Wednesday
Counts: Cross-section Extras

A. Colin Cameron
Univ. of Calif. - Davis

. .

Nonlinear Cross-section and Panel Regression
Models for Count Data

. .

Queens University, Canada
Department of Economics

May 13-16, 2019

1. Introduction

- Count data models are for dependent variable $y = 0, 1, 2, \ldots$
- Previously considered basic cross-section
  - Poisson, negative binomial, GLM's
- Now consider
  - richer models: truncated, censored, hurdle, zero-inflated, ..
  - mixture models
  - endogenous regressors.

Outline

2. 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.
- **Consistency in most fully parametric count models requires all parts of the distribution to be correctly specified**
  - **like the tobit model.**

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).
- Stata commands `tpoisson` (poisson) and `tnbreg` (negative binomial)
  - `tnbreg docvis $xlist if docvis>0, nolog`
  - `tpoisson` and `tnbreg` replace `ztp` and `ztnb`

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.
- Stata code up yourself
  - using command `ml` (for user-defined likelihood).

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.
- Stata has no command so need to code up.
- For convenience could instead use **ordered logit or probit** here.

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: Stata addon hnblogit

```
. 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

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

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

Zero-inflated negative binomial: Stata command zinb and zip

```
. 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 $\theta = \lambda v$ is the product of two components:
  - observed individual heterogeneity $\lambda = \exp(\mathbf{x}'\boldsymbol{\beta})$
  - unobserved individual heterogeneity $v\ \sim Gamma[1,\ \alpha]$.
- Integrating out

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

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

- Different distributions of v 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
- The Stata `gsem` command handles models with v and ε_{ij} normal
  - poisson-normal mixture: `gsem (docvis <- $xlist), poisson`
  - poisson-hierarchical model (with cluster on `educyr`): `gsem (docvis <- $xlist M1[educyr]), nbreg`

Model comparison for fully parametric models

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

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

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

- Compare predicted means: $\mathrm{E}[y|\mathbf{x}, \widehat{\boldsymbol{\theta}}]$
  - e.g. `predict dvnbreg, n` after `nbreg docvis $xlist`
- Compare observed frequencies $\overline{p}_j$ to average predicted frequencies

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

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

  - use user written add-on `countfit` or `chi2gof`

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
```

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
    - ★ This will be an integer.
  - Machado and Santos Silva (2005).
- Stata example
  - `qcount docvis $xlist, q(0.5) rep(1000)`
  - `qcount_mfx // Marginal effect at mean`
  - note: current data had variable one that needs to be dropped before `qcount_mfx`

3. 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

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

- For a 2-component model

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

- MLE maximizes

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

  - 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 $\mathbf{x}$)
  - 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.

Two-component negative binomial

- The latent class probabilities are obtained by a logit model
  - $\pi = 1/(1+\exp(-.2286522)) = 0.5569$

```
. * Finite-mixture model using fmm command with constant probabilities
. use mus17data.dta, clear

. fmm 2, nolog: nbreg docvis $xlist

Finite mixture model                             Number of obs     =      3,677
Log likelihood = -10534.237
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.Class	(base outcome)					
2.Class						
_cons	.2286522	.3567562	0.64	0.522	-.4705772	.9278815

- The next output gives the estimated coefficients in the two classes.

```
Class          : 1
Response       : docvis
Model          : nbreg, dispersion(mean)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
docvis						
private	.3292254	.090978	3.62	0.000	.1509117	.5075391
medicaid	.1319335	.1180946	1.12	0.264	-.0995277	.3633948
age	.3818359	.142241	2.68	0.007	.1030487	.6606231
age2	-.0024401	.0009451	-2.58	0.010	-.0042924	-.0005878
educyr	.0383352	.0106471	3.60	0.000	.0174674	.0592031
actlim	.066328	.0964388	0.69	0.492	-.1226887	.2553446
totchr	.5013939	.0493956	10.15	0.000	.4045802	.5982076
_cons	-15.15966	5.370014	-2.82	0.005	-25.6847	-4.634629
/docvis						
lnalpha	-.9786601	.2237814			-1.417264	-.5400566

```
Class          : 2
Response       : docvis
Model          : nbreg, dispersion(mean)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
docvis						
private	.0957578	.0460128	2.08	0.037	.0055744	.1859411
medicaid	.0901739	.0631541	1.43	0.153	-.0336059	.2139536
age	.2691574	.0818963	3.29	0.001	.1086437	.4296712
age2	-.0017901	.0005455	-3.28	0.001	-.0028593	-.000721
educyr	.0241906	.0058636	4.13	0.000	.0126982	.0356829
actlim	.219569	.0466449	4.71	0.000	.1281468	.3109913
totchr	.1758912	.0296181	5.94	0.000	.1178408	.2339417
_cons	-8.645314	3.070642	-2.82	0.005	-14.66366	-2.626967
/docvis						
lnalpha	-.7697485	.0835753			-.933553	-.605944

