Day 2A

Inference for Clustered Data: Part 1
With a Panel Data Example

A. Colin Cameron
Univ. of Calif. - Davis
... for
Center of Labor Economics
Norwegian School of Economics
Advanced Microeconometrics

Aug 28 - Sep 1, 2017
Introduction: Why Does Clustering Matter?

- By clustered data we mean data with the property that
  - observations in the same cluster or group are correlated
  - observations in the different clusters or group are uncorrelated
  - the clusters are known (unlike in cluster analysis).

- Failure to control for clustering in regression usually
  - underestimates standard errors
  - overstates t statistics and p-values and provides too narrow confidence intervals.

- These slides focus on panel data example
  - individuals uncorrelated but observations over time correlated for a given individual.

- The subsequent set of slides focus on cross-section example
  - individuals in groups such as a village uncorrelated but individuals in different villages.

- Focus is on OLS but most analysis generalizes to nonlinear models.
Outline

1. Introduction
2. Motivation
3. Summarizing Clustered Data: Panel Example
4. Pooled OLS for Clustered Data: Panel Example
5. Cluster-Robust Inference for OLS
6. Pooled Feasible GLS
7. Cluster-Specific Random Effects
8. Cluster-Specific Fixed Effects
9. Fixed versus Random Effects
10. First Differences
11. Estimator Comparison
12. Panel Bootstrap and Jackknife
13. Clustered Cross-section Data
2. Motivation: A simple example

- Suppose we have univariate data $y_i \sim (\mu, \sigma^2)$.
- We estimate $\mu$ by $\bar{y}$ and

$$\text{Var}[\bar{y}] = \text{Var} \left[ \frac{1}{N} \sum_{i=1}^{N} y_i \right] = \frac{1}{N^2} \left[ \sum_{i=1}^{N} \sum_{j=1}^{N} \text{Cov}(y_i, y_j) \right].$$

- **Given independence** over $i$
  - this simplifies to $\text{Var}[\bar{y}] = \frac{1}{N} \sigma^2$
  - since $\text{Cov}(y_i, y_j) = 0$ so we have just $\frac{1}{N^2} \left[ \sum_{i=1}^{N} \text{Var}(y_i) \right] = \frac{1}{N^2} \times N\sigma^2$.

- Now **suppose observations are positively correlated**
  - then there are lots of $\text{Cov}(y_i, y_j)$ terms to add in, increasing $\text{Var}[\bar{y}]$. 
2. Motivation

A simple example

Intuition (continued)

- Assume **equicorrelation** with \( \text{Cov}(y_i, y_j) = \rho \sigma^2 \) for \( i \neq j \)

\[
\begin{pmatrix}
    y_1 \\
y_2 \\
    \vdots \\
y_N
\end{pmatrix} =
\begin{pmatrix}
    \sigma_{11} & \cdots & \cdots & \sigma_{1N} \\
    \vdots & \ddots & \vdots & \vdots \\
    \vdots & \ddots & \ddots & \vdots \\
    \sigma_{N1} & \cdots & \cdots & \sigma_{NN}
\end{pmatrix}
= \sigma^2
\begin{pmatrix}
    1 & \rho & \cdots & \rho \\
    \rho & 1 & \cdots & \rho \\
    \vdots & \vdots & \ddots & \vdots \\
    \rho & \cdots & \cdots & 1
\end{pmatrix}
\]

- Then \( \text{Var}[\bar{y}] = \frac{1}{N^2} \sigma^2 [N + N(N - 1)\rho] = \frac{1}{N} \sigma^2 \{1 + (N - 1)\rho \} \).
- The variance is \( 1 + (N - 1)\rho \) times **larger**!
  - if \( N = 81 \) and \( \rho = 0.1 \) then \( \text{Var}[\bar{y}] \) is 9 times larger than \( \frac{1}{N} \sigma^2 \).
- Reason: **An extra observation is not providing a new independent piece of information.**
Example 1: Difference-in-Differences State-Year Panel

- Example: How do wages respond to a policy indicator variable $d_{ts}$ that varies by state
  - e.g. $d_{ts} = 1$ if minimum wage law in effect
- OLS estimate model for state $s$ at time $t$
  \[ y_{ts} = \alpha + x'_{ts}\beta + \gamma \times d_{ts} + u_{ts}. \]

