Panel data methods for microeconometrics using Stata

A. Colin Cameron
Univ. of California - Davis

Based on A. Colin Cameron and Pravin K. Trivedi,
Microeconometrics using Stata, Stata Press, forthcoming.

April 8, 2008

1. Introduction

Panel data are repeated measures on individuals (i) over time (t).

Regress y_{it} on $\mathbf{x}_{it}$ for $i = 1, ..., N$ and $t = 1, ..., T$.

Complications compared to cross-section data:

1. **Inference:** correct (inflate) standard errors.
   This is because each additional year of data is not independent of previous years.
2. **Modelling:** richer models and estimation methods are possible with repeated measures.
   Fixed effects and dynamic models are examples.
3. **Methodology:** different areas of applied statistics may apply different methods to the same panel data set.

This talk: **overview** of panel data methods and `xt` commands for **Stata 10** most commonly used by **microeconometricians**.

Three **specializations** to general panel methods:

1. **Short panel:** data on many individual units and few time periods. Then data viewed as clustered on the individual unit. Many panel methods also apply to clustered data such as cross-section individual-level surveys clustered at the village level.
2. **Causation from observational data**: use repeated measures to estimate key marginal effects that are causative rather than mere correlation. Fixed effects: assume time-invariant individual-specific effects. IV: use data from other periods as instruments.
3. **Dynamic models:** regressors include lagged dependent variables.

Outline

2.1 Example: wages

- PSID wage data 1976-82 on 595 individuals. Balanced.
- Source: Baltagi and Khanti-Akom (1990).
  [Corrected version of Cornwell and Rupert (1998).]
- Goal: estimate causative effect of education on wages.
- Complication: education is time-invariant in these data.
  Rules out fixed effects.
  Need to use IV methods (Hausman-Taylor).

2.2 Reading in panel data

- Data organization may be
  - long form: each observation is an individual-time (i, t) pair
  - wide form: each observation is data on i for all time periods
  - wide form: each observation is data on t for all individuals
- `xt` commands require data in long form
  - use `reshape long` command to convert from wide to long form.
- Data here are already in long form

```
.  * Read in data set
.  use mus08psidextract.dta, clear
(PSID wage data 1976-82 from Baltagi and Khanti-Akom
(1990))
```

2.3 Summarize data using usual commands

```
. * Describe dataset
. describe

Contains data from mus08psidextract.dta
  obs:         4,165                          PSID wage data 1976-82 from Baltagi ar
 vars:            15                          16 Aug 2007 16:29
 size:       283,220 (97.5% of memory free)   (_dta has notes)
-------------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------
exp             float  %9.0g                  years of full-time work experience
wks             float  %9.0g                  weeks worked
occ             float  %9.0g                  occupation; occ==1 if in a blue-collar
ind             float  %9.0g                  industry; ind==1 if working in a manuf
south           float  %9.0g                  residence; south==1 if in the South ar
smsa            float  %9.0g                  smsa==1 if in the Standard metropolita
ms              float  %9.0g                  marital status
fem             float  %9.0g                  female or male
union           float  %9.0g                  if wage set be a union contract
ed              float  %9.0g                  years of education
blk             float  %9.0g                  black
lwage           float  %9.0g                  log wage
id              float  %9.0g
t               float  %9.0g
exp2            float  %9.0g
-------------------------------------------------------------------------------------
```

```
. * Summarize dataset
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
exp	4165	19.85378	10.96637	1	51
wks	4165	46.81152	5.129098	5	52
occ	4165	.5111645	.4999354	0	1
ind	4165	.3954382	.4890033	0	1
south	4165	.2902761	.4539442	0	1
smsa	4165	.6537815	.475821	0	1
ms	4165	.8144058	.3888256	0	1
fem	4165	.112605	.3161473	0	1
union	4165	.3639856	.4812023	0	1
ed	4165	12.84538	2.787995	4	17
blk	4165	.0722689	.2589637	0	1
lwage	4165	6.676346	.4615122	4.60517	8.537
id	4165	298	171.7821	1	595
t	4165	4	2.00024	1	7
exp2	4165	514.405	496.9962	1	2601

Balanced and complete as $7 \times 595 = 4165$.