- Stata only gives marginal effects averaged across the two components.

```
. * Marginal effects averaged across the two components
. margins, dydx(*)

Average marginal effects                        Number of obs     =      3,677
Model VCE    : OIM

Expression   : Predicted mean (# doctor visits), using class probabilities, predict(mu
               outcome(docvis))
dy/dx w.r.t. : private medicaid age age2 educyr actlim totchr
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	1.123926	.2338631	4.81	0.000	.6655627	1.582289
medicaid	.705085	.3179559	2.22	0.027	.0819029	1.328267
age	2.080845	.4228884	4.92	0.000	1.251999	2.909691
age2	-.0136422	.0028106	-4.85	0.000	-.019151	-.0081334
educyr	.1949865	.0302031	6.46	0.000	.1357896	.2541835
actlim	1.211131	.243566	4.97	0.000	.7337508	1.688512
totchr	1.859463	.0975413	19.06	0.000	1.668286	2.050641

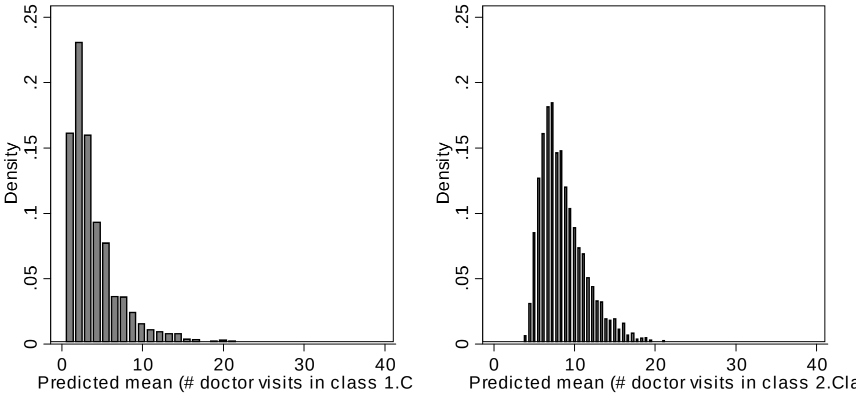
- Class 2 has higher mean and occurs with probability 0.55.

```
. * Marginal predicted means for each component
. estat lcmean

Latent class marginal means                     Number of obs     =      3,677

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
      docvis |   4.477236    .516289     8.67   0.000     3.465329    5.489144
-------------+----------------------------------------------------------------
2            |
      docvis |   8.828413   .4718929    18.71   0.000      7.90352    9.753307
------------------------------------------------------------------------------

.
. * Marginal predicted probabilities for each component
. estat lcprob

Latent class marginal probabilities             Number of obs     =      3,677

--------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       Class |
          1  |   .4430847   .0880334      .2833547    .6155204
          2  |   .5569153   .0880334      .3844796    .7166453
--------------------------------------------------------------
```

- Log-likelihood comparison across models:
  - Poisson -15019; 2-component Poisson -12100; 2-component NB2 -10534; 2-component NB1 -10493.
  - Last is almost exactly same as hurdle NB and ZINB (-10493).
- Obtain the predicted means for each component and plot

```
. * Predict y for two components
. estimates store FMM2

. predict dvfit*
(option mu assumed)

. summarize dvfit1 dvfit2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      dvfit1 |      3,677    4.477236    4.406067   .5638143   75.16145
      dvfit2 |      3,677    8.828413    2.976359   3.799716   31.43396

.
. * Create histograms of fitted values
. quietly histogram dvfit1 if dvfit1< 40, saving(graph1.gph, replace)

. quietly histogram dvfit2 if dvfit1< 40, saving(graph2.gph, replace)

