Some Stata commands for endogeneity in nonlinear panel-data models

David M. Drukker

Director of Econometrics
Stata

2014 German Stata Users Group meeting
June 13, 2014

Two approaches to endogeneity in nonlinear models

- Nonlinear instrumental variables, and control functions
    - Blundell et al. (2013) Chesher and Rosen (2013), Newey (2013), Wooldridge (2010), and Cameron and Trivedi (2005)
    - Only impose conditional moment restrictions
- Maximum likelihood
    - Wooldridge (2010), Cameron and Trivedi (2005), Skrondal and Rabe-Hesketh (2004), Rabe-Hesketh et al. (2004), Heckman (1978), and Heckman (1979)
    - Impose restrictions on the entire conditional distributions; less robust

Specific Stata solutions

- Stata has many commands to estimate the parameters of specific models
  - `ivregress`, `ivpoisson`, `ivprobit`, and `ivtobit`
  - `heckman`, `heckprobit`, and `heckoprobit`
- Two Stata commands that offer more general solutions are `gsem` and `gmm`

A GSEM solution for endogeneity

- Generalized structural equations models (GSEM) encompass many nonlinear triangular systems with unobserved components
  - A GSEM is a triangular system of nonlinear or linear equations that share unobserved random components
  - The `gsem` command can estimate the model parameters
    - `gsem` is new in Stata 13
  - The unobserved components can model random effects
    - Including nested effects, hierarchical effects, and random-coefficients
  - The unobserved components can also model endogeneity
    - Include the same unobserved component in two or more equations
  - Set up and estimation by maximum likelihood
  - Random-effects estimators and correlated-random-effects estimators
  - See Rabe-Hesketh and Skrondal (2012), Skrondal and Rabe-Hesketh (2004), Rabe-Hesketh et al. (2004), and Rabe-Hesketh et al. (2005)

A GMM solution for endogeneity or missing data

- Stata's `gmm` command can be used to stack the moment conditions from multistep estimators
  - Many control-function estimators for the parameters of models with endogeneity are described as multistep estimators
  - Many inverse-probability-weighted estimators, regression adjustment estimators, and combinations thereof, for the population-averaged effects from samples with missing data are described as multistep estimators
  - Converting multistep estimators into one-step estimators produces a consistent estimator for the variance-covariance of the estimator (VCE); see Newey (1984) and Wooldridge (2010) among others
  - Setup and estimation by GMM: Only the specified moment restrictions apply

GSEM structure

- GSEM handles endogeneity by including common, unobserved components into the equations for different variables

For example

$$\begin{pmatrix}\eta\\ \epsilon\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}\begin{bmatrix}1 & 0\\ 0 & \sigma^2\end{bmatrix}\right)$$

$$\mathbf{E}[y_1|\mathbf{x}, y_2, \eta] = F(\mathbf{x}\boldsymbol{\beta} + y_2\alpha + \eta\delta)$$

$$y_2 = \mathbf{x}\boldsymbol{\beta} + \mathbf{w}\boldsymbol{\gamma} + \eta + \epsilon$$

where

- $F()$ is smooth, nonlinear function
- $\mathbf{x}$ are exogenous covariates
- η is the common, unobserved component that gives rise to the endogeneity
- $\mathbf{w}$ are "instruments"
- ϵ is an error term

Bivariate probit with endogenous variable

- Two binary dependent variables, *school* and *work* for young people (20-30)
- Each is a function of *age* and parental socio-economic score (*ses*)
  - *age* is exogenous
  - *ses* is endogenous
    - *ses* is affected by an unobserved component that also affects each of the binary variables.
    - We believe that parental education *ped* affects *ses* but neither *school* nor *work*

$$ses_i = \alpha_0 + \alpha_1 ped_i + \alpha_2 \eta_i + \epsilon_1$$

$$work_i = \left((\beta_0 + \beta_1 ses_i + \beta_2 age_i + \beta_3 \eta_i + \epsilon_2) > 0 \right)$$

$$school_i = \left((\gamma_0 + \gamma_1 ses_i + \gamma_2 age_i + \gamma_3 \eta_i + \epsilon_3) > 0 \right)$$

