Advances in Count Data Regression: IV. Further Topics

A. Colin Cameron
Univ. of Calif. - Davis

28th Annual Workshop in Applied Statistics
Southern California Chapter of the American Statistical Association
Held at University of California - Los Angeles
http://cameron.econ.ucdavis.edu/racd/count.html

March 28 2009

Outline of all Lectures

I. Basic cross-section methods:
   - Poisson, GLM, negative binomial
II. More advanced cross-section methods:
   - Hurdle, zero-inflated, finite mixtures, endogeneity
III. Time series and panel methods
IV. Further Topics:
   - multivariate, maximum simulated likelihood, Bayesian

Outline of Further Topics

- Multivariate data
- Maximum simulated likelihood
- Bayesian methods (very brief)
- Summary

Multivariate data

- Example is number of doctor visits and number of hospital stays.
- Multivariate versions of Poisson and negative binomial exist but are too restrictive
  - e.g. may permit only positive correlation with bound much less than 1.
- Instead approaches are
  - moment-based approach that generalizes seemingly unrelated equations
  - parametric approach that incorporates correlated latent variables
  - parametric approach that uses copulas to introduce correlation given specified marginal distributions

Multivariate data: bivariate Poisson

- Define count variables y_1 and y_2 (Kocherlakota and Kocherlakota, 1993)

$$\begin{aligned} y_1 &= u_1 + u_3, \\ y_2 &= u_2 + u_3, \\ u_j &\sim Poisson[\mu_j],\ j = 1, 2, 3. \end{aligned}$$

- Then joint frequency distribution is

$$\Pr[y_1 = r, y_2 = s] = \exp\left[\mu_1 + \mu_2 + \mu_3\right] \sum_{l=0}^{\min(r,s)} \frac{\mu_1^{r-l}\mu_2^{s-l}\mu_3^{l}}{(r-l)!(s-l)!l!}.$$

  - with marginals $y_1 \sim \mathcal{P}[\mu_1 + \mu_3]$ and $y_2 \sim \mathrm{P}[\mu_2 + \mu_3]$

- Squared correlation coefficient is bounded

$$\begin{aligned} \rho_{12}^2 &= \mu_3^2/[(\mu_1 + \mu_3)(\mu_2 + \mu_3)] \\ 0 &\leq \rho_{12} \leq \mu_3/[\mu_3 + \min(\mu_1, \mu_2)] \end{aligned}$$

- Regression model specifies $\mu_j = \exp(\mathbf{x}_j'\boldsymbol{\beta}_j)$, $j = 1, 2, 3$.
  - The marginal conditional means $\mathrm{E}[y_1] = \mu_1 + \mu_3$ and $\mathrm{E}[y_2] = \mu_2 + \mu_3$ are no longer of simple exponential form.

Multivariate data: moment-based approach

- Extend seemingly unrelated regressions for linear models.
- Specify

$$\begin{aligned} \mathrm{E}[y_1|\mathbf{x}_1] &= \exp(\mathbf{x}_1'\boldsymbol{\beta}_1) \\ \mathrm{E}[y_2|\mathbf{x}_2] &= \exp(\mathbf{x}_2'\boldsymbol{\beta}_2) \end{aligned}$$

- For variances and covariance suppose

$$\begin{aligned} \mathrm{V}[y_1|\mathbf{x}_1] &= \alpha_1 \exp(\mathbf{x}_1'\boldsymbol{\beta}_1) \\ \mathrm{V}[y_2|\mathbf{x}_2] &= \alpha_2 \exp(\mathbf{x}_1'\boldsymbol{\beta}_1) \\ \mathrm{Cov}[y_1, y_2|\mathbf{x}_1, \mathbf{x}_1] &= \rho \exp(\mathbf{x}_1'\boldsymbol{\beta}_1)^{1/2} \exp(\mathbf{x}_1'\boldsymbol{\beta}_2)^{1/2} \end{aligned}$$

- Then estimate by GEE type estimator.
- Similar to univariate Poisson quasi-likelihood except improved efficiency is possible.