  - The regressor $d_{ts}$ will be highly correlated within cluster.
  - Problems if additionally errors $u_{ts}$ are correlated within cluster i.e. model systematically overpredicts (or underpredicts) wage in all years for state $s$
    - Cor[$u_{ts}, u_{t's'}$] ≠ 0 if $s = s'$
- Natural model for error is a time series model
  - e.g. AR(1): Var[$u_{ts}$] = $\sigma^2$ and Cor[$u_{ts}, u_{t+k,s}$] = $\rho^k \sigma^2$.
  - We want to do inference without specifying such a model.
- Same problems if individuals in states over time
  \[ y_{its} = \alpha + x'_{its}\beta + \gamma \times d_{ts} + u_{its}. \]
Example 2: Individuals in Cluster ("Moulton-type")

  - CPS individual data on male wages $N = 5960$.
  - But there is no individual data on job injury rate.
  - Instead aggregated data on industry injury rates for 211 industries.

- OLS estimate model for individual $i$ in industry $g$

\[ y_{ig} = \alpha + x^T_{ig} \beta + \gamma \times rind_g + u_{ig}. \]

- The regressor $rind_g$ is perfectly correlated within cluster.

- Problems if additionally errors $u_{ig}$ are correlated within cluster
  i.e. model systematically overpredicts (or underpredicts) wage for all
  individuals in industry $g$
  - $\text{Cor}[u_{ig}, u_{jh}] \neq 0$ if $g = h$

- Natural model for error is within-cluster equicorrelation
  - $\text{Var}[u_{ig}] = \sigma^2$ and $\text{Cor}[u_{ig}, u_{jg}] = \rho \sigma^2$.
  - We want to do inference without specifying such a model.
Moulton (1986, 1990) and Bertrand, Duflo & Mullainathan (2004) showed

- the practical importance of controlling for clustering
- clustering can arise in a wider range of settings than obvious.

To control for clustering

- originally use a restrictive one-way random effects model
- now use cluster-robust standard errors
3. Panel Data Summary: Wages Example

- This handout considers a panel data example
  - the next handout considers a cross-section clustering example
- Note: For investigating and estimating Moulton-type data
  - the xt panel commands can also be used
  - though only a subset are relevant.
- We present Stata xt commands in detail.
- We consider OLS, FGLS, random effects and fixed effects estimators.
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
  - see Cameron and Trivedi (2010) chapter 8.11.

- 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))
Summarize Data using Non-panel Commands

```
. * Describe dataset
. describe

Contains data from mus08psidextract.dta
obs: 4,165 vars: 15
size: 283,220 (97.5% of memory free) (_dta has notes)

+---------------------------------+---------------------------------+---------------------------------+
<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>type</th>
<th>format</th>
<th>label</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>years of full-time work experience</td>
</tr>
<tr>
<td>wks</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>weeks worked</td>
</tr>
<tr>
<td>occ</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>occupation; occ==1 if in a blue-collar occupation</td>
</tr>
<tr>
<td>ind</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>industry; ind==1 if working in a manufacturing industry</td>
</tr>
<tr>
<td>south</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>residence; south==1 if in the South area</td>
</tr>
<tr>
<td>smsa</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>smsa==1 if in the Standard metropolitan statistical area</td>
</tr>
<tr>
<td>ms</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>marital status</td>
</tr>
<tr>
<td>fem</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>female or male</td>
</tr>
<tr>
<td>union</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>if wage set be a union contract</td>
</tr>
<tr>
<td>ed</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>years of education</td>
</tr>
<tr>
<td>blk</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>black</td>
</tr>
<tr>
<td>lwage</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td>log wage</td>
</tr>
<tr>
<td>id</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp2</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
+---------------------------------+---------------------------------+---------------------------------+
```

PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990)
16 Aug 2007 16:29

Contains data from mus08psidextract.dta
Summary statistics combine variation over $i$ and $t$.

```
* Summarize dataset
summarize
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp</td>
<td>4165</td>
<td>19.85378</td>
<td>10.96637</td>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>wks</td>
<td>4165</td>
<td>46.81152</td>
<td>5.129098</td>
<td>5</td>
<td>52</td>
</tr>
<tr>
<td>occ</td>
<td>4165</td>
<td>.5111645</td>
<td>.4999354</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ind</td>
<td>4165</td>
<td>.3954382</td>
<td>.4890033</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>south</td>
<td>4165</td>
<td>.2902761</td>
<td>.4539442</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>smsa</td>
<td>4165</td>
<td>.6537815</td>
<td>.475821</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ms</td>
<td>4165</td>
<td>.8144058</td>
<td>.3888256</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>fem</td>
<td>4165</td>
<td>.112605</td>
<td>.3161473</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>union</td>
<td>4165</td>
<td>.3639856</td>
<td>.4812023</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>4165</td>
<td>12.84538</td>
<td>2.787995</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>blk</td>
<td>4165</td>
<td>.0722689</td>
<td>.2589637</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>lwage</td>
<td>4165</td>
<td>6.676346</td>
<td>.4615122</td>
<td>4.60517</td>
<td>8.5376</td>
</tr>
<tr>
<td>id</td>
<td>4165</td>
<td>298</td>
<td>171.7821</td>
<td>1</td>
<td>595</td>
</tr>
<tr>
<td>t</td>
<td>4165</td>
<td>4</td>
<td>2.00024</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>exp2</td>
<td>4165</td>
<td>514.405</td>
<td>496.9962</td>
<td>1</td>
<td>2601</td>
</tr>
</tbody>
</table>

Since 4165 ($= 7 \times 595$) observations for all variables the dataset is balanced and complete.
• Listing the first few observations is useful

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

<table>
<thead>
<tr>
<th>id</th>
<th>t</th>
<th>exp</th>
<th>wks</th>
<th>occ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>40</td>
<td>0</td>
</tr>
</tbody>
</table>

• Data are in long form, sorted by id and then by t
Stata Commands for Panel Data Summary

- 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.
3. Panel Data Summary

Summarize Data using Panel Commands

- `xtset` command defines $i$ and $t$.
  - Allows use of panel commands and some time series operators
  - For Moulton-type data only `xtset id`

- * Declare individual identifier and time identifier
- `xtset id t`
  - panel variable: id (strongly balanced)
  - time variable: t, 1 to 7
  - delta: 1 unit
**xtdescribe** command summarizes number of time periods each individual is observed.

```
. * 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  100.00  | xxxxxxxx
```

- Data are balanced with every individual $i$ having 7 time periods of data.
- **xtsum** command splits overall variation into
  - between variation: variation in $\bar{x}_i = T_i^{-1} \sum_i x_{it}$ across individuals
  - within variation: variation in $x_{it}$ around $\bar{x}_i$

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>lwage</td>
<td>6.676346</td>
<td>.4615122</td>
<td>4.60517</td>
<td>8.537</td>
<td>N = 4165</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>.3942387</td>
<td>5.3364</td>
<td>7.813596</td>
<td>n = 595</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>.2404023</td>
<td>4.781808</td>
<td>8.621092</td>
<td>T = 7</td>
</tr>
<tr>
<td>exp</td>
<td>19.85378</td>
<td>10.96637</td>
<td>1</td>
<td>51</td>
<td>N = 4165</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>10.79018</td>
<td>4</td>
<td>48</td>
<td>n = 595</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>2.00024</td>
<td>16.85378</td>
<td>22.85378</td>
<td>T = 7</td>
</tr>
<tr>
<td>ed</td>
<td>12.84538</td>
<td>2.787995</td>
<td>4</td>
<td>17</td>
<td>N = 4165</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>2.790006</td>
<td>4</td>
<td>17</td>
<td>n = 595</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>0</td>
<td>12.84538</td>
<td>12.84538</td>
<td>T = 7</td>
</tr>
<tr>
<td>t</td>
<td>4</td>
<td>2.00024</td>
<td>1</td>
<td>7</td>
<td>N = 4165</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>n = 595</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>2.00024</td>
<td>1</td>
<td>7</td>
<td>T = 7</td>
</tr>
</tbody>
</table>

- 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 is zero.
- `xttab` command provides more detail for discrete-valued variable.

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

<table>
<thead>
<tr>
<th>south</th>
<th>Overall</th>
<th>Between</th>
<th>Within</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq.</td>
<td>Percent</td>
<td>Freq.</td>
</tr>
<tr>
<td></td>
<td>2956</td>
<td>70.97</td>
<td>428</td>
</tr>
<tr>
<td>0</td>
<td>1209</td>
<td>29.03</td>
<td>182</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4165</td>
<td>100.00</td>
<td>610</td>
</tr>
</tbody>
</table>

  (n = 595)

- 29.03% on average were in the south.
- 30.59% were ever in the south.
- 94.9% of those ever in the south were always in the south.
\textbf{xttrans} provides transition probabilities for discrete-valued variable.

\begin{verbatim}
. * Transition probabilities for a variable
. xttrans south, freq

<table>
<thead>
<tr>
<th>residence; south==1 if in the South area</th>
<th>residence; south==1 if in the South area</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2,527</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>99.68</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,535</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>1,027</td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td>99.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>2,535</td>
<td>1,035</td>
</tr>
<tr>
<td></td>
<td>71.01</td>
<td>28.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
</tr>
</tbody>
</table>
\end{verbatim}

For the 28.99\% of the sample ever in the south, 99.23\% remained in the south the next period.
3. Panel Data Summary

Summarize Data using Panel Commands

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

- Much autocorrelation in each person’s wage.
Autocorrelations

- Because `xtset` set a time variable can use time series commands
  - `Lj.x` gives `x` lagged `j` periods.
- Can compute autocorrelations for a variable.