. quietly graph combine graph1.gph graph2.gph, ysize(3) xsize(6) ycommon xcommon
```

- Class 2 has higher mean values.



4. Diagnostics: residuals and influence measures

- Some diagnostics come out of the GLM literature.
- Residuals (for Poisson)
  - Raw: $r_i = (y_i - \widehat{\mu}_i)$
  - Pearson: $p_i = (y_i - \widehat{\mu}_i)/\sqrt{\widehat{\mu}_i}$
  - Deviance: $d_i = sign(y_i - \widehat{\mu}_i)\sqrt{2\{y_i \ln(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\}}$
  - Anscombe: $a_i = 1.5(y_i^{2/3} - \mu_i^{2/3})/\mu_i^{1/6}$
  - Last three will be standardized if $\mathsf{V}[y_i] = \mu_i$.
- Small-sample corrections (for Poisson)
  - Hat matrix: $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$; $\mathbf{W} = \mathrm{Diag}[\widehat{\mu}_i]$.
  - Studentized residual: $p_i^* = p_i/\sqrt{1 - h_{ii}}$ and $d_i^* = d_i/\sqrt{1 - h_{ii}}$.
- Influential observations:
  - Rule of thumb: $h_{ii} > 2K/N$
  - Cook's distance: $C_i = (p_i^*)^2 h_i / K(1 - h_{ii})$ measures change in $\widehat{\beta}$ when observation i is omitted.

- The raw residuals sum to zero due to f.o.c.
- The various residuals are highly correlated.

```
. summarize rraw rpearson rdeviance ranscombe hat cooksd, sep(10)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        rraw |       3677   -7.08e-10    6.808178  -29.12996   136.9007
    rpearson |       3677   -.0060737    2.509354  -4.914746   51.38051
   rdeviance |       3677   -.3438453    2.210397  -6.087067   24.35213
   ranscombe |       3677   -.3634087    2.254119  -6.153762   25.72892
         hat |       3677    .0021757    .0015322   .0007023    .027556
      cooksd |       3677    .0019357    .0177516   6.54e-11    .964198

.
. correlate rraw rpearson rdeviance ranscombe
(obs=3677)

             |     rraw rpearson rdevia~e ransco~e
-------------+------------------------------------
        rraw |   1.0000
    rpearson |   0.9792   1.0000
   rdeviance |   0.9454   0.9669   1.0000
   ranscombe |   0.9435   0.9661   0.9998   1.0000
```

Diagnostics: R-squared measures

- Different interpretations of R^2 in linear model lead to different R^2 in nonlinear model. Most are difficult to interpret in nonlinear models.
- **Simplest:** squared correlation coefficient between y_i and $\widehat{y}_i = \widehat{\mu}_i$

$$R^2_{\text{Cor}} = \widehat{\text{Cor}}^2[y_i, \widehat{y}_i]$$

- Sums of squares measures differ in nonlinear models

$$R^2_{\text{Res}} = 1 - \text{ ResSS}/\text{TotalSS} \neq R^2_{\text{Exp}} = \text{ ExpSS}/\text{TotalSS}$$

- Relative gain in log-likelihood (L_0 is intercept model only)

$$R^2_{\text{RG}} = \frac{\ln L_{fit} - \ln L_0}{\ln L_{\max} - \ln L_0} = 1 - \frac{\ln L_{\max} - \ln L_{fit}}{\ln L_{\max} - \ln L_0}.$$

  - Works for Poisson as $\ln L_{\max}$ occurs when $\mu_i = y_i$.
  - Unlike others $0 \leq R^2_{\text{RG}} < 1$ and R^2_{RG} always increases as add regressors.
- Stata measure is only applicable to binary and multinomial models

$$R^2_{\text{Pseudo}} = 1 - \ln L_{fit} / \ln L_0.$$

Diagnostics: overdispersion test

- In practice can skip this test and just do Poisson with robust s.e.'s.
- $H_0 : \mathsf{V}[y_i|\mathbf{x}_i] = \mathsf{E}[y_i|\mathbf{x}_i]$ versus
  $H_1 : \mathsf{V}[y_i|\mathbf{x}_i] = \mathsf{E}[y_i|\mathbf{x}_i] + \alpha(\mathsf{E}[y_i|\mathbf{x}_i])^2$.
- Test $H_0 : \alpha = 0$ against $H_1 : \alpha > 0$.
- Implement by auxiliary regression

$$((y_i - \widehat{\mu}_i)^2 - y_i)/\widehat{\mu}_i = \alpha\widehat{\mu}_i + \textit{ error}$$

  and do t test of whether the coefficient of $\widehat{\mu}_i$ is zero.
- Test is useful as can also use this test for test of underdispersion, whereas other tests (such as LM of Poisson vs. negative binomial) only test overdispersion.
- If we are just modelling the conditional mean, overdispersion is okay provided robust standard errors are calculated.

- Example of overdispersion test.

