate the double Poisson frequency function:

$$f_k(y_k|\mu_k, \theta_k) \simeq \theta_k \exp(-\theta_k \mu_k) \frac{(\theta_k \mu_k)^{\theta_k y_k}}{\Gamma(\theta_k y_k + 1)} ,$$

so that the approximation to the corresponding Laplace transform evaluated at $\zeta = 1$ is:

$$L_k(1|\mu_k, \theta_k) \simeq c(\mu_k, \theta_k)\theta_k \exp(-\theta_k \mu_k) \sum_{y_k=0}^{\infty} \frac{(\theta_k \mu_k)^{\theta_k y_k} \exp(-y_k)}{\Gamma(\theta_k y_k + 1)} .$$

More Approximations

The approximate mixing function for the double Poisson-Sarmanov distribution $\psi_k(y_k|\mu_k, \theta_k)$ is:

$$\exp(-y_k) - c(\mu_k, \theta_k)\theta_k \exp(-\theta_k \mu_k) \sum_{y_k=0}^{\infty} \frac{(\theta_k \mu_k)^{\theta_k y_k} \exp(-y_k)}{\Gamma(\theta_k y_k + 1)} ,$$

and the approximate mixing function weighted mean $\nu_k(\mu_k, \theta_k)$ is:

$$c(\mu_k, \theta_k)\theta_k \exp(-\theta_k \mu_k) \sum_{y_k=0}^{\infty} \frac{(\theta_k \mu_k)^{\theta_k y_k} \exp(-y_k)}{\Gamma(\theta_k y_k + 1)} (y_k - \mu_k) .$$

And more Approximations

Thus, the approximate correlation coefficient is:

$$\rho_{12} \simeq \omega_{12} \prod_{k=1}^{2} Q(\mu_k, \theta_k) \,,$$

where $Q(\mu_k, \theta_k)$ is:

$$\frac{c(\mu_k, \theta_k)\theta_k \exp(-\theta_k \mu_k)}{\sqrt{\mu_k/\theta_k}} \sum_{y_k=0}^{\infty} \frac{(\theta_k \mu_i)^{\theta_k y_k} \exp(-y_k)}{\Gamma(\theta_k y_k + 1)} (y_k - \mu_k) \,.$$

Double Poisson-Sarmanov Probability

The probability of observing simultaneously a pair of counts $\{y_1, y_2\}$ is:

$$f_{12}(y_1,y_2) \simeq \left(\prod_{k=1}^{2} \left\{ c(\mu_k,\theta_k)\theta_k \exp(-\theta_k\mu_k) \frac{(\theta_k\mu_k)^{\theta_k y_k}}{\Gamma(\theta_k y_k+1)} \right\} \right) \times$$

$$\left(1+\rho_{12} \frac{\prod_{m=1}^{2} \left\{ \exp(-y_m) - c(\mu_m,\theta_m)\theta_m \exp(-\theta_m\mu_m) \sum_{y_m=0}^{\infty} \frac{(\theta_m\mu_m)^{\theta_m y_m} \exp(-y_m)}{\Gamma(\theta_m y_m+1)} \right\}}{\prod_{m=1}^{2} Q(\mu_m,\theta_m)} \right)$$

subject to:

$$\underline{\omega}_{12} \prod_{k=1}^{2} Q(\mu_k,\theta_k) \leq \rho_{12} \leq \overline{\omega}_{12} \prod_{k=1}^{2} Q(\mu_k,\theta_k) \qquad \forall\ i.$$

Gamma-Sarmanov Count Data Model:

- There is no time to cover it.
- Nice: Because of the reproductive property of the gamma distribution, the evaluation of a multidimensional gamma-Sarmanov distribution reduces to evaluating a linear combination of products of single dimensional integrals (incomplete gamma functions).
- Not so nice: The acceptable range of the correlation coefficient is nil as soon as a single realization of one of the endogenous counts is zero.

Data Description

Early U.S. Cellular telephone pricing data 1984-1992:

- Number of tariff plans offered by competing duopolists in markets defined around SMSAs.
  - Industry consultancy sources.
- Some market and/or firm specific characteristics.
  - Census, Federal Communications Commission, and industry sources.
- Carrier ownership indicator.
  - Federal Communications Commission.