Multivariate data: correlated latent variables

- Introduce correlated latent variables (Marshall and Olkin, 1990).
- For example, for Poisson random variables

$$\begin{aligned} y_1|\mathbf{x}_1, \nu_1 &\sim \mathcal{P}[\exp(\nu_1 + \mathbf{x}_1'\boldsymbol{\beta}_1)] \\ y_2|\mathbf{x}_2, \nu_2 &\sim \mathcal{P}[\exp(\nu_2 + \mathbf{x}_2'\boldsymbol{\beta}_2)] \\ (\nu_1, \nu_2) &\sim \text{bivariate normal} \end{aligned}$$

  - This parametric model incorporates both overdispersion and correlation.
- Integrate out

$$f(y_1, y_2|\mathbf{x}_1, \mathbf{x}_2, \nu_1, \nu_2) = \int f_1(\mathbf{y}_1|\mathbf{x}_1, \nu_1) f_2(\mathbf{y}_2|\mathbf{x}_2, \nu_2) g(\nu_1, \nu_2) d\nu_1 d\nu_2.$$

- Estimation is by maximum simulated likelihood.

Multivariate data: copulas

- Specify appropriate marginals and introduce correlation via copulas (Sklar, 1973).
- We begin with the Gaussian copula.
- Convert Y_1 and Y_2 to Y_1^* and Y_2^* with standard normal marginals as follows
  - For $j = 1, 2$ convert Y_j with c.d.f. $F_j(\cdot)$ into $U_j \sim \text{Uniform}(0, 1)$

$$U_j = F_j(Y_j),\ j = 1, 2.$$

  - For $j = 1, 2$ convert U_j to Y_j^* with standard normal c.d.f. $G_j(\cdot)$

$$Y_j^* = G_j^{-1}(U_j),\ j = 1, 2.$$

- Then model Y_1^* and Y_2^* as regular bivariate normal with zero means, unit variances and covariance ρ.

- More generally for other copulas (Sklar, 1973)
  - A bivariate copula $C(u_1, u_2)$ is a bivariate c.d.f. with marginals that are Uniform$(0, 1)$.
  - There are many copulas other than the Gaussian that can be applied to U_1, U_2.
  - Extension to multivariate case is immediate for the Gaussian copula.
- For regression case distribution of Y_j depends on regressors and parameters to be estimated.
- For Guassian copula can do two-stage estimator
  - Estimate parameters of marginal distribution of $y_j|\mathbf{x}_j$, $j = 1, 2$.
  - Estimate parameters of the copula function (ρ for bivariate case).
- For Guassian copula FIML is difficult.
  Pitt, Chan and Kohn (2006) propose Bayesian method.

Multivariate data: copulas for counts

- The preceding implicitly assumed that Y_1 and Y_2 are continuous.
- For counts there is added complication that Y is discrete.
  - See for example Cameron, Li, Trivedi and Zimmer (2004).
- Also can use copulas to jointly model discrete and continuous random variables.

Maximum simulated likelihood: motivation

- Several models lead to low-dimensional integrals that have no closed-form solution.
- Cross-section example
  - Poisson with normally distributed intercept to capture overdispersion
- Panel example
  - Poisson with normally distributed intercept to capture correlation over time for a given individual
- Multivariate example
  - Poisson with multivariate normally distributed intercepts to capture correlation across variables.
- Can approximate integral by Gaussian quadrature or by Monte Carlo integration.

- Numerical integration using Gaussian quadrature:

$$\begin{aligned} I &= \int_a^b f(x)\,dx \\ &= \int_c^d w(x)r(x)dx \text{ where } r(x) = f(x)/w(x) \\ &\simeq \textstyle\sum_{j=1}^m w_j r(x_j), \end{aligned}$$

- Here three common choices of $w(x)$ are:
  - $w(x) = e^{-x^2}$ for $[c,d] = (-\infty,\infty)$ Gauss-Hermite
  - $w(x) = e^{-x}$ for $[c,d] = (0,\infty)$ Gauss-Laguerre
  - $w(x) = 1$ for $[c,d] = [-1,1]$ Gauss-Legendre
- Tables give good choices (for given m) of
  - weights w_j
  - evaluation points x_j.

