Advances in Count Data Regression:
II. Additional cross-section methods

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 additional cross-section count methods

- Introduction
- Censored and truncated data
- Richer parametric models
  - hurdle model
  - zero-inflated model
  - continuous mixtures
  - hierarchical models
  - model comparison
- Finite mixtures model
- Endogenous regressors
- Quantile regression

Counts left-truncated at zero

- Sampling rule is such that observe only y and $\mathbf{x}$ for $y \geq 1$
  i.e. only those who participate at least once are in sample.
- Truncated density (given untruncated density $f(y|\mathbf{x}, \boldsymbol{\theta})$) is

$$f(y|\mathbf{x}, \boldsymbol{\theta}, y \geq 0) = \frac{f(y|\mathbf{x}, \boldsymbol{\theta})}{\Pr[y \geq 0|\mathbf{x}, \boldsymbol{\theta}]} = \frac{f(y|\mathbf{x}, \boldsymbol{\theta})}{[1 - f(0|\mathbf{x}, \boldsymbol{\theta})]}.$$

- MLE is inconsistent if any aspect of the parametric model is misspecified.
- Need to assume that the process for nonzeroes is the same as zeroes.
  - e.g. If data are on annual number of hunting trips for only those who hunted this year, then a missing 0 is interpreted as being for a hunter who did not hunt this year (rather than for all people).

Counts right-censored

- Sampling rule is that observe only 0, 1, 2, ..., $c-1$, c or more i.e. Only record counts up to c and then any value above c.
- Censored density (given uncensored density $f(y|\mathbf{x},\boldsymbol{\theta})$ and cdf is $F(y|\mathbf{x},\boldsymbol{\theta})$)

$$\begin{cases} f(y|\mathbf{x},\boldsymbol{\theta}) & y \leq c-1 \\ 1-F(c-1|\mathbf{x},\boldsymbol{\theta}) = 1-\sum_{j=0}^{c-1} f(j|\mathbf{x},\boldsymbol{\theta}) & y = c \end{cases}$$

- Log-likelihood (where $d_i = 1$ if uncensored and $d_i = 0$ if censored)

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \{d_i \ln f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) + (1-d_i)\ln(1-\sum_{j=0}^{c-1} f(j|\mathbf{x}_i,\boldsymbol{\theta}))\}$$

- MLE is inconsistent if any aspect of the parametric model is misspecified
  - So pick a good density - at least negative binomial.

Counts recorded in intervals

- Sampling rule is that observe only counts in ranges. e.g. 0, 1-4, 5-9, 10 and above.
- Interval density is simply

$$\Pr[a \leq y \leq b] = \sum_{j=a}^{b} f(j|\mathbf{x},\boldsymbol{\theta}).$$

- Let interval ranges by $[a_0, a_1-1]$, $[a_1, a_2-1]$, ..., $[a_m, a_{m+1})$, where $a_0 = 0$, $a_{m+1} = \infty$.
  Let d_k be binary indicators for whether in interval k $(k = 0, ..., m)$.
  Then

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[\sum_{k=0}^{m} d_{ij} \ln \left(\sum_{k=a_k}^{a_{k+1}-1} f(j|\mathbf{x},\boldsymbol{\theta}) \right) \right].$$

- MLE is inconsistent if any aspect of the parametric model is misspecified.
- For convenience could instead use ordered logit or probit here.

Left-truncated at 0 (11% truncated) & right-censored at 10 (26% censored) are less efficient than NB on the complete data.

```
. estimates table NBREG ZTNB NBCENS10, equation(1) b(%10.4f) se  stats(N ll)
```

variable	NBREG	ZTNB	NBCENS10
#1			
private	0.1641	0.1096	0.1258
	0.0332	0.0345	0.0409
medicaid	0.1003	0.0972	0.0653
	0.0454	0.0470	0.0559
age	0.2941	0.2719	0.2276
	0.0602	0.0625	0.0735
age2	-0.0019	-0.0018	-0.0015
	0.0004	0.0004	0.0005
educyr	0.0287	0.0266	0.0197
	0.0042	0.0044	0.0052
actlim	0.1895	0.1955	0.0977
	0.0348	0.0355	0.0430
totchr	0.2776	0.2227	0.2147
	0.0121	0.0124	0.0151
_cons	-10.2975	-9.1902	-7.8000
	2.2474	2.3376	2.7462
lnalpha			
_cons	-0.4453	-0.5260	
	0.0307	0.0419	
eq2			
_cons			1.0134
			0.0378
statistics			
N	3677	3276	3677
ll	-1.059e+04	-9452.8990	-7796.8328

legend: b/se

Richer parametric models

- Data frequently exhibit "non-Poisson" features:
  - Overdispersion: conditional variance exceeds conditional mean whereas Poisson imposes equality.
  - Excess zeros: higher frequency of zeros than predicted by Poisson.
- This provides motivation for richer parametric models than basic Poisson.
- Some models still have $\mathrm{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$
  - Then richer model can provide more efficient estimates.
- Other models imply $\mathrm{E}[y|\mathbf{x}] \neq \exp(\mathbf{x}'\boldsymbol{\beta})$
  - Then Poisson QMLE is inconsistent
  - And marginal effects and coefficient interpretation more difficult.