$$\begin{pmatrix} \eta_i \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

```
. gsem (work <- ses age L, probit)        ///
>      (school <- ses age L, probit)      ///
>      (ses <- ped L),                    ///
>      var(L@1) nolog
Generalized structural equation model           Number of obs    =       5000
Log likelihood = -14078.848
 ( 1)  [var(L)]_cons = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
work <-						
ses	-.2405712	.0968634	-2.48	0.013	-.4304199	-.0507224
age	.1923723	.0148124	12.99	0.000	.1633406	.221404
L	.9237883	.1901529	4.86	0.000	.5510954	1.296481
_cons	-4.297587	.3235578	-13.28	0.000	-4.931748	-3.663425
school <-						
ses	.3839591	.084104	4.57	0.000	.2191182	.5488
age	-.1968823	.0156442	-12.58	0.000	-.2275444	-.1662201
L	.9276381	.2028112	4.57	0.000	.5301355	1.325141
_cons	3.934125	.5295485	7.43	0.000	2.896229	4.972021
ses <-						
ped	.2083431	.0145523	14.32	0.000	.1798212	.2368651
L	.923848	.0911936	10.13	0.000	.7451118	1.102584
_cons	.8938526	.1422065	6.29	0.000	.615133	1.172572
var(L)	1	(constrained)				
var(e.ses)	1.088828	.1668318			.8063745	1.470217

Fixed effects versus correlated random effects

- In the econometric parlance of panel data, fixed effects are generally defined to be individual-specific, unobserved random components that depend on observed covariates in an unspecified way
- Fixed effects are removed from the estimator to avoid the incidental parameters problem, so analysis is conditional on the unobserved fixed effects
- There is still some discussion as to whether fixed effects are random or fixed, but the modern approach views them as random (Wooldridge, 2010, page 286)
- Correlated random effects are a parametric approach to the problem of fixed effects

  The dependence between individual-specific effects and the covariates is modeled out, leaving common unobserved components (Cameron and Trivedi, 2005, pages 719 and 786) (Wooldridge, 2010, page 286)

Fixed effects versus correlated random effects

- At the cost of more parametric assumptions, correlated-random-effect (CRE) models identify average partial effects and many more functional forms for nonlinear dependent variables

Fixed-effects logit

- Main "job" is either work or school for young people aged 20–30
    - Variable $work_{it}$ is coded 0 for school, 1 for work
- We have 5 observations on each individual
- Logit probabilities that $work_{it} = 1$ are functions of age_{it}, and parental socio-economic score ses_{it}, and an unobserved individual-level component
    - age_{it} is exogenous
    - ses_{it} is endogenous, it is related to the unobserved individual-level component η_i

$$\epsilon_{it} \sim \text{Logistic}(0, \pi^2/3)$$
$$work_{it} = (\beta_0 + ses_{it}\beta_1 + age_{it}\beta_2 + \eta_i + \epsilon_{it}) > 0$$

- Except for regularity conditions, and $\eta_i \perp \epsilon_{it}$ no assumption is made about the distribution of η_i
- The distribution of η_i may depend on ses_{it} in an unspecified fashion

Conditional maximum-likelihood estimation

- The standard econometric approach is to maximize the log-likelihood function conditional on the sum $\sum_{t=1}^{T} y_{it}$
    - Chamberlain (1980), Chamberlain (1984), Wooldridge (2010) and Cameron and Trivedi (2005)
- This conditional log-likelihood function does not depend on the unobseved η_i, it is transformed out
- The estimator obtained by maximizing this conditional log-likelihood function is consistent for the coefficients on the time-varing covariates and it is asymptotically normal

```
. xtlogit w ses age, fe
note: multiple positive outcomes within groups encountered.
note: 185 groups (925 obs) dropped because of all positive or
      all negative outcomes.
Iteration 0:   log likelihood = -1513.9791
Iteration 1:   log likelihood = -1444.5811
Iteration 2:   log likelihood = -1444.4195
Iteration 3:   log likelihood = -1444.4195
Conditional fixed-effects logistic regression   Number of obs      =      4075
Group variable: id                              Number of groups   =       815
                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5
                                                LR chi2(2)         =    295.99
Log likelihood  = -1444.4195                    Prob > chi2        =    0.0000
```

work	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ses	-.5825966	.0392365	-14.85	0.000	-.6594987	-.5056946
age	.083444	.011576	7.21	0.000	.0607555	.1061325

A GSEM CRE logit

- A GSEM CRE logit specifies a distribution for η_i and how it enters the model for the related covariates
  - This estimator is better termed, a correlated-random-effects (CRE) estimator
  - Inference is not conditional on unobserved fixed effects and average partial effects, after averaging out CRE, are identified
- For example,

$$\begin{aligned} work_{it} &= (\beta_0 + ses_{it}\beta_1 + age_{it}\beta_2 + \eta_i + \epsilon_{it}) > 0 \\ ses_{it} &= \alpha_0 + \alpha_1 ped_i + \eta_i\alpha_2 + \xi_{it} \\ \eta_i &\sim \mathcal{N}(0,1) \\ \epsilon_{it} &\sim \text{Logistic}(0, \pi^2/3) \\ \xi_{it} &\sim \mathcal{N}(0,\sigma^2) \\ & (\eta_i, \epsilon_{it}, \xi_{it}) \text{ mutually independent} \end{aligned}$$

```
. gsem (work <- ses age L[id]@1, logit)   ///
>      (ses <- ped L[id]), vsquish nolog
Generalized structural equation model             Number of obs    =       5000
Log likelihood = -11172.491
 ( 1)  [work]L[id] = 1
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
work <-      |
         ses |  -.5902971   .0385655   -15.31   0.000     -.665884   -.5147101
         age |   .0875979   .0104571     8.38   0.000     .0671024    .1080934
       L[id] |          1  (constrained)
       _cons |  -2.047273   .2705777    -7.57   0.000    -2.577595    -1.51695
-------------+----------------------------------------------------------------
ses <-       |
         ped |   .0813543   .0118188     6.88   0.000     .0581898    .1045188
       L[id] |    1.48718   .1062063    14.00   0.000      1.27902    1.695341
       _cons |   1.151305   .1245313     9.25   0.000     .9072278    1.395381
-------------+----------------------------------------------------------------
   var(L[id])|   1.043044   .1547474                      .7798608    1.395044
-------------+----------------------------------------------------------------
  var(e.ses) |   .9936687   .0221993                      .9510978    1.038145
------------------------------------------------------------------------------
```

A CRE logit with an endogenous variable

- Now suppose that ses_{it} is endogenous and we have an instrument
  - ses_{it} is affected by the unobserved, individual-level component η_i and another unobserved component ξ_{it} that also affects $work_{it}$
  - We believe that parental education ped_{it} affects ses_{it} but not $work_{it}$
  - Some would not define η_i to FE, but rather RE that are related to the observed covariates

$$
\begin{aligned}
work_{it} &= (\beta_0 + ses_{it}\beta_1 + age_{it}\beta_2 + \eta_i + \xi_{it}\beta_3 + \epsilon_{1it}) > 0 \\
ses_{it} &= \alpha_0 + ped_{it}\alpha_1 + \eta_i\alpha_2 + \xi_{it} + \epsilon_{2it} \\
\epsilon_{1it} &\sim \text{Logistic}(0, \pi^2/3) \\
\epsilon_{2it} &\sim \mathcal{N}(0, \sigma^2) \\
\eta_i &\sim Normal(0,1) \\
\xi_i &\sim Normal(0,1) \\
& (\epsilon_{1it}, \epsilon_{2it}, \eta_i, \xi_i) \text{ mutually independent}
\end{aligned}
$$