- Monte Carlo integration when $g(x)$ is a density:

$$\begin{aligned} \mathrm{E}[h(x)] &= \int_a^b h(x)g(x)dx \\ &\simeq S^{-1}\textstyle\sum_{s=1}^S h(x^s) \end{aligned}$$

- Here $\{x^s,\ s = 1,...,S\}$ is a Monte Carlo sample of S pseudo-random numbers from the density $g(x)$.
- So to compute $\mathrm{E}[h(x)]$ when x has density $g(x)$
  - make many draws of x
  - compute $h(x)$ each time
  - average these.

Maximum simulated likelihood

- Consider Poisson with normally distributed intercept

$$\begin{aligned} y &\sim Poisson[\exp(\mathbf{x}'\boldsymbol{\beta} + v)] \text{ with density } g(y|\mathbf{x},\boldsymbol{\beta},v) \\ v &\sim \mathcal{N}[0,\sigma^2] \text{ with density } h(\sigma). \end{aligned}$$

- Then a typical observation has density

$$\begin{aligned} f(y|\mathbf{x},\boldsymbol{\beta},\sigma) &= \int g(y|\mathbf{x},\boldsymbol{\beta},v)h(v|\sigma) \\ &= \int_{-\infty}^{\infty} \frac{\exp(-e^{\mathbf{x}'\boldsymbol{\beta}+v})(e^{\mathbf{x}'\boldsymbol{\beta}+v})^y}{y!}\frac{1}{\sqrt{2\pi\sigma}}e^{-v^2/2\sigma^2}dv \\ &= \int_{-\infty}^{\infty} \frac{\exp(-e^{\mathbf{x}'\boldsymbol{\beta}+\sigma u})(e^{\mathbf{x}'\boldsymbol{\beta}+\sigma u})^y}{y!}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du \\ &\simeq S^{-1}\textstyle\sum_{s=1}^S \dfrac{\exp(-e^{\mathbf{x}'\boldsymbol{\beta}+\sigma u^s})(e^{\mathbf{x}'\boldsymbol{\beta}+\sigma u^s})^y}{y!} \end{aligned}$$

where u^s, $s = 1,...,S$ are S draws from $\mathcal{N}[0,1]$ and $S \to \infty$.

- The MLE of $(\boldsymbol{\beta},\sigma)$ maximizes

$$\textstyle\sum_{i=1}^N \ln f(y_i|\mathbf{x}_i,\boldsymbol{\beta},\sigma).$$

- The maximum simulated likelihood estimator (MSL) maximizes

$$\begin{aligned} &\textstyle\sum_{i=1}^N \ln\left(S^{-1}\sum_{s=1}^S g(y_i|\mathbf{x}_i,\boldsymbol{\beta},v_i^s)\right) \\ &= \textstyle\sum_{i=1}^N \ln\left(S^{-1}\sum_{s=1}^S \dfrac{\exp(-e^{\mathbf{x}_i'\boldsymbol{\beta}+\sigma u_i^s})(e^{\mathbf{x}_i'\boldsymbol{\beta}+\sigma u_i^s})^y}{y_i!}\right) \end{aligned}$$

- Same distribution as MLE as $S \to \infty$.
- Biased for finite S as unbiased for $f(y_i)$ but not for $\ln f(y_i)$
  - Make S large but then computationally expensive as need to do this for each round of Newton's iterative method.

Bayesian methods

- Bayesian methods
  - Density or likelihood $f(\mathbf{y}|\boldsymbol{\theta})$
  - Prior density $\pi(\boldsymbol{\theta})$
  - Posterior density

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})}$$

- Problem is that closed form solution for $f(\boldsymbol{\theta}|\mathbf{y})$ is rarely available.

- Bayesian methods can be used to enable Bayesian inference
  - different from classical statistical inference
  - can incorporate prior beliefs on parameters $\boldsymbol{\theta}$.
- Bayesian methods can be used to enable classical inference
  - MLE is the Bayesian posterior mode if prior is diffuse
  - computationally faster then MSL in high-dimensional problems

- Recent computational advances make it possible to do Bayesian analysis when there is no closed form solution for the posterior density.
  - Importance sampling
  - Markov chain Monte Carlo (MCMC) methods
  - data augmentation
- Count example is Chib, Greenberg and Winkelmann (1998).