Hurdle model or two-part model

- Suppose zero counts are determined by a different process to positive counts.
  - Zeros: density $f_1(y|\mathbf{x}_1,\boldsymbol{\theta}_1)$ so $\Pr[y=0]=f_1(0)$ and $\Pr[y>0]=1-f_1(0)$.
  - Positives: density $f_2(y|\mathbf{x}_2,\boldsymbol{\theta}_2)$ so truncated density $f_2(y)/(1-f_2(0))$.
- e.g. First - do I hunt this year or not?
  Second - given I chose to hunt, how many times (≥ 1)?
- Combined density is

$$f(y|\mathbf{x}_1,\mathbf{x}_1,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\begin{cases} f_1(y|\mathbf{x}_1,\boldsymbol{\theta}_1) & y=0 \\ \dfrac{1-f_1(0|\mathbf{x}_1,\boldsymbol{\theta}_1)}{1-f_2(0|\mathbf{x}_2,\boldsymbol{\theta}_2)}\times f_2(y|\mathbf{x}_2,\boldsymbol{\theta}_2) & y\geq 1 \end{cases}$$

- MLE is inconsistent if any aspect of model misspecified.

- Conditional mean is now

$$\mathrm{E}[y|\mathbf{x}]=\Pr[y_1>0|\mathbf{x}_1]\times \mathrm{E}_{y_2>0}[y_2|y_2>0,\mathbf{x}_2].$$

- This makes marginal effects more complicated.
- Example: $f_1(\cdot)$ is logit and $f_2(\cdot)$ is negative binomial.
- Then

$$\mathrm{E}[y|\mathbf{x}]=\Lambda(\mathbf{x}_1'\boldsymbol{\beta})\times\exp(\mathbf{x}_2'\boldsymbol{\beta})/[1-(1+\alpha_2\exp(\mathbf{x}_2'\boldsymbol{\beta}))^{-1/\alpha_2}],$$

where $\Lambda(z)=e^z/(1+e^z)$.

Hurdle model - logit and negative binomial

```
. hnblogit docvis $xlist, nolog

Negative Binomial-Logit Hurdle Regression        Number of obs   =      3677
                                                 Wald chi2(7)    =    309.90
Log likelihood = -10493.225                      Prob > chi2     =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logit						
private	.6586978	.1264608	5.21	0.000	.4108393	.9065563
medicaid	.0554225	.1726694	0.32	0.748	-.2830032	.3938483
age	.542878	.2238845	2.42	0.015	.1040724	.9816835
age2	-.0034989	.0014957	-2.34	0.019	-.0064304	-.0005673
educyr	.047035	.0155706	3.02	0.003	.0165171	.0775529
actlim	.1623927	.1523743	1.07	0.287	-.1362554	.4610408
totchr	1.050562	.0671922	15.64	0.000	.9188676	1.182256
_cons	-20.94163	8.335138	-2.51	0.012	-37.2782	-4.605058
negbinomial						
private	.1095566	.0345239	3.17	0.002	.041891	.1772222
medicaid	.0972308	.0470358	2.07	0.039	.0050423	.1894193
age	.2719031	.0625359	4.35	0.000	.149335	.3944712
age2	-.0017959	.000416	-4.32	0.000	-.0026113	-.0009805
educyr	.0265974	.0043937	6.05	0.000	.0179859	.035209
actlim	.1955384	.0355161	5.51	0.000	.125928	.2651487
totchr	.2226967	.0124128	17.94	0.000	.1983681	.2470252
_cons	-9.190165	2.337592	-3.93	0.000	-13.77176	-4.608569
/lnalpha	-.525962	.0418671	-12.56	0.000	-.60802	-.443904

```
AIC Statistic =      5.712
```


Zero-inflated model (or with-zeroes model)

- Suppose there is an additional reason for zero counts
  - Extra model for 0: density $f_1(y|\mathbf{x}_1,\boldsymbol{\theta}_1)$
  - Usual model for 0: realization of 0 from density $f_2(y|\mathbf{x}_2,\boldsymbol{\theta}_2)$.
- e.g. Some zeroes are mismeasurement and some are true zeros.
- Zero-inflated model has density

$$f(y|\mathbf{x}_1,\mathbf{x}_1,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \begin{cases} f_1(0|\mathbf{x}_1,\boldsymbol{\theta}_1)+[1-f_1(0|\mathbf{x}_1,\boldsymbol{\theta}_1)]\times f_2(0|\mathbf{x}_2,\boldsymbol{\theta}_2) & y=0 \\ [1-f_1(0|\mathbf{x}_1,\boldsymbol{\theta}_1)]\times\cdot f_2(y|\mathbf{x}_2,\boldsymbol{\theta}_2) & y\geq 1 \end{cases}$$

- MLE is inconsistent if any aspect of model misspecified.
- Not used much in econometrics - hurdle model more popular.

Zero-inflated negative binomial