```
. gsem (work <- ses age L[id]@1 X, logit)          ///
>       (ses <- ped L[id] X@1), var(X@1)vsquish    ///
>       from(var(e.ses):_cons = 1) nolog
Generalized structural equation model              Number of obs    =       5000
Log likelihood =  -12851.37
 ( 1)  [work]L[id] = 1
 ( 2)  [ses]X = 1
 ( 3)  [var(X)]_cons = 1
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
work <-      |
         ses |   -.593026   .0496495   -11.94   0.000    -.6903373   -.4957148
         age |   .1019323   .0149949     6.80   0.000     .0725429    .1313217
       L[id] |          1  (constrained)
           X |   2.150414   .2074175    10.37   0.000     1.743883    2.556945
       _cons |   9.282667   .9335425     9.94   0.000     7.452957    11.11238
-------------+----------------------------------------------------------------
ses <-       |
         ped |   2.020729   .0168226   120.12   0.000     1.987757    2.053701
       L[id] |   1.515159   .1373711    11.03   0.000     1.245916    1.784401
           X |          1  (constrained)
       _cons |    .741761   .1704414     4.35   0.000     .4077019     1.07582
-------------+----------------------------------------------------------------
   var(L[id])|   .9920447   .1891004                      .6827755      1.4414
       var(X)|          1  (constrained)
-------------+----------------------------------------------------------------
   var(e.ses)|   1.066483   .0459968                      .9800357    1.160555
------------------------------------------------------------------------------
```

Panel probit with endogenous variable and CRE

- Binary dependent variables $school_{it}$ for young people (20-30, at first interview)
  - $school_{it}$ is a function of age_{it} and time-varying parental socio-economic score ses_{it}
  - age_{it} is exogenous
  - ses_{it} is endogenous
    - ses_{it} is affected by an unobserved component individual-level effect η_i and by a time-varying unobserved component ξ_{it}, both of which also affect $school_{it}$
    - We believe that time-varying parental education ped_{it} affects ses_{it} but not $school_{it}$.
- We have 5 observations on each young person

$$ses_{it} = \alpha_0 + \alpha_1 ped_{it} + \xi_{it} + \eta_i + \epsilon_{1,it}$$

$$school_{it} = \Big((\beta_0 + \beta_1 ses_{it} + \beta_2 age_{it} + \beta_3 \xi_{it} + \eta_i + \epsilon_{2,it}) > 0\Big)$$

$$\eta_i \sim Normal(0, \sigma_\eta) \qquad \epsilon_{1,it} \sim Normal(0, \sigma_{ses})$$

$$\xi_{it} \sim Normal(0, 1) \qquad \epsilon_{2,it} \sim Normal(0, 1)$$

```
. gsem (school <- ses age L M1[id]@1, probit)      ///
>       (ses <- ped L@1 M1[id]@1),                          ///
>       var(L@1) from(var(e.ses):_cons=1)  nolog
Generalized structural equation model             Number of obs    =        5000
Log likelihood = -10377.715
 ( 1)  [school]M1[id] = 1
 ( 2)  [ses]M1[id] = 1
 ( 3)  [ses]L = 1
 ( 4)  [var(L)]_cons = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
school <-						
ses	.6098294	.0447354	13.63	0.000	.5221496	.6975093
age	-.4142175	.0201581	-20.55	0.000	-.4537266	-.3747085
M1[id]	1	(constrained)				
L	1.123539	.1016453	11.05	0.000	.9243183	1.322761
_cons	10.69246	.5345878	20.00	0.000	9.644685	11.74023
ses <-						
ped	.5016687	.0150045	33.43	0.000	.4722603	.531077
M1[id]	1	(constrained)				
L	1	(constrained)				
_cons	.9645122	.1500038	6.43	0.000	.6705102	1.258514
var(M1[id])	1.042761	.0646625			.9234241	1.177521
var(L)	1	(constrained)				
var(e.ses)	.9568585	.0433915			.8754826	1.045798

Multinomial logit with endogenous variable

- Main "job" is either work, school, or home for young people aged 20–30
  - job_i is coded, 0 for home, 1 for work, and 2 for school
- Multinomial-logit probabilities are functions of age_i, and parental socio-economic score ses_i, and an unobserved individual-level component η_i
  - age_i is exogenous
  - ses_i is endogenous,
    - ses_i is affected by η_i that also affects the multinomial-logit probabilities
    - We believe that parental education ped_i affects ses_i but not the multinomial-logit probabilities

$$Pr[job = j] = \frac{exp(\beta_{0j} + ses_i\beta_{1j} + age_i\beta_{2j} + \eta_i\beta_{4j})}{1 + \sum_{j=1}^{2} exp(\beta_{0j} + ses_i\beta_{1j} + age_i\beta_{2j} + \eta_i\beta_{4j})} \qquad j \in \{1, 2\}$$