Table 1: Frequency Distributions of Number of Tariff Options

	1984–1988				*1992*			
	Incumbent		*Entrant*		*Incumbent*		*Entrant*	
Tariff Options	Cases	Rel.Freq.	Cases	Rel.Freq.	Cases	Rel.Freq.	Cases	Rel.Freq.
1	14	0.0269	3	0.0423	51	0.0979	5	0.0704
2	71	0.1363	7	0.0986	76	0.1459	3	0.0423
3	198	0.3800	5	0.0704	122	0.2342	13	0.1831
4	128	0.2457	16	0.2254	162	0.3109	18	0.2535
5	63	0.1209	40	0.5634	55	0.1056	32	0.4507
6	47	0.0902	0	0.0000	55	0.1056	0	0.0000
Mean, (Var.)	3.5681	(1.4651)	4.1690	(1.3996)	3.4971	(1.9774)	3.9718	(1.4563)

Absolute and relative frequency distributions of the number of tariff options offered by each active firm.

- Entrants offers more options than incumbents.
- Slight reduction of options offered over time.
- The unconditional distribution of counts is underdispersed.

Table 2: Correlation Among Number of Tariff Options

	1984–1988							*1992*						
Plans	1	2	3	4	5	6	All	1	2	3	4	5	6	All
1	9	0	1	4	0	0	14	0	0	1	1	1	0	3
2	20	35	11	4	0	1	71	2	1	1	2	1	0	7
3	9	15	55	68	26	25	198	1	0	2	1	1	0	5
4	8	19	42	36	9	14	128	0	0	4	3	9	0	16
5	5	7	9	34	7	1	63	2	2	5	11	20	0	40
6	0	0	4	16	13	14	47	0	0	0	0	0	0	0
All	15	76	122	162	55	55	521	5	3	14	18	32	0	72
Kendall's τ	0.2928				(9.99)			0.1836				(2.26)		

Total cases for each combination of tariff options offered by the incumbent and entrant firm. Rows indicate the number of options of the entrant and columns those of the incumbent. Kendall's τ measures the association among the number of tariff options. The corresponding absolute value t-statistics are shown in parentheses. There are 521 pairs of tariff strategies in the 1984–1988 sample and 72 in the 1992 sample.

- Positive unconditional association suggests that the number of tariff options are strategic complements.
- Firms frequently offer the same number of options:
  - $1984-1988 \Rightarrow 30\%$ of cases.
  - $1992 \Rightarrow 36\%$ of cases.
- Firms frequently offer almost the same number of options:
  - $1984-1988 \Rightarrow 71\%$ of cases.
  - $1992 \Rightarrow 75\%$ of cases.

Table 3: Descriptive Statistics

	Incumbent		*Entrant*	
Variables	Mean	Std.Dev.	Mean	Std.Dev.
PLANS	3.6402	1.2219	3.5541	1.3915
YEAR92	0.1199	0.3252	0.1199	0.3252
COMMUTING	3.1428	0.1512	3.1428	0.1512
POPULATION	0.0793	0.9583	0.0793	0.9583
EDUCATION	2.5752	0.0352	2.5752	0.0352
BUSINESS	3.2840	0.8876	3.2840	0.8876
GROWTH	0.9361	1.0274	0.9361	1.0274
INCOME	3.6406	0.1318	3.6406	0.1318
MULTIMARKET	3.1824	2.2808	3.1824	2.2808
REGULATED	0.5270	0.4997	0.5270	0.4997
AMERITECH	0.1554	0.3626	0.0942	0.2206
BELLATL	0.0574	0.2329	0.0671	0.1725
BELLSTH	0.0878	0.2833	0.0600	0.1652
CENTEL	0.0895	0.2857	0.0541	0.1623
CONTEL	0.0507	0.2195	0.0270	0.1204
GTE	0.1436	0.3510	0.0777	0.1970
MCCAW	0.0000	0.0000	0.2782	0.2473
NYNEX	0.0963	0.2952	0.0550	0.1734
PACTEL	0.0220	0.1467	0.0388	0.1354
SWBELL	0.1334	0.3403	0.0802	0.2174
USWEST	0.0895	0.2857	0.0566	0.1638

All variables are defined in the text. The number of observations is 592.