```
. zinb docvis $xlist, inflate($xlist) vuong nolog

Zero-inflated negative binomial regression        Number of obs   =      3677
                                                  Nonzero obs     =      3276
                                                  Zero obs        =       401

Inflation model = logit                           LR chi2(7)      =    588.93
Log likelihood  = -10492.88                       Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
docvis       |
     private |   .1289797    .032987     3.91   0.000     .0643264     .193633
    medicaid |   .1091956    .044511     2.45   0.014     .0219556    .1964356
         age |   .2847325   .0589577     4.83   0.000     .1691776    .4002874
        age2 |  -.0018781   .0003922    -4.79   0.000    -.0026469   -.0011093
      educyr |   .0253991   .0041432     6.13   0.000     .0172786    .0335196
      actlim |   .1737716   .0336464     5.16   0.000     .1078258    .2397173
      totchr |    .229991   .0120795    19.04   0.000     .2063156    .2536663
       _cons |  -9.680235   2.204161    -4.39   0.000    -14.00031   -5.36016
-------------+----------------------------------------------------------------
inflate      |
     private |  -.9152675   .2758402    -3.32   0.001    -1.455904   -.3746307
    medicaid |   .3487142   .3372848     1.03   0.301    -.3123519     1.00978
         age |  -.4357439   .5156094    -0.85   0.398     -1.44632    .5748319
        age2 |    .002805   .0034886     0.80   0.421    -.0040326    .0096426
      educyr |   -.08423    .0339273    -2.48   0.013    -.1507263   -.0177336
      actlim |  -.8241735   .4825621    -1.71   0.088    -1.769978    .1216309
      totchr |  -2.985208   .6860952    -4.35   0.000     -4.32993   -1.640486
       _cons |   17.09618   18.97318     0.90   0.368    -20.09057    54.28294
-------------+----------------------------------------------------------------
    /lnalpha |  -.5848279   .0349792   -16.72   0.000    -.6533859   -.5162699
-------------+----------------------------------------------------------------
       alpha |   .5572017   .0194905                      .5202812    .5967423
------------------------------------------------------------------------------
Vuong test of zinb vs. standard negative binomial: z =     6.48  Pr>z = 0.0000
```


Continuous mixture models

- Mixture motivation for negative binomial assumes $y|\theta \sim Poisson\ (\theta)$ where $\theta = \lambda\upsilon$ is the product of two components:
  - observed individual heterogeneity $\lambda = \exp(\mathbf{x}'\boldsymbol{\beta})$
  - unobserved individual heterogeneity $\upsilon \sim Gamma[1,\ \alpha]$.
- Integrating out

$$h(y|\lambda) = \int f(y|\lambda, \upsilon)g(\upsilon)d\upsilon = \int [e^{-\lambda\upsilon}(\lambda\upsilon)^y/y!] \times g(\upsilon)d\upsilon$$

gives $y|\lambda \sim NB\ [\lambda,\ \lambda + \alpha\lambda^2]$ if $\upsilon \sim Gamma[1,\ \alpha]$.

- Different distributions of υ lead to different models
  - e.g. Poisson-lognormal mixture (random effects model)
  - e.g. Poisson-Inverse Gaussian.
- Even if no closed form solution can estimate using
  - numerical integration (one-dimensional) e.g. Gaussian quadrature.
  - Monte Carlo integration e.g. maximum simulated likelihood.

Hierarchical models

- For multi-level surveys cross-section data individuals i may be in cluster j
  - e.g. patient i in hospital j
  - e.g. individual i in household j or village j
- Hierarchical model or generalized linear mixed model example

$$\begin{aligned} y_i &\sim \text{Poisson}[\mu_{ij} = \exp(\mathbf{x}_{ij}'\boldsymbol{\beta}_j + \varepsilon_{ij})] \\ \boldsymbol{\beta}_j &= \mathbf{W}_j\gamma + \mathbf{v}_j \\ \varepsilon_{ij} &\sim \mathcal{N}[0, \sigma_\varepsilon^2] \\ \mathbf{v}_j &\sim \mathcal{N}[\mathbf{0}, \text{Diag}[\sigma_{jk}^2]] \end{aligned}$$

  - Estimate by MLE or by Bayesian methods.

Model comparison for fully parametric models

- Choice between nested models using likelihood ratio tests
  - e.g. Poisson versus negative binomial.
- Choice between non-nested models using Vuong's (1989) likelihood ratio test
  - e.g. Zero-inflated NB versus NB
- Choice between non-nested mixture models using penalized log-likelihood
  - Akaike's information criterion (AIC) and extensions ($q = \#$ parameters)

$$\begin{aligned} AIC &= -2\ln L + 2q \\ BIC &= -2\ln L + qk\ln N \\ CAIC &= -2\ln L + q(1+\ln)N \end{aligned}$$

  - Prefer model with small AIC or BIC.
  - AIC penalty for larger model too small. Bayesian IC (BIC) better.

- Compare predicted means: $\mathsf{E}[y|\mathbf{x},\widehat{\boldsymbol{\theta}}]$.
- Compare observed frequencies $\overline{p}_j$ to average predicted frequencies

$$\widehat{p}_j = N^{-1}\sum_{i=1}^{N}\widehat{p}_{ij},$$

where $\widehat{p}_{ij} = \widehat{\Pr}[y_i = j]$.

Compare AIC, BIC for regular NB, hurdle logit/NB and zero-inflated NB.

Statistics	NBREG	HURDLENB	ZINB
N	3677	3677	3677
ll	-10589.3	-10493.2	-10492.9
aic	21196.7	21020.4	21019.8
bic	21252.6	21126.0	21125.3

Hurdle NB and ZINB are big improvement on regular NB
- lnL is approximately 100 higher than for NB
- AIC and BIC is much smaller (with only 9 extra parameters)
Little difference between Hurdle NB and ZINB.

The conditional means from the three models are similar.