$$ses_i = \alpha_0 + \alpha_1 ped_i + \eta_i + \epsilon_i$$

$$\eta_i \sim \ Normal(0, 1) \qquad \epsilon_i \sim \ Normal(0, \sigma_{ses})$$

```
. gsem (job <- ses age L, mlogit) (ses <- ped L@1), var(L@1) nolog
Generalized structural equation model            Number of obs    =       3000
Log likelihood = -8130.9865
 ( 1)  [ses]L = 1
 ( 2)  [var(L)]_cons = 1
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0.job        |  (base outcome)
-------------+----------------------------------------------------------------
1.job <-     |
         ses |   .1680505    .079434     2.12   0.034     .0123627    .3237383
         age |   .1977622   .0176799    11.19   0.000     .1631103    .2324141
           L |   .4178895   .1825025     2.29   0.022     .0601912    .7755879
       _cons |  -5.667666   .5556052   -10.20   0.000    -6.756632   -4.578699
-------------+----------------------------------------------------------------
2.job <-     |
         ses |   .5734593   .0834707     6.87   0.000     .4098598    .7370588
         age |  -.2094759   .0201765   -10.38   0.000    -.2490211   -.1699306
           L |  -.6267227   .1836712    -3.41   0.001    -.9867115   -.2667338
       _cons |    1.21761   .6033821     2.02   0.044      .035003    2.400217
-------------+----------------------------------------------------------------
ses <-       |
         ped |   .6313673   .0197324    32.00   0.000     .5926925     .670042
           L |          1  (constrained)
       _cons |   .6768382   .1919967     3.53   0.000     .3005317    1.053145
-------------+----------------------------------------------------------------
      var(L)|          1  (constrained)
-------------+----------------------------------------------------------------
  var(e.ses)|   1.007182   .0518205                      .9105691    1.114046
------------------------------------------------------------------------------
```

Multinomial logit with CRE and an endogenous variable

- Main "job" is either work, school, or home for young people
  - job_{it} is coded, 0 for home, 1 for work, and 2 for school
- Multinomial-logit probabilities are functions of age_{it}, and parental socio-economic score ses_{it}, an unobserved individual-level component η_i, and an unobserved component that varies over individuals and time ξ_{it}
  - age_{it} is exogenous, ses_{it} is endogenous
    - ses_{it} is affected by η_i and by ξ_{it}, both of which also affect the multinomial-logit probabilities
    - We believe that parental education ped_{it} affects ses_{it} but not the multinomial-logit probabilities

$$xb_{itj} = \beta_{0j} + ses_{it}\beta_{1j} + age_{it}\beta_{2j} + \eta_i + \xi_{it}\beta_{4j}$$

$$Pr[job_{it} = j] = \frac{exp(xb_{ijt})}{1 + \sum_{j=1}^{2} exp(xb_{itj})} \quad j \in \{1, 2\}$$

$$ses_i = \alpha_0 + \alpha_1 ped_i + \eta_i + \xi_{it} + \epsilon_{it}$$

$$\eta_i \sim \ Normal(0, \sigma_\eta) \quad \xi_{it} \ \sim \ Normal(0, 1) \quad \epsilon_{it} \sim \ Normal(0, \sigma_{ses})$$

```
. gsem (job <- ses age L P1[id]@1, mlogit) (ses <- ped L@1 P1[id]@1),     ///
>         var(L@1)  vsquish  nolog
Generalized structural equation model             Number of obs    =       5000
Log likelihood = -13691.986
 ( 1)  [1.job]P1[id] = 1
 ( 2)  [2.job]P1[id] = 1
 ( 3)  [ses]P1[id] = 1
 ( 4)  [ses]L = 1
 ( 5)  [var(L)]_cons = 1
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0.job        |  (base outcome)
-------------+----------------------------------------------------------------
1.job <-     |
         ses |    .082676   .0381896     2.16   0.030     .0078257    .1575262
         age |   .2072062   .0150389    13.78   0.000     .1777304    .2366819
      P1[id] |          1  (constrained)
           L |   .6057244   .1070445     5.66   0.000      .395921    .8155277
       _cons |  -5.398094   .4560614   -11.84   0.000    -6.291958    -4.50423
-------------+----------------------------------------------------------------
2.job <-     |
         ses |   .4291914   .0422678    10.15   0.000      .346348    .5120348
         age |  -.1651801   .0164842   -10.02   0.000    -.1974885   -.1328717
      P1[id] |          1  (constrained)
           L |  -.2399792   .1115573    -2.15   0.031    -.4586274    -.021331
       _cons |   1.206197   .4645158     2.60   0.009     .2957623    2.116631
-------------+----------------------------------------------------------------
ses <-       |
         ped |   .8193806   .0206827    39.62   0.000     .7788433    .8599179
      P1[id] |          1  (constrained)
           L |          1  (constrained)
       _cons |   .7655727   .2146381     3.57   0.000     .3448897    1.186256
-------------+----------------------------------------------------------------
  var(P1[id])|   1.012727   .0616391                      .8988445    1.141039
       var(L)|          1  (constrained)
-------------+----------------------------------------------------------------
   var(e.ses)|   .9701532   .0435647                      .8884176    1.059409
------------------------------------------------------------------------------
```

A CRE probit with sample-selection

- Binary variable for school or work $sowork_{it}$ is missing if the young person is at home
- We believe that parental education ped_{it} and parental SES score ses_{it} affect the choice between school or work
- We believe that that ses_{it} and an attachment-to-home score ath_{it} affect whether the young person stays home, making $sowork_{it}$ missing.
- We allow for Heckman-type endogenous selection and CRE

$$sowork_{it} = \begin{cases} (\beta_0 + \beta_1 ses_{it} + \beta_2 ped_{it} + \beta_3 \xi_{it} + \eta_i + \epsilon_{1it} > 0), & \text{if } home_{it} = 0 \\ . & \text{otherwise} \end{cases}$$

$$home_{it} = (\gamma_0 + \gamma_1 ses_{it} + \gamma_2 ath_{it} + \xi_{it} + \eta_{it} + \epsilon_{2it} > 0)$$

$$ses_{it} = \alpha_0 + \eta_i + \epsilon_{3it} \qquad ped_{it} = \alpha_0 + \eta_i + \epsilon_{4it}$$

$$ath_{it} = \alpha_0 + \eta_i + \epsilon_{5it}$$

$$\eta_i \sim Normal(0,1) \quad \epsilon_{1it} \sim Normal(0,1) \quad \epsilon_{2it} \sim Normal(0,1)$$