```
. * Organization of data set
. list id t exp wks occ in 1/3, clean

       id   t   exp   wks   occ
  1.    1   1     3    32     0
  2.    1   2     4    43     0
  3.    1   3     5    40     0
```

Data are sorted by `id` and then by `t`

2.3 Summarize data using xt commands

- `xtset` command defines i and t.
- Allows use of panel commands and some time series operators

```
.  * Declare individual identifier and time identifier
.  xtset id t
 panel variable:  id (strongly balanced)
 time variable:  t, 1 to 7
 delta:  1 unit
```

```
. * Panel description of data set
. xtdescribe

      id:  1, 2, ..., 595                                    n =        595
       t:  1, 2, ..., 7                                      T =          7
           Delta(t) = 1 unit
           Span(t)  = 7 periods
           (id*t uniquely identifies each observation)

Distribution of T_i:   min      5%      25%       50%       75%     95%     max
                         7       7        7         7         7       7       7

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      595    100.00  100.00 |  1111111
 ---------------------------+---------
      595    100.00         |  XXXXXXX
```

Data are balanced with every individual *i* having 7 time periods of data.

```
. * Panel summary statistics: within and between variation
. xtsum lwage exp ed t

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
lwage    overall |  6.676346   .4615122    4.60517      8.537 |     N =    4165
         between |             .3942387     5.3364   7.813596 |     n =     595
         within  |             .2404023   4.781808   8.621092 |     T =       7
                 |                                            |
exp      overall |  19.85378   10.96637          1         51 |     N =    4165
         between |             10.79018          4         48 |     n =     595
         within  |              2.00024   16.85378   22.85378 |     T =       7
                 |                                            |
ed       overall |  12.84538   2.787995          4         17 |     N =    4165
         between |             2.790006          4         17 |     n =     595
         within  |                    0   12.84538   12.84538 |     T =       7
                 |                                            |
t        overall |         4    2.00024          1          7 |     N =    4165
         between |                    0          4          4 |     n =     595
         within  |              2.00024          1          7 |     T =       7
```

For time-invariant variable **ed** the within variation is zero.

For individual-invariant variable **t** the between variation is zero.

For `lwage` the within variation $<$ between variation.

```
. * Panel tabulation for a variable
. xttab south
```

south	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	2956	70.97	428	71.93	98.66
1	1209	29.03	182	30.59	94.90
Total	4165	100.00	610	102.52	97.54
			(n = 595)		

29.03% on average were in the south.

20.59% were ever in the south.

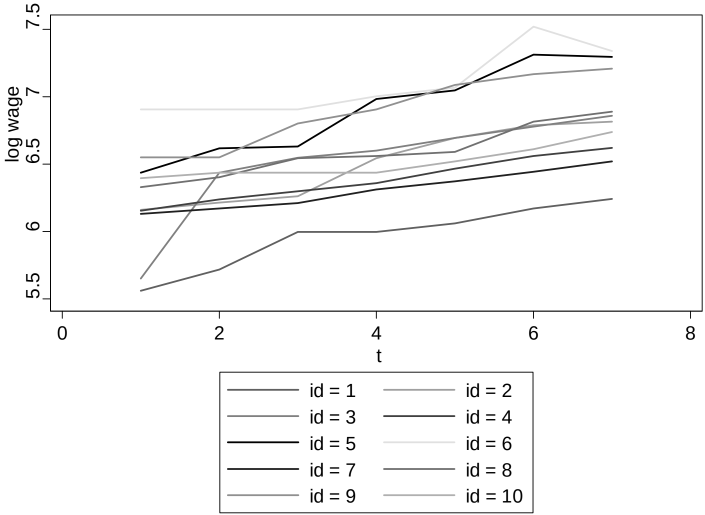
94.9% of those ever in the south were always in the south.

```
. * Transition probabilities for a variable
. xttrans south, freq
```

residence; south==1 if in the South area	residence; south==1 if in the South area 0	1	Total
0	2,527 99.68	8 0.32	2,535 100.00
1	8 0.77	1,027 99.23	1,035 100.00
Total	2,535 71.01	1,035 28.99	3,570 100.00

For the 28.99% of the sample ever in the south, 99.23% remained in the south the next period.