Estimation

The likelihood function can be written as follows:

$$\mathcal{L}\left(\gamma_{\mathbf{1}}, \gamma_{\mathbf{2}}, \omega_{12}\right) = \sum_{i=1}^{n}\sum_{k=1}^{2} \ln f_k\left(y_{ki}|\mathbf{x_{ki}}, \gamma_{\mathbf{ki}}\right) + \sum_{i=1}^{n} \ln\left[1 + \omega_{12}\prod_{k=1}^{2}\psi_k\left(y_{ki}|\mathbf{x_{ki}}, \gamma_{\mathbf{ki}}\right)\right] .$$

- Since ω_{12} only enters the term between brackets the gradient of the log-likelihood function is block-recursive.
- Iterative estimation alternatively fixing the value of ω_{12} or $\gamma_{\mathbf{1}}$ and $\gamma_{\mathbf{2}}$.
- The infinite sum of the probability of the double Poisson-Sarmanov distribution always converges, although more rapidly for underdispersed distribution of counts.

Table 4: Double Poisson – Sarmanov Regression

	Independent Regressions				*Sarmanov Regression*			
Variables	*Incumbent*		*Entrant*		*Incumbent*		*Entrant*	
CONSTANT	2.3986	(0.47)	–12.5831	(1.79)	2.3967	(0.58)	–12.6123	(2.45)
YEAR92	0.7212	(6.87)	0.6151	(4.18)	0.7115	(4.22)	0.6216	(3.64)
COMMUTING	–1.1548	(1.92)	1.5559	(2.06)	–1.1798	(1.80)	1.5674	(2.57)
POPULATION	–0.0489	(0.40)	0.0743	(0.59)	–0.0475	(0.44)	0.0652	(0.54)
EDUCATION	0.1227	(0.06)	2.8608	(0.97)	0.1284	(0.07)	2.8691	(1.19)
BUSINESS	0.0333	(0.26)	–0.2681	(2.01)	0.0237	(0.22)	–0.2694	(1.86)
GROWTH	0.0891	(1.51)	–0.4534	(6.76)	0.0874	(1.38)	–0.4500	(5.81)
INCOME	1.4633	(2.32)	1.3248	(1.64)	1.4941	(1.86)	1.3347	(1.56)
MULTIMARKET	0.0409	(1.80)	0.1082	(4.06)	0.0394	(1.30)	0.1037	(3.31)
REGULATED	0.0928	(0.87)	0.6520	(4.66)	0.0815	(0.71)	0.6342	(4.88)
AMERITECH	–0.2183	(0.87)	0.3169	(0.63)	–0.2299	(0.82)	0.2406	(0.67)
BELLATL	1.0770	(4.97)	0.1317	(0.31)	1.0957	(3.38)	0.0848	(0.17)
BELLSTH	–1.2825	(6.09)	–0.9200	(2.07)	–1.2362	(3.43)	–0.9414	(1.53)
CENTEL	–0.2719	(1.26)	1.3981	(2.94)	–0.2827	(0.97)	1.3036	(2.71)
CONTEL	–0.8500	(3.70)	–0.7116	(1.42)	–0.8524	(2.07)	–0.7340	(0.92)
GTE	–1.1022	(6.38)	–0.1997	(0.52)	–1.0929	(3.30)	–0.2429	(0.52)
MCCAW			0.8311	(2.76)			0.8508	(2.25)
NYNEX	0.9543	(5.30)	0.9591	(2.37)	0.9632	(3.27)	0.8880	(1.69)
PACTEL	–1.2295	(4.07)	–0.0734	(0.11)	–1.1967	(2.36)	–0.0748	(0.17)
SWBELL	–0.5886	(2.53)	0.0341	(0.06)	–0.5839	(1.83)	–0.0037	(0.01)
USWEST	–0.0150	(0.08)	0.7996	(1.69)	–0.0048	(0.01)	0.7491	(1.52)
θ	3.6324	(16.37)	2.3895	(15.24)	3.6585	(12.93)	2.4113	(16.62)
ρ					0.0396		(3.40)	
$-\ln L$	830.16		942.49		1,766.60			

Marginal effects evaluated at the sample mean of regressors. Endogenous variables are the number of tariff options of of each competing firm. Absolute value, bootsrapped t-statistics are reported between parentheses.

Double Poisson-Sarmanov Estimates

Results:

- The independent count data regression specification is rejected in favor of the double Poisson-Sarmanov model (0.01 p-value).
- Estimate of correlation is positive and small but significant.
- The number of tariff options are strategic complements.
- The distribution of counts is always underdispersed.
- Other estimates differ in a non-systematic way but they normally become less significant once we account for the possibility of correlated counts.

Summary

- The double Poisson-Sarmanov model is the most flexible model of multivariate count data regression available.
- It can easily be extended beyond two dimensions.

- Things to do:
  - k-step bootstrap.
  - Subsampling.
  - Anything else?