$$\epsilon_{3it} \sim Normal(0,\sigma_3^2) \quad \epsilon_{4it} \sim Normal(0,\sigma_3^2) \quad \epsilon_{5it} \sim Normal(0,\sigma_5^2)$$

$$\xi_{it} \sim Normal(0,1)$$

```
. gsem (sowork <- ses ped L M[id]@1, probit)                    ///
>       (home   <- ses ath L@1 M[id]@1, probit)                 ///
>       (ses <- M[id]@1)                                         ///
>       (ped <- M[id]@1)                                         ///
>       (ath <- M[id]@1)                                         ///
>          , var(L@1) nologGeneralized structural equation model             Number of obs     =       7500
Log likelihood = -38532.664
 ( 1)  [sowork]M[id] = 1
 ( 2)  [home]M[id] = 1
 ( 3)  [home]L = 1
 ( 4)  [ses]M[id] = 1
 ( 5)  [ped]M[id] = 1
 ( 6)  [ath]M[id] = 1
 ( 7)  [var(L)]_cons = 1
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sowork <-    |
         ses |   .9927245   .0810946    12.24   0.000     .8337821    1.151667
         ped |   .9831526   .0816976    12.03   0.000     .8230283    1.143277
             |
       M[id] |          1  (constrained)
             |
           L |    1.06312   .1247585     8.52   0.000     .8185974    1.307642
       _cons |  -2.024637   .1560467   -12.97   0.000    -2.330483   -1.718791
-------------+----------------------------------------------------------------
home <-      |
         ses |   -.989918   .0236261   -41.90   0.000    -1.036224   -.9436117
         ath |   .9893967   .0292436    33.83   0.000     .9320802    1.046713
             |
       M[id] |          1  (constrained)
             |
           L |          1  (constrained)
       _cons |  -1.034227   .0484887   -21.33   0.000    -1.129263   -.9391909
-------------+----------------------------------------------------------------
ses <-       |
       M[id] |          1  (constrained)
             |
       _cons |   .9617187   .0288255    33.36   0.000     .9052217    1.018216
-------------+----------------------------------------------------------------
ped <-       |
```

More GSEM examples

- All the documentation in online.
  - `http://www.stata.com/support/documentation/`
- For an example of a cross-sectional Heckman model, see `http://www.stata.com/bookstore/structural-equation-modeling-reference-manual/` and click on **example43g**
- For an example of a cross-sectional endogenous treatment effects, see `http://www.stata.com/bookstore/structural-equation-modeling-reference-manual/` and click on **example44g**

Two-step estimators as GMM estimators

- Many two-step estimators have the form
  1. Estimate nuisance parameters γ by an M estimator
  2. Estimate parameters of interest β by an M estimator or a method of moments estimator that depends on the original data and $\widehat{\gamma}$
- In general, the distribution of $\widehat{\beta}$ depends on the first stage estimation
  - The correction is well known, e.g. Wooldridge (2010)
- Another way solving the two-step estimation problem is to stack the moment conditions from the two estimation problems and solve them jointly

Definition of GMM estimator

- Our research question implies q population moment conditions

$$E[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$$

  - $\mathbf{m}$ is $q \times 1$ vector of functions whose expected values are zero in the population
  - $\mathbf{w}_i$ is the data on person i
  - $\boldsymbol{\theta}$ is $k \times 1$ vector of parameters, $k \leq q$

- The sample moments that correspond to the population moments are

$$\overline{\mathbf{m}}(\boldsymbol{\theta}) = (1/N) \sum_{i=1}^{N} \mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})$$

- When $k < q$, the GMM choses the parameters that are as close as possible to solving the over-identified system of moment conditions

$$\widehat{\boldsymbol{\theta}}_{GMM} \equiv \arg\min_{\boldsymbol{\theta}} \quad \overline{\mathbf{m}}(\boldsymbol{\theta})' \mathbf{W} \overline{\mathbf{m}}(\boldsymbol{\theta})$$

Some properties of the GMM estimator

$$\widehat{\boldsymbol{\theta}}_{GMM} \equiv \text{arg min}_{\boldsymbol{\theta}} \quad \overline{\mathbf{m}}(\boldsymbol{\theta})'\mathbf{W}\overline{\mathbf{m}}(\boldsymbol{\theta})$$

- When $k = q$, the MM estimator solves $\overline{\mathbf{m}}(\boldsymbol{\theta})$ exactly so $\overline{\mathbf{m}}(\boldsymbol{\theta})'\mathbf{W}\overline{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0}$
- $\mathbf{W}$ only affects the efficiency of the GMM estimator
  - Setting $\mathbf{W} = \mathbf{I}$ yields consistent, but inefficient estimates
  - Setting $\mathbf{W} = \text{Cov}[\overline{\mathbf{m}}(\theta)]^{-1}$ yields an efficient GMM estimator
  - We can take multiple steps to get an efficient GMM estimator
    1. Let $\mathbf{W} = \mathbf{I}$ and get
    $$\widehat{\boldsymbol{\theta}}_{GMM1} \equiv \text{arg min}_{\boldsymbol{\theta}} \quad \overline{\mathbf{m}}(\boldsymbol{\theta})'\overline{\mathbf{m}}(\boldsymbol{\theta})$$
    2. Use $\widehat{\boldsymbol{\theta}}_{GMM1}$ to get $\widehat{\mathbf{W}}$, which is an estimate of $\text{Cov}[\overline{\mathbf{m}}(\theta)]^{-1}$
    3. Get
    $$\widehat{\boldsymbol{\theta}}_{GMM2} \equiv \text{arg min}_{\boldsymbol{\theta}} \quad \overline{\mathbf{m}}(\boldsymbol{\theta})'\widehat{\mathbf{W}}\overline{\mathbf{m}}(\boldsymbol{\theta})$$
    4. Repeat steps 2 and 3 using $\widehat{\boldsymbol{\theta}}_{GMM2}$ in place of $\widehat{\boldsymbol{\theta}}_{GMM1}$