```
. * Time series plots of log wage for first 10 individuals
. xtline lwage if id<=10, overlay
```



```
. * First-order autocorrelation in a variable
. sort id t

. correlate lwage L.lwage L2.lwage L3.lwage L4.lwage L5.lwage L6.lwage
(obs=595)

             |                L.       L2.       L3.       L4.       L5.       L6.
             |    lwage     lwage     lwage     lwage     lwage     lwage     lwage
-------------+----------------------------------------------------------------------
       lwage |
         --. |   1.0000
         L1. |   0.9238    1.0000
         L2. |   0.9083    0.9271    1.0000
         L3. |   0.8753    0.8843    0.9067    1.0000
         L4. |   0.8471    0.8551    0.8833    0.8990    1.0000
         L5. |   0.8261    0.8347    0.8721    0.8641    0.8667    1.0000
         L6. |   0.8033    0.8163    0.8518    0.8465    0.8594    0.9418    1.0000
```

High serial correlation: $\mathrm{Cor}[y_t, y_{t-6}] = 0.80$.
Can also estimate correlations without imposing stationarity.

- Commands **describe, summarize and tabulate** confound cross-section and time series variation.
- Instead use **specialized panel commands** after `xtset`:
    - `xtdescribe`: extent to which panel is unbalanced
    - `xtsum`: separate within (over time) and between (over individuals) variation
    - `xttab`: tabulations within and between for discrete data e.g. binary
    - `xttrans`: transition frequencies for discrete data
    - `xtline`: time series plot for each individual on one chart
    - `xtdata`: scatterplots for within and between variation.

2.4 Pooled OLS

- Do regular OLS of y_{it} on $\mathbf{x}_{it}$.

```
. * Pooled OLS with incorrect default standard errors
. regress lwage exp exp2 wks ed, noheader
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	.044675	.0023929	18.67	0.000	.0399838	.0493663
exp2	-.0007156	.0000528	-13.56	0.000	-.0008191	-.0006121
wks	.005827	.0011827	4.93	0.000	.0035084	.0081456
ed	.0760407	.0022266	34.15	0.000	.0716754	.080406
_cons	4.907961	.0673297	72.89	0.000	4.775959	5.039963

The default standard errors erroneously assume errors are independent over i for given t.
Assumes more information content from data then is the case.

```
. * Pooled OLS with cluster-robust standard errors
. regress lwage exp exp2 wks ed, noheader vce(cluster id)
                                 (Std. Err. adjusted for 595 clusters in id)
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	.044675	.0054385	8.21	0.000	.0339941	.055356
exp2	-.0007156	.0001285	-5.57	0.000	-.0009679	-.0004633
wks	.005827	.0019284	3.02	0.003	.0020396	.0096144
ed	.0760407	.0052122	14.59	0.000	.0658042	.0862772
_cons	4.907961	.1399887	35.06	0.000	4.633028	5.182894

Cluster-robust standard errors are twice as large as default.

Cluster-robust t-statistics are half as large as default.

Typical result. Need to use cluster-robust se's if use pooled OLS.

3.1 Some basic considerations

1. **Regular time intervals** assumed.
2. **Unbalanced** panel okay (xt commands handle unbalanced data). [Should then rule out selection/attrition bias].
3. **Short panel** assumed, with T small and $N \rightarrow \infty$. [Versus long panels, with $T \rightarrow \infty$ and N small or $N \rightarrow \infty$.]
4. **Errors are correlated.** [For short panel: correlated over t for given i, but not over i.]
5. **Parameters** may vary over individuals or time. **Intercept:** Individual-specific effects model (fixed or random effects). **Slopes:** Pooling and random coefficients models.
6. **Regressors**: time-invariant, individual-invariant, or vary over both.
7. **Prediction**: ignored. [Not always possible even if marginal effects computed.]
8. **Dynamic models**: possible. [Usually static models are estimated.]

3.2 Basic linear panel models

- **Pooled model (or population-averaged)**

$$y_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it}. \tag{1}$$

- **Two-way effects model** allows intercept to vary over i and t

$$y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}. \tag{2}$$

- **Individual-specific effects model**

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \tag{3}$$

  where α_i may be fixed effect or random effect.

- **Mixed model or random coefficients model** allows slopes to vary over i

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}_i + \varepsilon_{it}. \tag{4}$$

3.3 Fixed effects versus random effects

- **Individual-specific effects model:**

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (\alpha_i + \varepsilon_{it}).$$

- **Fixed effects (FE):**
  - α_i is a random variable possibly correlated with $\mathbf{x}_{it}$
  - so regressor $\mathbf{x}_{it}$ may be **endogenous** (wrt to α_i but not ε_{it}) e.g. education is correlated with time-invariant ability
  - pooled OLS, pooled GLS, RE are inconsistent for $\boldsymbol{\beta}$
  - within (FE) and first difference estimators are consistent.
- **Random effects (RE) or population-averaged (PA):**
  - α_i is purely random (usually iid $(0, \sigma_\alpha^2)$) unrelated to $\mathbf{x}_{it}$
  - so regressor $\mathbf{x}_{it}$ is exogenous
  - all estimators are consistent for $\boldsymbol{\beta}$
- **Fundamental divide:** microeconometricians FE versus others RE.

3.4 Cluster-robust inference

- Many methods assume ε_{it} and α_i (if present) are iid.
- Yields **wrong standard errors** if heteroskedasticity or if errors not equicorrelated over time for a given individual.
- For short panel can relax and use **cluster-robust inference**.
  - Allows heteroskedasticity and general correlation over time for given i.
  - Independence over i is still assumed.
- For `xtreg` use option `vce(robust)` does cluster-robust
- For some other `xt` commands use option `vce(cluster)`
- And for some other `xt` commands there is no option but may be able to do a cluster bootstrap.

4.1 Pooled GLS estimator: xtgee

- Regress y_{it} on $\mathbf{x}_{it}$ using feasible GLS as error is not iid.