```
. summarize docvis dvnbreg dvhurdle dvzinb
```

Variable	Obs	Mean	Std. Dev.	Min	Max
docvis	3677	6.822682	7.394937	0	144
dvnbreg	3677	6.890034	3.486562	2.078925	41.31503
dvhurdle	3677	6.840676	3.134925	1.35431	31.86874
dvzinb	3677	6.838704	3.135122	.9473827	32.98153

```
. correlate docvis dvnbreg dvhurdle dvzinb
(obs=3677)
```

	docvis	dvnbreg	dvhurdle	dvzinb
docvis	1.0000			
dvnbreg	0.3870	1.0000		
dvhurdle	0.3990	0.9894	1.0000	
dvzinb	0.3983	0.9882	0.9982	1.0000

Finite mixtures model

- Density is weighted sum of two (or more) densities
  - Permits flexible models e.g. bimodal from Poissons.
- For an m-component model

$$f(y|\mathbf{x},\boldsymbol{\theta},\boldsymbol{\pi}) = \textstyle\sum_{j=1}^{m}\pi_j f_j(y|\mathbf{x},\boldsymbol{\theta}_j), \qquad 0 \le \pi_j \le 1,\ \sum_{j=1}^{m}\pi_j = 1.$$

- For a 2-component model

$$f(y|\mathbf{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\pi) = \pi f_1(y|\mathbf{x},\boldsymbol{\theta}_1) + (1-\pi)\pi f_2(y|\mathbf{x},\boldsymbol{\theta}_2)$$

- MLE maximizes

$$\ln L(\boldsymbol{\theta}) = \textstyle\sum_{i=1}^{N}\ln(\pi f_1(y|\mathbf{x},\boldsymbol{\theta}_1) + (1-\pi)\pi f_2(y|\mathbf{x},\boldsymbol{\theta}_2)).$$

  - Can restrict some parameters to be the same. e.g. only intercept differs
  - EM algorithm often used rather than Newton-Raphson.

- Determining the number of components is a nonstandard inference problem as testing at boundary of parameter space.
  - Simple approach is to use BIC or CAIC.
  - Or do appropriate bootstrap for the likelihood ratio test.
- An alternative to MLE is minimum Hellinger distance estimation.

$$d(\boldsymbol{\theta}) = \sum_{k=0}^{\infty} \left[(\overline{p}_k)^{1/2} - \left(\frac{1}{N} \sum_{i=1}^{N} f(y_i = k | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\pi}) \right)^{1/2} \right]^2$$

  - where $\overline{p}_k$ equals fraction of observations with $y_i = k$.
  - attraction is that it is less influenced by outlying observations
  - estimate using an iterative method (HELMIX)

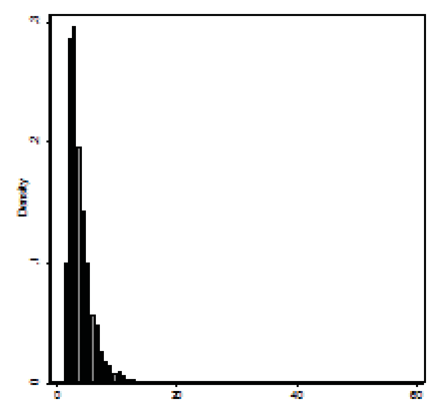
Latent class model

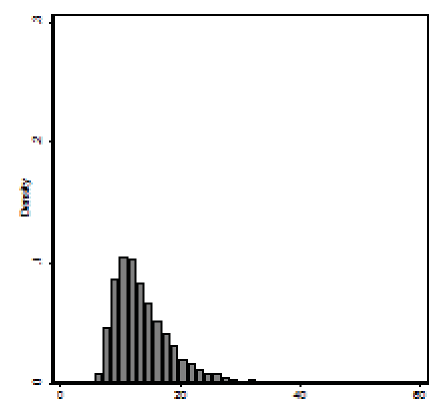
- Finite mixture model can be interpreted as a latent class model.
- There are two types of people (given observables **x**)
  - e.g. "sick" type and "healthy" type
  - there is a probability of being drawn from either type.
- Similar to unobserved heterogeneity in duration data models.