The `gmm` command

- The command `gmm` estimates parameters by GMM
- `gmm` is similar to `nl`, you specify the sample moment conditions as substitutable expressions
- Substitutable expressions enclose the model parameters in braces $\{\}$

The syntax of gmm I

- For many models, the population moment conditions have the form
$$E[\mathbf{z}e(\beta)] = \mathbf{0}$$
where $\mathbf{z}$ is a $q \times 1$ vector of instrumental variables and $e(\beta)$ is a scalar function of the data and the parameters β
- The corresponding syntax of *gmm* is

gmm (*eb_expression*) [*if*] [*in*] [*weight*] ,
    instruments(*instrument_varlist*) [*options*]

where some options are

onestep	use one-step estimator (default is two-step estimator)
winitial(*wmtype*)	initial weight-matrix $\mathbf{W}$
wmatrix(*witype*)	weight-matrix $\mathbf{W}$ computation after first step
vce(*vcetype*)	*vcetype* may be robust, cluster, bootstrap, hac

Modeling crime data I

- We have data

```
. use cscrime, clear
. describe

Contains data from cscrime.dta
  obs:        10,000
 vars:             5                          24 May 2008 17:01
 size:       400,000                          (_dta has notes)
-------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-------------------------------------------------------------------------------
policepc        double  %10.0g               police officers per thousand
arrestp         double  %10.0g               arrests/crimes
convictp        double  %10.0g               convictions/arrests
legalwage       double  %10.0g               legal wage index 0-20 scale
crime           double  %10.0g               property-crime index 0-50 scale
-------------------------------------------------------------------------------
Sorted by:
```

Modeling crime data II

- We specify that

$$\texttt{crime}_i = \beta_0 + \texttt{policepc}_i\beta_1 + \texttt{legalwage}_i\beta_2 + \epsilon_i$$

- We want to model

$$E[\texttt{crime}|\texttt{policepc}, \texttt{legalwage}] = \beta_0 + \texttt{policepc}\beta_1 + \texttt{legalwage}\beta_2$$

- If $E[\epsilon|\texttt{policepc}, \texttt{legalwage}] = 0$, the population moment conditions

$$E\left[\begin{pmatrix}\texttt{policepc}\\ \texttt{legalwage}\end{pmatrix}\epsilon\right] = \begin{pmatrix}0\\ 0\end{pmatrix}$$

  hold

OLS by GMM I

```
. gmm (crime - policepc*{b1} - legalwage*{b2} - {b3}),          ///
>         instruments(policepc legalwage) nolog

Final GMM criterion Q(b) =  6.61e-32

GMM estimation

Number of parameters =   3
Number of moments    =   3
Initial weight matrix: Unadjusted                  Number of obs   =    10000
GMM weight matrix:     Robust

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         /b1 |  -.4203287   .0053645   -78.35   0.000    -.4308431   -.4098144
         /b2 |  -7.365905   .2411545   -30.54   0.000    -7.838559   -6.893251
         /b3 |   27.75419   .0311028   892.34   0.000     27.69323    27.81515
------------------------------------------------------------------------------
Instruments for equation 1: policepc legalwage _cons
```

OLS by GMM II

```
. regress crime policepc legalwage, robust

Linear regression                               Number of obs =     10000
                                                F(  2,  9997) =   4422.19
                                                Prob > F      =    0.0000
                                                R-squared     =    0.6092
                                                Root MSE      =    1.8032

------------------------------------------------------------------------------
             |               Robust
       crime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    policepc |  -.4203287   .0053653   -78.34   0.000    -.4308459   -.4098116
   legalwage |  -7.365905   .2411907   -30.54   0.000    -7.838688   -6.893123
       _cons |   27.75419   .0311075   892.20   0.000     27.69321    27.81517
------------------------------------------------------------------------------
```

OLS by GMM III

```
. generate cons = 1
. gmm (crime - {xb:police legalwage cons}),          ///
>         instruments(police legalwage ) nolog onestep
Final GMM criterion Q(b) =  1.84e-31
GMM estimation
Number of parameters =   3
Number of moments    =   3
Initial weight matrix: Unadjusted                     Number of obs   =    10000
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
/xb_policepc	-.4203287	.0053645	-78.35	0.000	-.4308431	-.4098144
/xb_legalw~e	-7.365905	.2411545	-30.54	0.000	-7.838559	-6.893251
/xb_cons	27.75419	.0311028	892.34	0.000	27.69323	27.81515

```
Instruments for equation 1: policepc legalwage _cons
```

IV and 2SLS

- For some variables, the assumption $E[\epsilon|x] = 0$ is too strong and we need to allow for $E[\epsilon|x] \neq 0$
- If we have q variables $\mathbf{z}$ for which $E[\epsilon|\mathbf{z}] = \mathbf{0}$ and the correlation between $\mathbf{z}$ and $\mathbf{x}$ is sufficiently strong, we can estimate $\boldsymbol{\beta}$ from the population moment conditions
$$E[\mathbf{z}(y - \mathbf{x}\boldsymbol{\beta})] = \mathbf{0}$$
- $\mathbf{z}$ are known as instrumental variables
- If the number of variables in $\mathbf{z}$ and $\mathbf{x}$ is the same $(q = k)$, solving the sample moment conditions yield the MM estimator known as the instrumental variables (IV) estimator
- If there are more variables in $\mathbf{z}$ than in $\mathbf{x}$ $(q > k)$ and we let $\mathbf{W} = \left(\sum_{i=1}^{N} \mathbf{z}_i'\mathbf{z}_i\right)^{-1}$ in our GMM estimator, we obtain the two-stage least-squares (2SLS) estimator

2SLS on crime data I

- The assumption that $E[\epsilon|\texttt{policepc}] = 0$ is false, if communities increase `policepc` in response to an increase in crime (an increase in ϵ_i)
- The variables `arrestp` and `convictp` are valid instruments, if they measure some components of communities' toughness-on crime that are unrelated to ϵ but are related to `policepc`
- We will continue to maintain that $E[\epsilon|\texttt{legalwage}] = 0$

2SLS by GMM I