```
.  * Pooled FGLS estimator with AR(2) error &
cluster-robust se's
.  xtgee lwage exp exp2 wks ed, corr(ar 2) vce(robust)
```

4.2 Between GLS estimator: xtreg, be

- OLS of $\bar{y}_i$ on $\bar{\mathbf{x}}_i$. i.e. Regression using each individual's averages.

```
. * Between estimator with default standard errors
. xtreg lwage exp exp2 wks ed, be
```

4.3 Random effects estimator: xtreg, re

- FGLS in RE model assuming $\alpha_i$iid $(0, \sigma_\alpha^2)$ and ε_i iid $(0, \sigma_\varepsilon^2)$.
- Equals OLS of $(y_{it} - \widehat{\theta}_i \bar{y}_i)$ on $(\mathbf{x}_{it} - \widehat{\theta}_i \overline{\mathbf{x}}_i)$;
  $\theta_i = 1 - \sqrt{\sigma_\varepsilon^2 / (T_i \sigma_\alpha^2 + \sigma_\varepsilon^2)}$.

```
.  * Random effects estimator with cluster-robust se's
.  xtreg lwage exp exp2 wks ed, re vce(robust) theta
```

This gives $\widehat{\theta} = 0.82$.

4.4 Fixed effects (or within) estimator: xtreg, fe

- OLS regress $(y_{it} - \bar{y}_i)$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$.
- Mean-differencing eliminates α_i in $y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$

```
.   * Within or FE estimator with cluster-robust se's
.   xtreg lwage exp exp2 wks ed, fe vce(robust)
```

4.5 First differences estimator: regress with differences

- OLS regress $(y_{it} - y_{i,t-1})$ on $(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})$.
- First-differencing eliminates α_i in $y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$.

```
. * First difference estimator with cluster-robust se's
. regress D.(lwage $xlist), vce(cluster id)
```

4.6 Estimator comparison