```
2 component Poisson regression                    Number of obs   =      3677
                                                  Wald chi2(14)   =    576.86
Log pseudolikelihood = -11502.686                 Prob > chi2     =    0.0000
```

docvis	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
component1						
private	.2077415	.0560256	3.71	0.000	.0979333	.3175497
medicaid	.1071618	.0964233	1.11	0.266	-.0818245	.2961481
age	.3798087	.100821	3.77	0.000	.1822032	.5774143
age2	-.0024869	.0006711	-3.71	0.000	-.0038022	-.0011717
educyr	.029099	.0067908	4.29	0.000	.0157893	.0424087
actlim	.1244235	.0558883	2.23	0.026	.0148844	.2339625
totchr	.3191166	.0184744	17.27	0.000	.2829074	.3553259
_cons	-14.25713	3.759845	-3.79	0.000	-21.62629	-6.887972
component2						
private	.138229	.0614901	2.25	0.025	.0177106	.2587474
medicaid	.1269723	.1329626	0.95	0.340	-.1336297	.3875742
age	.2628874	.1140355	2.31	0.021	.0393819	.486393
age2	-.0017418	.0007542	-2.31	0.021	-.00322	-.0002636
educyr	.0241679	.0076208	3.17	0.002	.0092314	.0391045
actlim	.1831598	.0622267	2.94	0.003	.0611977	.3051218
totchr	.1970511	.0263763	7.47	0.000	.1453545	.2487477
_cons	-8.051256	4.28211	-1.88	0.060	-16.44404	.3415266
/imlogitpi1	.877227	.0952018	9.21	0.000	.690635	1.063819
pi1	.7062473	.0197508			.6661082	.7434197
pi2	.2937527	.0197508			.2565803	.3338918

Component 1 occurs with probability 0.71 and is low use.
Component 2 occurs with probability 0.29 and is high use.





```
. quietly fmm docvis $xlist, vce(robust) components(2) mixtureof(poisson)
. quietly predict dvfit1, equation(component1)
. quietly predict dvfit2, equation(component2)
. quietly predict dvcombinedfit
. summarize dvfit1 dvfit2 dvcombinedfit docvis