``` stata
. * 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)
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>lwage</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1.</td>
<td>0.9238</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2.</td>
<td>0.9083</td>
<td>0.9271</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3.</td>
<td>0.8753</td>
<td>0.8843</td>
<td>0.9067</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L4.</td>
<td>0.8471</td>
<td>0.8551</td>
<td>0.8833</td>
<td>0.8990</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L5.</td>
<td>0.8261</td>
<td>0.8347</td>
<td>0.8721</td>
<td>0.8641</td>
<td>0.8667</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>L6.</td>
<td>0.8033</td>
<td>0.8163</td>
<td>0.8518</td>
<td>0.8465</td>
<td>0.8594</td>
<td>0.9418</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

- High serial correlation: \( \text{Cor}[y_t, y_{t-6}] = 0.80 \) and not AR(1).
- Note that estimated autocorrelations without imposing stationarity.
- Weakness: Only 595 observations used as needed `L6.lwage`
Autocorrelations - better

- Command `pwcorr` uses all the available data

```
L6.lwage  0.8033  0.8163  0.8518  0.8465  0.8594  0.9418  1.0000
L5.lwage  0.8220  0.8444  0.8602  0.8631  0.8944  1.0000
L4.lwage  0.8460  0.8565  0.8684  0.8988  1.0000
L3.lwage  0.8649  0.8748  0.9044  1.0000
L2.lwage  0.8858  0.9128  1.0000
L.lwage   0.9189  1.0000
lwage     1.0000
```

- The following uses as much data as is available
  `pwcorr lwage L.lwage L2.lwage L3.lwage L4.lwage L5.lwage L6.lwage`

```
lwage    L.lwage L2.lwage L3.lwage L4.lwage L5.lwage L6.lwage
lwage     1.0000
L.lwage   0.9189  1.0000
L2.lwage  0.8858  0.9128  1.0000
L3.lwage  0.8649  0.8748  0.9044  1.0000
L4.lwage  0.8460  0.8565  0.8684  0.8988  1.0000
L5.lwage  0.8220  0.8444  0.8602  0.8631  0.8944  1.0000
L6.lwage  0.8033  0.8163  0.8518  0.8465  0.8594  0.9418  1.0000
```
Autocorrelations - nonstationary

- The preceding imposed stationarity: \( \text{Corr}[y_{it}, y_{it-j}] = \text{Corr}[y_{it+k}, y_{it+k-j}] \).
- Following does not constrain e.g. correlation between years 1 and 2 to equal that between 2 and 3 (so no longer stationary)

```stata
* First-order autocorrelation differs in different year pairs
. forvalues s = 2/7 {
    2.    quietly corr lwage L1.lwage if t == `s'
    3.    display "Autocorrelation at lag 1 in year `s' = " %6.3f r(rho)
    4. }
```

Autocorrelation at lag 1 in year 2 = 0.942
Autocorrelation at lag 1 in year 3 = 0.867
Autocorrelation at lag 1 in year 4 = 0.899
Autocorrelation at lag 1 in year 5 = 0.907
Autocorrelation at lag 1 in year 6 = 0.927
Autocorrelation at lag 1 in year 7 = 0.924
4. Pooled OLS (a Population-Averaged Estimator)

- Pooled OLS is regular OLS of $y_{it}$ on $x_{it}$
  - Consistent if $x_{it}$ is uncorrelated with the error $u_{it}$.