```
. * Compare various estimators (with cluster-robust se's)
. global xlist exp exp2 wks ed
. quietly regress lwage $xlist, vce(cluster id)
. estimates store OLS
. quietly xtgee lwage exp exp2 wks ed, corr(ar 2)
vce(robust)
. estimates store PFGLS
. quietly xtreg lwage $xlist, be
. estimates store BE
. quietly xtreg lwage $xlist, re vce(robust)
. estimates store RE
. quietly xtreg lwage $xlist, fe vce(robust)
. estimates store FE
. estimates table OLS PFGLS BE RE FE, b(%9.4f) se stats(N)
```

Variable	OLS	PFGLS	BE	RE	FE
exp	0.0447	0.0719	0.0382	0.0889	0.1138
	0.0054	0.0040	0.0057	0.0029	0.0040
exp2	-0.0007	-0.0009	-0.0006	-0.0008	-0.0004
	0.0001	0.0001	0.0001	0.0001	0.0001
wks	0.0058	0.0003	0.0131	0.0010	0.0008
	0.0019	0.0011	0.0041	0.0009	0.0009
ed	0.0760	0.0905	0.0738	0.1117	0.0000
	0.0052	0.0060	0.0049	0.0063	0.0000
_cons	4.9080	4.5264	4.6830	3.8294	4.5964
	0.1400	0.1057	0.2101	0.1039	0.0601
N	4165.0000	4165.0000	4165.0000	4165.0000	4165.0000

legend: b/se

- Coefficients vary considerably across OLS, RE, FE and RE estimators.
- FE and RE similar as $\widehat{\theta} = 0.82 \simeq 1$.
- Not shown is that even for FE and RE cluster-robust changes se's.
- Coefficient of `ed` not identified for FE as time-invariant regressor!

4.7 Fixed effects versus random effects

- Prefer RE as can estimate all parameters and more efficient.
- But RE is inconsistent if fixed effects present.
- Use **Hausman test** to discriminate between FE and RE.
  - This tests difference between FE and RE estimates is statistically significantly different from zero.
- Problem: `hausman` command gives incorrect statistic as it assumes RE estimator is fully efficient, usually not the case.
- Solution: do a panel bootstrap of the Hausman test
  or use the Wooldridge (2002) robust version of Hausman test.

4.8 Stata linear panel commands

Panel summary	`xtset; xtdescribe; xtsum; xtdata; xtline; xttab; xttran`
Pooled OLS	`regress`
Feasible GLS	`xtgee, family(gaussian)` `xtgls; xtpcse`
Random effects	`xtreg, re; xtregar, re`
Fixed effects	`xtreg, fe; xtregar, fe`
Random slopes	`xtmixed; quadchk; xtrc`
First differences	`regress` (with differenced data)
Static IV	`xtivreg; xthtaylor`
Dynamic IV	`xtabond; xtdpdsys; xtdpd`

5.1 Long panels

- For **short panels** asymptotics are T fixed and $N \rightarrow \infty$.
- For **long panels** asymptotics are for $T \rightarrow \infty$
  - A dynamic model for the errors is specified, such as AR(1) error
  - Errors may be correlated over individuals
  - Individual-specific effects can be just individual dummies
  - Furthermore if N is small and T large can allow slopes to differ across individuals and test for poolability.

5.2 Commands for long panels

- Models with **stationary errors:**
  - `xtgls` allows several different models for the error
  - `xtpcse` is a variation of `xtgls`
  - `xtregar` does FE and RE with AR(1) error
  - Add-on `xtscc` gives HAC se's with spatial correlation.
- Models with **nonstationary errors** (currently active area):
  - As yet no Stata commands
  - Add-on `levinlin` does Levin-Lin-Chu (2002) panel unit root test
  - Add-on `ipshin` does Im-Pesaran-Shin (1997) panel unit root test in heterogeneous panels
  - Add-on `xtpmg` for does Pesaran-Smith and Pesaran-Shin-Smith estimation for nonstationary heterogeneous panels with both N and T large.

6.1 Panel IV: xtivreg

- Command `xtivreg` is natural extension of `ivregress` to panels.
- Consider model with possibly transformed variables:

$$y_{it}^* = \alpha + \mathbf{x}_{it}^{*\prime}\boldsymbol{\beta} + u_{it},$$

where transformations are

OLS	$y_{it}^* = y_{it}$
Between	$y_{it}^* = \bar{y}_i$
Fixed effects	$y_{it}^* = (y_{it} - \bar{y}_i)$
Random effects	$y_{it}^* = (y_{it} - \theta_i \bar{y}_i)$

- OLS is **inconsistent** if $\mathrm{E}[u_{it}|\mathbf{x}_{it}^*] = 0$.
- **IV estimation** with **instruments** $\mathbf{z}_{it}^*$ satisfy $\mathrm{E}[u_{it}|\mathbf{z}_{it}^*] = 0$.
- Example: `xtivreg lwage exp exp2 (wks = ms), fe`

6.2 Hausman-Taylor IV estimator: xthtaylor

- Command `xthtaylor` uses exogenous time-varying regressors $\mathbf{x}_{it}$ from periods other than the current as instruments.
- This enables estimation of coefficient of a time-invariant regressor in a fixed effects model (not possible using FE estimator).
- Example: allows estimation of coefficient of time-invariant regressor ed