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      dvfit1 |      3677    3.801692    2.176922   .9815563   27.28715
      dvfit2 |      3677    13.95943    5.077463   5.615584   55.13366
dvcombined~t |      3677    6.785555    3.013985   2.342815   35.46714
      docvis |      3677    6.822682    7.394937          0        144
```

Log-likelihood comparison across models:
Poisson -15019; 2-component Poisson -11052; 2-component NB2 -10534; 2-component NB1 -10493.
Last is almost exactly same as hurdle NB and ZINB (-10493).

Endogenous regressor: linear model

- Begin with review of the linear regression model: $y_i = \mathbf{x}_i'\beta + u_i$.
- If regressors are correlated with error then OLS is inconsistent.
  - Reason: OLS $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ so

$$\begin{aligned} \text{plim}\,\widehat{\beta} &= \beta + (\text{plim}\,N^{-1}\mathbf{X}'\mathbf{X})^{-1}\,\text{plim}\,N^{-1}\mathbf{X}'\mathbf{u} \\ &\neq \beta \text{ if } \text{plim}\,N^{-1}\mathbf{X}'\mathbf{u} \neq \mathbf{0}. \end{aligned}$$

- Solution: Assume the existence of an instrument z where
  - changes in z are associated with changes in x
  - but changes in z do not led to change in y (aside from indirectly via x)

$$\begin{array}{ccccc} z & \longrightarrow & x & \longrightarrow & y \\ & & \uparrow & \nearrow & \\ & & u & & \end{array}$$

- Leads to instrumental variables (IV) estimator and two-stage least squares (2SLS) estimator.

- Formally key assumption is:

$$\text{E}[u_i|\mathbf{z}_i] = 0$$

- Just-identified case (# instruments = # endogenous)

$$\widehat{\beta}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

- Over-identified case (# instruments > # endogenous)

$$\widehat{\beta}_{\text{2SLS}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.$$

- Example: log-earnings (y) regressed on years of school (x)
  - ability is an omitted regressor so part of error (u) and clearly correlated with x
  - instrument z is correlated with years of school but not directly with earnings
  - example of z may be distance from school or college.

Linear IV interpretation 1: method of moments

- We have $\text{E}[u_i|\mathbf{z}_i] = 0 \Rightarrow \text{E}[\mathbf{z}_i u_i] = \mathbf{0} \Rightarrow \text{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\beta)] = \mathbf{0}$.
- The IV estimator solves the corresponding sample moment condition

$$\textstyle\sum_{i=1}^N \mathbf{z}_i(y_i - \mathbf{x}_i'\beta) = \mathbf{0}.$$

- Just-identified case can solve

$$\widehat{\beta}_{\text{IV}} = \left(\textstyle\sum_i \mathbf{z}_i\mathbf{x}_i'\right)^{-1}\textstyle\sum_i \mathbf{z}_i y_i = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

- Over-identified case cannot solve so minimize the quadratic form:

$$Q(\beta) = \left(\textstyle\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\beta)\right)'\mathbf{W}\left(\textstyle\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\beta)\right)$$

leads to generalized method of moments (GMM) estimator

$$\widehat{\beta}_{\text{GMM}} = [\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y}$$

  - 2SLS is special case $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$.
- Essentially method of moments based on $\text{E}[\mathbf{z}_i u_i] = \mathbf{0}$.
  - This generalizes to nonlinear models such as Poisson.

Linear IV interpretation 2: two-stage least squares

- Replace endogenous regressor by its predicted value.
- Specify structural equation for y_1 and reduced form equation for y_2
  - Split $\mathbf{x}$ into endogenous regressor y_2 and exogenous regressors $\mathbf{z}_1$
  - Split $\mathbf{z}$ into instrument z_2 for y_2 and other exogenous regressors $\mathbf{z}_1$

$$\begin{array}{ll} \text{Structural eqn:} & y_{1i} = \beta_1 y_{2i} + \mathbf{z}_{1i}'\boldsymbol{\beta}_2 + u_{1i} \\ \text{Reduced-form eqn:} & y_{2i} = \gamma_1 z_{2i} + \mathbf{z}_{1i}'\gamma_2 + v_{2i} \end{array}$$

- Two-stage least squares
  - 1. OLS of y_2 on z_2 and $\mathbf{z}_1$ gives prediction $\widehat{y}_{2i} = \widehat{\gamma}_1 z_{1i} + \mathbf{z}_{2i}'\widehat{\gamma}_2$.
  - 2. OLS of y_{1i} on $\widehat{y}_{2i}$ and $\mathbf{z}_{1i}$ gives estimates equal to IV/2SLS.
- Essentially OLS with y_{2i} replaced by $\widehat{y}_{2i}$
  - This does not generalize to nonlinear models such as Poisson.
  - In particular, it leads to inconsistent estimates.

Linear IV interpretation 3: control function

- Add predicted residual to control for endogeneity.
- Model relationship between structural model error and reduced form error

$$u_{1i} = \alpha v_{2i} + \varepsilon_i$$

where ε_i is independent of v_{2i}, y_{2i} and $\mathbf{z}_{1i}$.

- Then

$$y_{1i} = \beta_1 y_{2i} + \mathbf{z}_{1i}'\boldsymbol{\beta}_2 + \alpha v_{2i} + \varepsilon_i$$

- Control function approach
  - 1. OLS of y_2 on z_2 and $\mathbf{z}_1$ gives residual $\widehat{v}_{2i} = y_{2i} - \widehat{\gamma}_1 z_{1i} - \mathbf{z}_{2i}'\widehat{\gamma}_2$.
  - 2. OLS of y_{1i} on y_{2i}, $\mathbf{z}_{1i}$ and $\widehat{v}_{2i}$ gives estimates equal to IV/2SLS.
- Essentially OLS with y_{2i} augmented by the control for endogeneity $\widehat{v}_{2i}$
  - This generalizes to nonlinear models such as Poisson

Poisson endogenous method 1: nonlinear GMM

- Problem is

$$\mathrm{E}[(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))|\mathbf{x}_i] \neq \mathbf{0}.$$

- Assume existence of instruments $\mathbf{z}_i$ such that

$$\begin{array}{rl} & \mathrm{E}[(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))|\mathbf{z}_i] = \mathbf{0} \\ \Rightarrow & \mathrm{E}[\mathbf{z}_i(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))] = \mathbf{0} \end{array}$$

- Just-identified case: $\widehat{\boldsymbol{\beta}}_{\mathrm{MM}}$ solves

$$\sum_{i=1}^n (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i = \mathbf{0}.$$

- Over-identified case $\widehat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ minimizes

$$\left(\sum_{i=1}^n (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i\right)' \mathbf{W} \left(\sum_{i=1}^n (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i\right)$$

  - usually $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ (called nonlinear 2SLS).

- Literature exists on weighting matrix $\mathbf{W}$ and whether to use different moment condition such as

$$\mathrm{E}\left[\frac{(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})}{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\mathbf{z}_i\right] = \mathbf{0}$$

  - Mullahy (1997), Windmeijer and Santos Silva (1997), Windmeijer (2008).

Poisson endogenous: method 2 control function

- Add error in Poisson model (allows for overdispersion and endogeneity)

$$\begin{array}{ll} \text{Structural eqn:} & y_{1i} \sim \text{Poisson}[\mu_i = \exp(\beta_1 y_{2i} + \mathbf{z}_{1i}'\boldsymbol{\beta}_2 + u_{1i})] \\ \text{Reduced-form eqn:} & y_{2i} = \gamma_1 z_{2i} + \mathbf{z}_{1i}'\gamma_2 + v_{2i} \\ \text{Error model:} & u_{1i} = \alpha v_{2i} + \varepsilon_i \end{array}$$

- Then

$$\begin{aligned} \mu_i | y_{2i}, \mathbf{z}_{1i}, v_{2i}, \varepsilon_i &= \exp(\beta_1 y_{2i} + \mathbf{z}_{1i}'\boldsymbol{\beta}_2 + \alpha v_{2i} + \varepsilon_i) \\ &= \exp(\varepsilon_i)\exp(\beta_1 y_{2i} + \mathbf{z}_{1i}'\boldsymbol{\beta}_2 + \alpha v_{2i}) \\ \mu_i | y_{2i}, \mathbf{z}_{1i}, v_{2i} &= \mathrm{E}[\exp(\varepsilon_i)]\exp(\beta_1 y_{2i} + \mathbf{z}_{1i}'\boldsymbol{\beta}_2 + \alpha v_{2i}) \\ &= \exp(\beta_1 y_{2i} + \mathbf{z}_{1i}'\boldsymbol{\beta}_2 + \alpha v_{2i}) \end{aligned}$$

where if ε_i is i.i.d. then $\mathrm{E}[\exp(\varepsilon_i)]$ is a constant that is absorbed in $\boldsymbol{\beta}_2$.

- Control function approach
  - 1. OLS of y_2 on z_2 and $\mathbf{z}_1$ gives residual $\widehat{v}_{2i} = y_{2i} - \widehat{\gamma}_1 z_{1i} - \mathbf{z}_{2i}'\widehat{\gamma}_2$.
  - 2. Poisson of y_{1i} on y_{2i}, $\mathbf{z}_{1i}$ and $\widehat{v}_{2i}$ gives IV estimate.

NL2SLS: Example with `private` (private insurance) endogenous
Instruments are `income` and `ssiratio` (soc sec income / total income)
Estimate by nonlinear 2SLS:

```
. ereturn display
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	.5920658	.3401151	1.74	0.082	-.0745475	1.258679
medicaid	.3186961	.1912099	1.67	0.096	-.0560685	.6934607
age	.3323219	.0706128	4.71	0.000	.1939233	.4707205
age2	-.002176	.0004648	-4.68	0.000	-.003087	-.001265
educyr	.0190875	.0092318	2.07	0.039	.0009935	.0371815
actlim	.2084997	.0434233	4.80	0.000	.1233916	.2936079
totchr	.2418424	.013001	18.60	0.000	.2163608	.267324
cons	-11.86341	2.735737	-4.34	0.000	-17.22535	-6.50146

`private` was 0.142 (0.036) and is now 0.592 (0.340)
standard errors much larger with IV
Also `medicaid` changes a lot. Others change little.

Control function approach for same example.
First-stage: OLS for reduced form

```
. global xlist2 medicaid age age2 educyr actlim totchr

. regress private $xlist2 income ssiratio, vce(robust)

Linear regression                               Number of obs =     3677
                                                F(  8,  3668) =   249.61
                                                Prob > F      =   0.0000
                                                R-squared     =   0.2108
                                                Root MSE      =   .44472
```

private	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
medicaid	-.3934477	.0173623	-22.66	0.000	-.4274884	-.3594071
age	-.0831201	.0293734	-2.83	0.005	-.1407098	-.0255303
age2	.0005257	.0001959	2.68	0.007	.0001417	.0009098
educyr	.0212523	.0020492	10.37	0.000	.0172345	.02527
actlim	-.0300936	.0176874	-1.70	0.089	-.0647718	.0045845
totchr	.0185063	.005743	3.22	0.001	.0072465	.0297662
income	.0027416	.0004736	5.79	0.000	.0018131	.0036702
ssiratio	-.0647637	.0211178	-3.07	0.002	-.1061675	-.0233599
_cons	3.531058	1.09581	3.22	0.001	1.3826	5.679516

Second stage: Poisson with first-stage predicted residual as regressor

```
. predict lpuhat, residual

.
. * Second-stage Poisson with robust SEs
. poisson docvis private $xlist2 lpuhat, vce(robust) nolog

Poisson regression                              Number of obs   =      3677
                                                Wald chi2(8)    =    718.87
                                                Prob > chi2     =    0.0000
Log pseudolikelihood = -15010.614               Pseudo R2       =    0.1303
```

docvis	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	.5505541	.2453175	2.24	0.025	.0697407	1.031368
medicaid	.2628822	.1197162	2.20	0.028	.0282428	.4975217
age	.3350604	.0696064	4.81	0.000	.1986344	.4714865
age2	-.0021923	.0004576	-4.79	0.000	-.0030893	-.0012954
educyr	.018606	.0080461	2.31	0.021	.002836	.034376
actlim	.2053417	.0414248	4.96	0.000	.1241505	.286533
totchr	.24147	.0129175	18.69	0.000	.2161523	.2667878
lpuhat	-.4166838	.249347	-1.67	0.095	-.9053949	.0720272
_cons	-11.90647	2.661445	-4.47	0.000	-17.1228	-6.69013