```
. gmm (crime - {xb:police legalwage cons}),          ///
>        instruments(arrestp convictp legalwage ) nolog onestep
Final GMM criterion Q(b) =   .001454
GMM estimation
Number of parameters =   3
Number of moments    =   4
Initial weight matrix: Unadjusted                    Number of obs   =    10000
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
/xb_policepc	-1.002431	.0455469	-22.01	0.000	-1.091701	-.9131606
/xb_legalw~e	-1.281091	.5890977	-2.17	0.030	-2.435702	-.1264811
/xb_cons	30.0494	.1830541	164.16	0.000	29.69062	30.40818

Instruments for equation 1: arrestp convictp legalwage _cons

2SLS by GMM II

```
. ivregress 2sls crime legalwage (policepc = arrestp convictp) , robust

Instrumental variables (2SLS) regression          Number of obs =    10000
                                                  Wald chi2(2)  = 1891.83
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =       .
                                                  Root MSE      =   3.216

------------------------------------------------------------------------------
             |               Robust
       crime |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    policepc |  -1.002431   .0455469   -22.01   0.000    -1.091701   -.9131606
   legalwage |  -1.281091   .5890977    -2.17   0.030    -2.435702   -.1264811
       _cons |    30.0494   .1830541   164.16   0.000     29.69062    30.40818
------------------------------------------------------------------------------
Instrumented:  policepc
Instruments:   legalwage arrestp convictp
```

CF estimator for Poisson model endogenous variables

- Cross-sectional CF estimator for Poisson model endogenous variables
- See Wooldridge (2010), and `ivpoisson` documentation

$$
\begin{aligned}
y_i &= \exp(\beta_0 + x_i\beta_1 + \epsilon_i) \\
x_i &= \alpha_0 + z_i\alpha_1 + \xi_i \\
\epsilon_i &= \xi_i\rho + \eta_i
\end{aligned}
$$

- (η_i is independent of ξ and $E[\exp(\eta_i)] = 1$)
- Implied model

$$E[y_i|z, x, \xi_i] = \exp(\beta_0 + x_i\beta_1 + \xi_i\rho)$$

  So we could estimate β_1 if we knew ξ_i
- CF estimator
  1. Estimates α_0 and α_1 by OLS,
  2. Computes residuals $\widehat{\epsilon}_i$
  3. Plug $\widehat{\epsilon}_i$ in for ξ
  4. Now estimate β_1 by multiplicative moment condition as $E[\exp(\eta_i)] = 1$

GMM with evaluator programs

- Up to this point, all the problems have fit into the residual-instrument syntax
- We want to use `gmm` to estimator more difficult models
- We need to use the program-evaluator syntax

gmm program evaluator syntax

gmm *evaluator_program_name*, nequations(#)
    parameters(*parameter_name_list*) [*options*]

```
program define ivp_m
   version 13
   syntax varlist if, at(name)
   forvalues i=1/5{
    local m`i' : word `i' of `varlist'
   }
   quietly {
      tempvar r1 r2
      generate double `r2' = x - `at'[1,4]*z - `at'[1,5]
      generate double `r1' = y/exp(`at'[1,1]*x + `at'[1,2] +`at'[1,3]*`r2') - 1
      replace `m1' = `r2'
      replace `m2' = `r2'*z
      replace `m3' = `r1'
      replace `m4' = `r1'*x
      replace `m5' = `r1'*`r2'
  }
end
```

```
. gmm ivp_m , nequations(5) parameters(y:x y:_cons rho:_cons x:z x:_cons) winit
> ial(identity) onestep nolog
Final GMM criterion Q(b) =  4.05e-15
GMM estimation
Number of parameters =   5
Number of moments    =   5
Initial weight matrix: Identity                       Number of obs   =    5000
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
y						
x	1.037235	.062547	16.58	0.000	.914645	1.159825
_cons	.0112318	.0272029	0.41	0.680	-.0420849	.0645485
rho						
_cons	.0947202	.0657478	1.44	0.150	-.0341431	.2235835
x						
z	.3890606	.0137986	28.20	0.000	.3620159	.4161053
_cons	.1003455	.0144203	6.96	0.000	.0720821	.1286088

```
Instruments for equation 1: _cons
Instruments for equation 2: _cons
Instruments for equation 3: _cons
Instruments for equation 4: _cons
Instruments for equation 5: _cons
```