```
xthtaylor lwage occ south smsa ind exp exp2 wks ms union
fem blk ed, ///
endog(exp exp2 wks ms union ed)
```

7.1 Linear dynamic panel models

- Simple dynamic model regresses y_{it} in **polynomial in time**.
  - e.g. Growth curve of child height or IQ as grow older
  - use previous models with $\mathbf{x}_{it}$ polynomial in time or age.
- Richer dynamic model regresses y_{it} on **lags** of y_{it}.

7.2 Linear dynamic panel models with individual effects

- **Leading example**: AR(1) model with individual specific effects

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}.$$

- Four reasons for y_{it} being serially correlated over time:
  - **True state dependence:** via $y_{i,t-1}$
  - **Observed heterogeneity:** via $\mathbf{x}_{it}$ which may be serially correlated
  - **Unobserved heterogeneity:** via α_i
  - Error correlation: via ε_{it}
- Focus on case where α_i is a **fixed effect**
  - FE estimator is now inconsistent (if short panel)
  - Instead use Arellano-Bond estimator

7.3 Arellano-Bond estimator

- **First-difference** to eliminate α_i (rather than mean-difference)

$$(y_{it} - y_{i,t-1}) = \gamma(y_{i,t-1} - y_{i,t-2}) + (\mathbf{x}_{it} - \mathbf{x}'_{i,t-1})\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}).$$

- **OLS inconsistent** as $(y_{i,t-1} - y_{i,t-2})$ correlated with $(\varepsilon_{it} - \varepsilon_{i,t-1})$ (even under assumption ε_{it} is serially uncorrelated).
- But $y_{i,t-2}$ is not correlated with $(\varepsilon_{it} - \varepsilon_{i,t-1})$, so can use $y_{i,t-2}$ as an **instrument** for $(y_{i,t-1} - y_{i,t-2})$.
- Arellano-Bond is a variation that uses **unbalanced set** of instruments with further lags as instruments.
  For $t = 3$ can use y_{i1}, for $t = 4$ can use y_{i1} and y_{i2}, and so on.
- Stata commands
  - `xtabond` for Arellano-Bond
  - `xtdpdsys` for Blundell-Bond (more efficient than `xtabond`)
  - `xtdpd` for more complicated models than `xtabond` and `xtdpdsys`.

```
.  * Optimal or two-step GMM for a dynamic panel model
.  xtabond lwage occ south smsa ind, lags(2) maxldep(3) ///
.  pre(wks,lag(1,2)) endogenous(ms,lag(0,2)) ///
.  endogenous(union,lag(0,2)) twostep vce(robust)
artests(3)
.  * Test whether error is serially correlated
.  estat abond
.  * Test of overidentifying restrictions
.  estat sargan
.  * Arellano/Bover or Blundell/Bond for a dynamic panel
model
.  xtdpdsys lwage occ south smsa ind, lags(2) maxldep(3)
///
.  pre(wks,lag(1,2)) endogenous(ms,lag(0,2)) ///
.  endogenous(union,lag(0,2)) twostep vce(robust)
artests(3)
```

8.1 Random coefficients model: xtrc or xtmixed

- Generalize random effects model to **random slopes**.
- Command `xtrc` estimates the **random coefficients model**

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}_i + \varepsilon_{it},$$

where $(\alpha_i, \boldsymbol{\beta}_i)$ are iid with mean $(\alpha, \boldsymbol{\beta})$ and variance matrix Σ and ε_{it} is iid.

- Not used in microeconometrics but used in many other disciplines.
- Stack all observations for individual i and specify

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

  where $\mathbf{u}_i$ is iid $(\mathbf{0}, \mathbf{G})$ and $\mathbf{Z}_i$ is called a design matrix.
- Random effects: $\mathbf{Z}_i = \mathbf{e}$ (a vector of ones) and $\mathbf{u}_i = \alpha_i$
- Random coefficients: $\mathbf{Z}_i = \mathbf{X}_i$.
- Example:

```
xtmixed lwage exp exp2 wks ed || id:  exp wks,
covar(unstructured) mle
```

9.1 Clustered data

- Consider data on individual i in village j with **clustering on village**.
- A **cluster-specific model** (here village-specific) specifies

$$y_{ji} = \alpha_i + \mathbf{x}'_{ji}\boldsymbol{\beta} + \varepsilon_{ji}.$$

- Here clustering is on village (not individual) and the repeated measures are over individuals (not time).
- Use `xtset village id`
- Assuming **equicorrelated errors** can be more reasonable here than with panel data (where correlation dampens over time).
  So perhaps less need for `vce(cluster)` after `xtreg`

9.2 Estimators for clustered data

- First use `xtset village person` (versus `xtset id t` for panel).
- If α_i is **random** use:
  - `regress` with option `vce(cluster village)`
  - `xtreg,re`
  - `xtgee` with option `exchangeable`
  - `xtmixed` for richer models of error structure
- If α_i is **fixed** use:
  - `xtreg,fe`

10.1 Nonlinear panel models overview

- **General approaches** similar to linear case
  - Pooled estimation or population-averaged
  - Random effects
  - Fixed effects
- **Complications**
  - Random effects often not tractable so need numerical integration
  - Fixed effects models in short panels are generally not estimable due to the incidental parameters problem.
- Here we consider **short panels** throughout.
- **Standard nonlinear models** are:
  - Binary: logit and probit
  - Counts: Poisson and negative binomial
  - Truncated: Tobit

10.2 Nonlinear panel models

- A **pooled** or **population-averaged model** may be used.
  This is same model as in cross-section case, with adjustment for correlation over time for a given individual.
- A **fully parametric model** may be specified, with conditional density

$$f(y_{it}|\alpha_i, \mathbf{x}_{it}) = f(y_{it}, \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}, \gamma), \quad t = 1, ..., T_i, \; i = 1, ..., N, \quad (5)$$

  where γ denotes additional model parameters such as variance parameters and α_i is an individual effect.
- A **conditional mean model** may be specified, with **additive effects**

$$\mathrm{E}[y_{it}|\alpha_i, \mathbf{x}_{it}] = \alpha_i + g(\mathbf{x}_{it}'\boldsymbol{\beta}) \quad (6)$$

  or **multiplicative effects**

$$\mathrm{E}[y_{it}|\alpha_i, \mathbf{x}_{it}] = \alpha_i \times g(\mathbf{x}_{it}'\boldsymbol{\beta}). \quad (7)$$

10.3 Nonlinear panel commands

	Counts	Binary
Pooled	poisson	logit
	nbreg	probit
GEE (PA)	xtgee,family(poisson)	xtgee,family(binomial) link(logit
	xtgee,family(nbinomial)	xtgee,family(poisson) link(probit
RE	xtpoisson, re	xtlogit, re
	xtnbreg, fe	xtprobit, re
Random slopes	xtmepoisson	xtmelogit
FE	xtpoisson, fe	xtlogit, fe
	xtnbreg, fe	

plus `tobit` and `xttobit`.

11.1 Pooled or Population-averaged estimation

- Extend pooled OLS
  - Give the usual cross-section command for conditional mean models or conditional density models but then get cluster-robust standard errors
  - Probit example:

    ```
    probit y x, vce(cluster id)
    ```

    or

    ```
    xtgee y x, fam(binomial) link(probit) corr(ind)
    vce(cluster id)
    ```

- Extend pooled feasible GLS
  - Estimate with an assumed correlation structure over time
  - Equicorrelated probit example:

    ```
    xtprobit y x, pa vce(boot)
    ```

    or

    ```
    xtgee y x, fam(binomial) link(probit) corr(exch)
    vce(cluster id)
    ```

11.2 Random effects estimation

- Assume individual-specific effect α_i has specified distribution $g(\alpha_i|\boldsymbol{\eta})$.
- Then the unconditional density for the i^{th} observation is

$$
\begin{aligned}
& f(y_{it}, \dots, y_{iT}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, \boldsymbol{\beta}, \gamma, \boldsymbol{\eta}) \\