`private` is 0.551 (0.245) compared to (0.340) for NL2SLS

Should bootstrap to get correct s.e.'s (`lpuhat` is a generated regressor)

```
. * Program and bootstrap for Poisson two-step estimator
. program endogtwostep, eclass
  1.   version 10.1
  2.   tempname b
  3.   capture drop lpuhat2
  4.   regress private $xlist2 income ssiratio
  5.   predict lpuhat2, residual
  6.   poisson docvis private $xlist2 lpuhat2
  7.   matrix `b' = e(b)
  8.   ereturn post `b'
  9. end

. bootstrap _b, reps(400) seed(10101) nodots nowarn: endogtwostep

Bootstrap results                          Number of obs      =      3677
                                           Replications       =       400
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
private	.5505541	.2567815	2.14	0.032	.0472716	1.053837
medicaid	.2628822	.1205813	2.18	0.029	.0265473	.4992172
age	.3350604	.0707275	4.74	0.000	.1964371	.4736838
age2	-.0021923	.0004667	-4.70	0.000	-.0031071	-.0012776
educyr	.018606	.0083042	2.24	0.025	.0023301	.034882
actlim	.2053417	.0412756	4.97	0.000	.124443	.2862405
totchr	.24147	.0134522	17.95	0.000	.2151042	.2678359
lpuhat2	-.4166838	.2617964	-1.59	0.111	-.9297953	.0964276
_cons	-11.90647	2.698704	-4.41	0.000	-17.19583	-6.617104

Here little change in standard errors.

Poisson endogenous method 3: structural approach

- Example with binary endogenous regressor y_{2i} is

Outcome eqn:	$y_{1i} \sim \text{Poisson}[\mu_i = \exp(\beta_1 y_{2i} + \mathbf{z}'_{1i}\boldsymbol{\beta}_2 + \delta_1 u_i)]$
Participation eqn:	$\Pr[y_{2i} = 1] = \Lambda(\mathbf{z}'_{2i}\boldsymbol{\beta}_2 + \lambda_1 u_i)$
Error model:	$u_i \sim \mathcal{N}[0,1]$

  - Estimate by simulated maximum likelihood.
  - Deb and Trivedi (2006).
- Can also extend the two-part (hurdle) model to incorporate selection
  - This allows for correlation due to unobservables between process for $y = 0$ or not and process for positives.
  - Terza (1998).

Quantile regression

- The q^{th} quantile regression estimator $\widehat{\boldsymbol{\beta}}_q$ minimizes over $\boldsymbol{\beta}_q$

$$Q(\boldsymbol{\beta}_q) = \sum_{i:y_i \geq \mathbf{x}'_i\boldsymbol{\beta}}^{N} q|y_i - \mathbf{x}'_i\boldsymbol{\beta}_q| + \sum_{i:y_i < \mathbf{x}'_i\boldsymbol{\beta}}^{N} (1-q)|y_i - \mathbf{x}'_i\boldsymbol{\beta}_q|, \quad 0 < q < 1.$$

  - Example: median regression with $q = 0.5$.
- For count y adapt standard methods for continuous y by:
  - Replace count y by continuous variable $z = y + u$ where $u \sim Uniform[0,1]$.
  - Then reconvert predicted z-quantile to y-quantile using ceiling function.
  - Machado and Santos Silva (2005).

References

- Censored, truncated, hurdle and zero-inflated is in standard tests.
- Finite mixtures:

  - Deb, P. and P.K. Trivedi (1997), "Demand for Medical Care by the Elderly: A Finite Mixture Approach," *Journal of Applied Econometrics*, 12, 313-326.
  - Bago d'Uva, T. (2006), "Latent class models for utilisation of health care," *Health Economics*, 15, 329-343.
  - Böhning, D., and W. Seidel, "Editorial: recent developments in mixture models," *Computational Statistics and Data Analysis*, 41, 349-357.
  - Lu, Z., Y.V. Hui, and A. H. Lee (2003), "Minimum Hellinger Distance Estimation for Finite Mixtures of Poisson Regression Models and Its Applications," *Biometrics*, 59, 1016-1026.
  - Xiang, L., K.K.W. Yau, Y.V. Hui, and A. H. Lee (2008), "Minimum Hellinger Distance Estimation for k-Component Poisson Mixture with Random Effects," *Biometrics*, 64, 508-518.

- Endogenous regressors:
  - Mullahy, J. (1997), "Instrumental Variable Estimation of Poisson Regression Models: Application to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics*, 79, 586-593.
  - Windmeijer, F.A.G., and J.M.C. Santos Silva (1997), "Endogeneity in Count Data Models; an Application to Demand for Health Care," *Journal of Applied Econometrics*, 12, 281-294.
  - Windmeijer, F.A.G. (2008), "GMM for Panel Count Data Models," ch.18 in L. Matyas and P. Sivestre eds., *The Econometrics of Panel Data*, Springer.
  - Deb, P., and Trivedi, P.K. (2006), "Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization," *Econometrics Journal*, 9, 307-331.
  - Terza, J.V. (1998), "Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects," *Journal of Econometrics*, 84, 129-139.
- Quantile regression:
  - Machado J., and J. Santos Silva (2005), "Quantiles for counts," *Journal of American Statistical Association*, 100, 1226–1237.