```
. * Overdispersion test against V[y|x] = E[y|x] + a*(E[y|x]^2)
. quietly poisson docvis $xlist, vce(robust)

. predict muhat, n

. quietly generate ystar = ((docvis-muhat)^2 - docvis)/muhat

. regress ystar muhat, noconstant noheader
```

ystar	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
muhat	.7047319	.1035926	6.80	0.000	.5016273	.9078365

- Very strongly reject H_0. Data here are overdispersed.

Diagnostics: predicted probabilities

- Now suppose we want to predict probability of 0 doctor visits, 1 doctor visits, ...
- Observed frequency $\overline{p}_j$ (fraction of observations with $y_i = j$).
- Fitted frequency $\widehat{p}_j = N^{-1} \sum_{i=1}^{N} \widehat{p}_{ij}$
  - predicted probability $\widehat{p}_{ij} = \Pr[y_i = j] = e^{-\widehat{\mu}_i} \widehat{\mu}_i^j / j!$ for Poisson.
- Expect $\widehat{p}_j$ close to $\bar{p}_j$, $j = 0, 1, 2, ...$
- Informal statistic is Pearson's chi-square test

$$\sum_j \frac{(n\overline{p}_j - n\widehat{p}_j)^2}{n\widehat{p}_j}$$

but this is not χ^2 distributed due to estimation to get $\widehat{p}_j$.

- Instead do a formal chi-square goodness of fit test.
- Assuming that the density is correctly specified (so more applicable to models more general than Poisson) this can be computed as NR_u^2 (uncentered R^2) from the artificial regression

$$1 = \mathbf{s}_i(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}})'\gamma + \sum\nolimits_j (d_{ij}(y_i) - \widehat{p}_{ij})'\delta_j + error$$

- where
  - j denotes cells (e.g. values 0, 1, 2, 3, and 4 or more)
  - $d_{ij}(y_i)$ equals 1 if y_i is in cell j and 0 otherwise
  - $\widehat{p}_{ij}$ equals predicted probability for that cell
  - $\mathbf{s}_i(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \partial \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ $\quad (= (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))$ for Poisson).
- Reject at level α if $NR_u^2 > \chi^2_\alpha(J-1)$ where J is number of cells.
- Use Stata add-on chi2gof: e.g. `chi2gof, cells(11) table`

- Stata add-on `chi2gof, cells(11) table`
  - yields $\chi^2_\alpha(10)$ statistic equals 1103.43
  - and Poisson greatly underpredicts low counts e.g. for $y = 0$.

```
. chi2gof, cells(11) table

Chi-square Goodness-of-Fit Test for Poisson Model:

         Chi-square chi2(10) = 1103.43
         Prob>chi2           =    0.00
```

Cells	Abs. Freq.	Rel. Freq.	Fitted Rel. Freq.	Abs. Dif.
0	401	.1091	.0074	.1017
1	314	.0854	.0296	.0558
2	358	.0974	.063	.0344
3	334	.0908	.0954	.0045
4	339	.0922	.1159	.0237
5	266	.0723	.1212	.0489
6	231	.0628	.114	.0511
7	202	.0549	.0994	.0444
8	179	.0487	.0821	.0335
9	154	.0419	.0655	.0236
10 or more	791	.2445	.2067	.0378

5. Endogenous regressors: linear model review

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

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

- 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: $\mathrm{E}[u_i|\mathbf{z}_i] = 0$.
- Define $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix}$ and $\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_N' \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1' \\ \vdots \\ y_N' \end{bmatrix}$
- Just-identified case (# instruments = # endogenous)

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

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

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

- 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.

Several Interpretations of Linear IV/2SLS

- 1. Method of Moments
  - $\mathrm{E}[u_i|\mathbf{z}_i] = 0 \Rightarrow \mathrm{E}[\mathbf{z}_i u_i] = \mathbf{0} \Rightarrow \mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$.
  - IV solves corresponding sample moment condition $\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}$
  - And if overidentified do generalized method of moments (GMM).
- 2. Control Function
  - add predicted residual to control for endogeneity.
  - OLS with additional regressor the residual from first-stage OLS regression of y_2 on all exogenous regressors.
- 3. Two-Stage Least Squares
  - OLS with endogenous regressor y_2 replaced by its predicted value $\widehat{y}_2$ from first-stage OLS regression of y_2 on all exogenous regressors.
- Only methods 1 and 2 extend to nonlinear models such as Poisson.

Poisson endogenous method 1: nonlinear GMM

- Problem is
$$\mathsf{E}[(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))|\mathbf{x}_i] \neq \mathbf{0}.$$
- Assume existence of instruments $\mathbf{z}_i$ such that
$$\begin{aligned} & \mathsf{E}[(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))|\mathbf{z}_i] = \mathbf{0} \\ \Rightarrow \quad & \mathsf{E}[\mathbf{z}_i(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))] = \mathbf{0}. \end{aligned}$$
- Just-identified case: $\widehat{\boldsymbol{\beta}}_{\text{MM}}$ solves
$$\sum_{i=1}^n (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i = \mathbf{0}.$$
- Over-identified case $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$ 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).
- We use the simpler $\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i = \mathbf{0}$.
- Example: `docvis` with `private` (private insurance) endogenous
  - instruments are `income` and `ssiratio` (soc sec income / total income)
- The nonlinear GMM estimator was obtained in two ways:
  - first using Stata command GMM
  - second using own code written using Mata - an illustration of Mata.

- Nonlinear GMM (NL2SLS) using command gmm