= & \int \left[\prod_{t=1}^{T} f(y_{it}|\mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}, \gamma)\right] g(\alpha_i|\boldsymbol{\eta}) d\alpha_i. \qquad (8)
\end{aligned}
$$

- **Analytical solution:**
  - For Poisson with gamma random effect
  - For negative binomial with gamma effect
  - Use `xtpoisson, re` and `xtnbreg, re`
- **No analytical solution:**
  - For other models.
  - Instead use numerical integration (only univariate integration is required).
  - Assume normally distributed random effects.
  - Use `re` option for `xtlogit`, `xtprobit`
  - Use `normal` option for `xtpoisson` and `xtnbreg`

11.2 Random slopes estimation

- Can extend to **random slopes**.
  - Nonlinear generalization of `xtmixed`
  - Then higher-dimensional numerical integral.
  - Use adaptive Gaussian quadrature
- Stata commands are:
  - `xtmelogit` for binary data
  - `xtmepoisson` for counts
- Stata add-on that is very rich:
  - `gllamm` (generalized linear and latent mixed models)
  - Developed by Sophia Rabe-Hesketh and Anders Skrondal.

11.3 Fixed effects estimation

- In general not possible in short panels.
- **Incidental parameters problem:**
  - N fixed effects α_i plus K regressors means $(N+K)$ parameters
  - But $(N+K) \rightarrow \infty$ as $N \rightarrow \infty$
  - Need to eliminate α_i by some sort of differencing
  - possible for Poisson, negative binomial and logit.
- Stata commands
  - `xtlogit, fe`
  - `xtpoisson, fe` (better to use `xtpqml` as robust se's)
  - `xtnbreg, fe`
- Fixed effects extended to **dynamic models** for logit and probit. No Stata command.

12. Conclusion

- Stata provides commands for panel models and estimators commonly used in microeconometrics and biostatistics.
- Stata also provides diagnostics and postestimation commands, not presented here.
- The emphasis is on short panels. Some commands provide cluster-robust standard errors, some do not.
- A big distinction is between fixed effects models, emphasized by microeconometricians, and random effects and mixed models favored by many others.
- Extensions to nonlinear panel models exist, though FE models may not be estimable with short panels.
- This presentation draws on two chapters in Cameron and Trivedi, *Microeconometrics using Stata*, forthcoming.

Book Outline

For Cameron and Trivedi, *Microeconometrics using Stata*, forthcoming.

1. Stata basics
2. Data management and graphics
3. Linear regression basics
4. Simulation
5. GLS regression
6. Linear instrumental variable regression
7. Quantile regression
8. Linear panel models
9. Nonlinear regression methods
10. Nonlinear optimization methods
11. Testing methods
12. Bootstrap methods

Book Outline (continued)

13. Binary outcome models
14. Multinomial models
15. Tobit and selection models
16. Count models
17. Nonlinear panel models
18. Topics

A. Programming in Stata

B. Mata

Econometrics graduate-level panel data texts

- Comprehensive panel texts
  - Baltagi, B.H. (1995, 2001, 200?), *Econometric Analysis of Panel Data*, 1st and 2nd editions, New York, John Wiley.
  - Hsiao, C. (1986, 2003), *Analysis of Panel Data*, 1st and 2nd editions, Cambridge, UK, Cambridge University Press.
- More selective advanced panel texts
  - Arellano, M. (2003), *Panel Data Econometrics*, Oxford, Oxford University Press.
  - Lee, M.-J. (2002), *Panel Data Econometrics: Methods-of-Moments and Limited Dependent Variables*, San Diego, Academic Press.
- Texts with several chapters on panel
  - Cameron, A.C. and P.K. Trivedi (2005), *Microeconometrics: Methods and Applications*, New York, Cambridge University Press.
  - Greene, W.H. (2003), *Econometric Analysis*, fifth edition, Upper Saddle River, NJ, Prentice-Hall.
  - Wooldridge, J.M. (2002, 200?), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA, MIT Press.