Bayesian example

- Consider Poisson with normally distributed intercept

$$\begin{aligned} y &\sim Poisson[\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)] \\ v_i &\sim \mathcal{N}[0, \sigma^2]. \end{aligned}$$

- Problem is that because v_i is not observed we need to integrate it out.
- A Bayesian solution is data augmentation
  - generate values of the latent variables v_i and treat as data (like y_i)
- Actually more convenient to do this for

$$z_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i).$$

- The Bayesian procedure comprises
- 1. Joint density of data: y_i and z_i (given $\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2$) is

$$y_i, z_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2 \sim Poisson[\mu_i = e^{z_i}] \times \mathcal{N}[z_i, \sigma^2].$$

- 2. Priors for the parameters: $\boldsymbol{\beta}$ and σ^2

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}[\mathbf{0}, k\mathbf{I}] \\ \sigma^{-2} &\sim Gamma\left[\frac{b}{2} \times \left(\frac{c}{2}\right)^{-1}\right] \end{aligned}$$

  where $k = 10$ and $b = 5$ and $c = 10$ give diffuse priors.
- 3. Posterior density for $\boldsymbol{\beta}, \sigma^2, z_i|y_i, \mathbf{x}_i$ is computed recursively using MCMC algorithm.

MCMC for this example

- MCMC algorithm blocks as $\mathbf{z}_i, \boldsymbol{\beta}, \sigma^2$
  - $p(\mathbf{z}_i|\boldsymbol{\beta}, \sigma^2, \mathbf{x}_i, y_i)$ needs Metropolis Hastings
  - $p(\boldsymbol{\beta}|-)$ is $\boldsymbol{\beta} \sim \mathcal{N}[\overline{\boldsymbol{\beta}}, \overline{\mathbf{H}}_{\boldsymbol{\beta}}^{-1}]$
  - $p(1/\sigma^2|-)$ is $\sigma^{-2} \sim Gamma\left[? \times (?)^{-1}\right]$

Summary of count regression

- Cross-section count data basic approaches:
  - Moment-based
    - ★ $\mathrm{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$ and do Poisson QMLE with robust s.e.'s
  - Fully parametric
    - ★ MLE of models richer than Poisson.
- Time series count data:
  - No standard preferred model.

- Panel count data
  - Econometricians focus on multiplicative individual specific effect.
- Fixed effects
  - Use quasi-difference $\mathrm{E}\left[\left(y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_i\right)|\mathbf{x}_{i1,...,}\mathbf{x}_{iT}\right] = 0$ with robust s.e.'s
  - For dynamic models use related quasi first-difference.
- Random effects
  - Population-averaged approach
  - Fully parametric MLE of models such as Poisson-gamma.

- The cross-section and static panel count models can be estimated in STATA, LIMDEP and TSP.
- Count methods also exist (though no off-the-shelf programs) for the usual complications
  - Endogenous regressors
  - Sample selection
  - Multivariate data
  - Measurement error
  - Bayesian methods
  - Semiparametric methods

References

- Materials for these lectures are at http://cameron.econ.ucdavis.edu/racd/count.html
  - Stata program, dataset and output used for talk
  - References from draft version of Pravin K. Trivedi and Murat K. Munkin, "Recent Developments in Cross section and Panel Count Models,"' January 27, 2009.
  - Pravin Trivedi's website (he will post final version of his paper soon).
- Multivariate data: Copulas

  - Cameron, A.C., T. Li, P.K. Trivedi, and D.M. Zimmer (2004), "Modeling the Differences in Counted Outcomes using Bivariate Copula Models: with Application to Mismeasured Counts," *Econometrics Journal*, 7, 566-584.
  - Pitt, M., D. Chan, and R. Kohn (2006), "Efficient Bayesian inference for Gaussian copula regression," *Biometrika*, 93, 537-554.
  - Trivedi, P.K. and D.M. Zimmer (2007), "Copula modeling: an introduction for practitioners," *Foundations and Trends in Econometrics*, 1, 1-110.

- Simulated maximum likelihood

  - Gouriéroux, C., and A. Monfort (1996), *Simulation Based Econometrics Methods*, New York, Oxford University Press.
  - Trivedi, P.K. and Munkin, M.K. (1999), "Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with Application", *Econometrics Journal*, 2(1), 29-48.

- Bayesian methods

  - Chib, S., E.Greenberg and R. Winkelmann (1998), "Posterior Simulation and Bayes Factors in Panel Count Data," *Journal of Econometrics*, 86, 33-54.