```
. gmm (docvis - exp({xb:$xlist}+{b0})), instruments($zlist) onestep nolog

Final GMM criterion Q(b) =  .0495772

GMM estimation

Number of parameters =   8
Number of moments    =   9
Initial weight matrix: Unadjusted                 Number of obs   =      3,677

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  /xb_private |   .5920142   .3397345     1.74   0.081    -.0738533    1.257882
 /xb_medicaid |   .3186685   .1909951     1.67   0.095    -.0556751     .693012
      /xb_age |   .3323179   .0705348     4.71   0.000     .1940723    .4705634
     /xb_age2 |   -.002176   .0004643    -4.69   0.000     -.003086   -.001266
   /xb_educyr |   .0190887   .0092216     2.07   0.038     .0010147    .0371626
   /xb_actlim |   .2084978   .0433758     4.81   0.000     .1234828    .2935128
   /xb_totchr |    .241843   .0129869    18.62   0.000     .2163892    .2672968
          /b0 |  -11.86323   2.732711    -4.34   0.000    -17.21924   -6.507211
------------------------------------------------------------------------------
Instruments for equation 1: income ssiratio medicaid age age2 educyr actlim totchr _cons
```

- Nonlinear GMM (NL2SLS) using own code in Mata

```
. 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.
- Slightly different se's to gmm due to different degrees of freedom.

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.

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

```
. 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

------------------------------------------------------------------------------
             |               Robust
      docvis |      Coef.   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
------------------------------------------------------------------------------
```

- Should bootstrap to get correct s.e.'s
  - as `lpuhat` is a generated regressor.
  - Here little change in standard errors.

```
. * 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

Poisson endogenous method 3: fully structural approach

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

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

  - 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).

- Use gsem command.

```
. * Fully structural model of endogeneity
. gsem (docvis <- private $xlist2 L, poisson) ///
>    (private <- $xlist2 income ssiratio L), nolog

Generalized structural equation model           Number of obs     =      3,677

Response       : docvis
Family         : Poisson
Link           : log

Response       : private
Family         : Gaussian
Link           : identity

Log likelihood = -12796.769

 ( 1)  [docvis]L = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
docvis						
private	.6331611	.2211577	2.86	0.004	.1997001	1.066622
medicaid	.2675884	.1007469	2.66	0.008	.070128	.4650487
age	.3687153	.0671669	5.49	0.000	.2370706	.50036
age2	-.0023977	.0004442	-5.40	0.000	-.0032682	-.0015271
educyr	.0176344	.0073809	2.39	0.017	.0031681	.0321006
actlim	.1852541	.0383531	4.83	0.000	.1100835	.2604248
totchr	.3038513	.0130363	23.31	0.000	.2783006	.329402
L	1	(constrained)				
_cons	-13.71873	2.545692	-5.39	0.000	-18.7082	-8.729267

- Further output from gsem command.

private						
medicaid	-.3945375	.0216178	-18.25	0.000	-.4369076	-.3521673
age	-.0829381	.0293255	-2.83	0.005	-.1404151	-.0254611
age2	.0005247	.000195	2.69	0.007	.0001424	.0009069
educyr	.0213581	.0021595	9.89	0.000	.0171255	.0255907
actlim	-.0297923	.0176412	-1.69	0.091	-.0643686	.0047839
totchr	.0185451	.0058515	3.17	0.002	.0070763	.0300138
income	.0026301	.000402	6.54	0.000	.0018422	.0034179
ssiratio	-.0724733	.0211615	-3.42	0.001	-.1139492	-.0309975
L	-.1355801	.0580826	-2.33	0.020	-.2494199	-.0217404
_cons	3.528705	1.095672	3.22	0.001	1.381227	5.676184
var(L)	.6679536	.0465516			.5826712	.7657183
var(e.private)	.1850348	.0119217			.1630837	.2099404