```
. ivpoisson cfunction y (x = z)
Step 1
Iteration 0:   GMM criterion Q(b) =  .01255627
Iteration 1:   GMM criterion Q(b) =  .00003538
Iteration 2:   GMM criterion Q(b) =  4.202e-10
Iteration 3:   GMM criterion Q(b) =  6.188e-20
Exponential mean model with endogenous regressors
Number of parameters =   5                          Number of obs   =    5000
Number of moments    =   5
Initial weight matrix: Unadjusted
GMM weight matrix:     Robust
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
           x |   1.037235    .062547    16.58   0.000     .9146451    1.159825
       _cons |   .0112319   .0272029     0.41   0.680    -.0420848    .0645486
-------------+----------------------------------------------------------------
x            |
           z |   .3890606   .0137986    28.20   0.000     .3620159    .4161053
       _cons |   .1003455   .0144203     6.96   0.000     .0720821    .1286088
-------------+----------------------------------------------------------------
        /c_x |   .0947201   .0657478     1.44   0.150    -.0341432    .2235834
------------------------------------------------------------------------------
Instrumented:  x
Instruments:   z
```

Fixed-effects Poisson estimator

- Wooldridge (1999, 2010); Blundell, Griffith, and Windmeijer (2002) discuss estimating the fixed-effects Poisson model for panel data by GMM.
- In the Poisson panel-data model we are modeling

$$E[y_{it}|\mathbf{x}_{it}, \eta_i] = \exp(\mathbf{x}_{it}\boldsymbol{\beta} + \eta_i)$$

- Hausman, Hall, and Griliches (1984) derived a conditional log-likelihood function when the outcome is assumed to come from a Poisson distribution with mean $\exp(\mathbf{x}_{it}\boldsymbol{\beta} + \eta_i)$ and η_i is an observed component that is correlated with the $\mathbf{x}_{it}$

- Wooldridge (1999) showed that you could estimate the parameters of this model by solving the sample moment equations

$$\sum_i \sum_t \mathbf{x}_{it} \left(y_{it} - \mu_{it} \frac{\overline{y}_i}{\overline{\mu}_i} \right) = \mathbf{0}$$

- These moment conditions do not fit into the interactive syntax because the term $\overline{\mu}_i$ depends on the parameters
- Need to use moment-evaluator program syntax

```
program xtfe
    version 13
    syntax varlist if, at(name)
    quietly {
        tempvar mu mubar ybar
        generate double `mu' = exp(kids*`at'[1,1]   ///
            + cvalue*`at'[1,2]                      ///
            + tickets*`at'[1,3]) `if'
        egen double `mubar'  = mean(`mu') `if', by(id)
        egen double `ybar'   = mean(accidents) `if', by(id)
        replace `varlist'    = accidents            ///
                             - `mu'*`ybar'/`mubar' `if'
    }
end
```

FE Poisson by gmm

```
. use xtaccidents, clear
. by id: egen max_a = max(accidents )
. drop if max_a ==0
(3750 observations deleted)
. gmm xtfe , equations(accidents) parameters(kids cvalue tickets)    ///
>         instruments(kids cvalue tickets, noconstant)               ///
>         vce(cluster id) onestep nolog
Final GMM criterion Q(b) =  1.50e-16
GMM estimation
Number of parameters =   3
Number of moments    =   3
Initial weight matrix: Unadjusted                    Number of obs  =     1250
                                  (Std. Err. adjusted for 250 clusters in id)
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
/kids	-.4506245	.0969133	-4.65	0.000	-.6405711	-.2606779
/cvalue	-.5079946	.0615506	-8.25	0.000	-.6286315	-.3873577
/tickets	.151354	.0873677	1.73	0.083	-.0198835	.3225914

```
Instruments for equation 1: kids cvalue tickets
```

FE Poisson by xtpoisson, fe

```
. xtpoisson accidents kids cvalue tickets, fe nolog vce(robust)
Conditional fixed-effects Poisson regression    Number of obs      =      1250
Group variable: id                              Number of groups   =       250
                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5
                                                Wald chi2(3)       =     84.89
Log pseudolikelihood  = -351.11739              Prob > chi2        =    0.0000
                                     (Std. Err. adjusted for clustering on id)
```

accidents	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
kids	-.4506245	.0969133	-4.65	0.000	-.6405712	-.2606779
cvalue	-.5079949	.0615506	-8.25	0.000	-.6286319	-.3873579
tickets	.151354	.0873677	1.73	0.083	-.0198835	.3225914

Blundell, Richard, Rachel Griffith, and Frank Windmeijer. 2002. “Individual effects and dynamics in count data models,” *Journal of Econometrics*, 108, 113–131.

Blundell, Richard, Dennis Kristensen, and Rosa L Matzkin. 2013. “Control Functions and Simultaneous Equations Methods,” *American Economic Review*, 103(3), 563–569.

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and applications*, Cambridge: Cambridge University Press.

Chamberlain, Gary. 1980. “Analysis of Covariance with Qualitative Data,” *Review of Economic Studies*, 47, 225–238.

———. 1984. “Panel Data,” in Zvi Grilliches and Micheal D. Intrilligaor (eds.), *Handbook of Econometrics*, vol. II, Amsterdam: Elsevier, pp. 1247–1318.

Chesher, Andrew and Adam M. Rosen. 2013. “What do instrumental variable models deliver with discrete dependent variables?” *American Economic Review*, 103(3), 557–562.

Hausman, Jerry A., Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric models for count data with an application to the patents–R & D relationship," *Econometrica*, 52(4), 909–938.

Heckman, James J. 1978. "Dummy exogenous variables in a simulation equation system," *Econometrica*, 46(2), 403–426.

———. 1979. "Sample selection bias as a specification error," *Econometrica*, 153–161.

Newey, Whitney K. 1984. "A method of moments interpretation of sequential estimators," *Economics Letters*, 14(2), 201–206.

———. 2013. "Nonparametric instrumental variables estimation," *American Economic Review*, 103(3), 550–556.

Rabe-Hesketh, Sophia and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata, Volume II: Categorical Responses, Counts, and Survival*, College Station, Tx: Stata Press, 3d ed.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles. 2004. "Generalized multilevel structural equation modeling," *Psychometrika*, 69(2), 167–190.

———. 2005. "Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects," *Journal of Econometrics*, 128(2), 301–323.

Skrondal, Anders and Sophia Rabe-Hesketh. 2004. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, Boca Raton, Florida: Chapman and Hall/CRC.

Wooldridge, Jeffrey M. 1999. "Distribution-free estimation of some nonlinear panel-data models," *Journal of Econometrics*, 90, 77–90.

———. 2010. *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Massachusetts: MIT Press, second ed.