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

|        | 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 |

- Important: The default standard errors are too small
  - they erroneously assume errors are independent over $t$ for given $i$.
  - this assumes more information content from data then is the case.
Cluster-Robust Standard Errors

- Should instead use cluster-robust standard errors

``` stata
. * 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)
```

|   | Coef. | Robust Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|---|-------|------------------|------|------|---------------------|
| lwage |         |                  |      |      |                     |
| 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 here 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.
OLS with Clustered Errors Theory

- Model for $G$ clusters with $N_g$ individuals per cluster:

  \[
  y_{ig} = x_{ig}' \beta + u_{ig}, \quad i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \\
  y_g = X_g \beta + u_g, \quad g = 1, \ldots, G, \\
  y = X \beta + u.
  \]

- OLS estimator

  \[
  \hat{\beta} = \left( \sum_{g=1}^{G} \sum_{i=1}^{N_g} x_{ig} x_{ig}' \right)^{-1} \left( \sum_{g=1}^{G} \sum_{i=1}^{N_g} x_{ig} y_{ig} \right) \\
  = \left( \sum_{g=1}^{G} x_g' x_g \right)^{-1} \left( \sum_{g=1}^{G} x_g' y_g \right) \\
  = (X'X)^{-1} X'y.
  \]
As usual

\[ \hat{\beta} = \beta + (X'X)^{-1}X'u \]

\[ = \beta + (X'X)^{-1}\left(\sum_{g=1}^{G} X_g u_g\right). \]

Assume independence over \( g \) and correlation within \( g \)

\[ E[u_{ig}u_{jg'}|x_{ig}, x_{jg'}] = 0, \text{ unless } g = g'. \]

Then \( \hat{\beta} \sim \mathcal{N} [\beta, V[\hat{\beta}]] \) with asymptotic variance

\[ \text{Avar}[\hat{\beta}] = (E[X'X])^{-1}(\sum_{g=1}^{G} E[X'g' u_g u_g' X_g])(E[X'X])^{-1} \]

\[ \neq \sigma_u^2 (E[X'X])^{-1}. \]
Consequences - KEY RESULT FOR INSIGHT

- Suppose equicorrelation within cluster $g$

$$\text{Cor}[u_{ig}, u_{jg} | x_{ig}, x_{jg}] = \begin{cases} 1 & i = j \\ \rho_{u} & i \neq j \end{cases}$$

- this arises in a random effects model with $u_{ig} = \alpha_g + \varepsilon_{ig}$, where $\alpha_g$ and $\varepsilon_{ig}$ are i.i.d. errors.
- an example is individual $i$ in village $g$ or student $i$ in school $g$.

- The incorrect default OLS variance estimate should be inflated by

$$\tau_j \simeq 1 + \rho_{xj} \rho_u (\bar{N}_g - 1),$$

- (1) $\rho_{xj}$ is the within cluster correlation of $x_j$
- (2) $\rho_u$ is the within cluster error correlation
- (3) $\bar{N}_g$ is the average cluster size.
- Need both (1) and (2) and it also increases with (3)
Moulton (1986, 1990) showed that the inflation can be large even if $\rho_u$ is small

- especially with a grouped regressor (same for all individuals in group) so that $\rho_x = 1$.
- CPS data example: $N_g = 81$, $\rho_x = 1$ and $\rho_u = 0.1$
  then $\tau_j \approx 1 + \rho_x \rho_u (N_g - 1) = 1 + 1 \times 0.1 \times 80 = 9$.
  - true standard errors are three times the default!

So should correct for clustering even in settings where not obviously a problem.

Bertrand, Duflo and Mullainathan (2004) showed such problems also arise for difference-in-differences analysis with individual in state-year panel

- and cluster on state, not state-year pair.
The Cluster-Robust Variance Matrix Estimate

- Recall for OLS with independent heteroskedastic errors
  \[
  \text{Avar}[\hat{\beta}] = (E[XX'])^{-1} \left( \sum_{i=1}^{N} E[u_i^2 x_i' x_i'] \right) (E[XX'])^{-1}
  \]
  can be consistently estimated (White (1980)) as \( N \to \infty \) by
  \[
  \hat{V}[\hat{\beta}] = (X'X)^{-1} \left( \sum_{i=1}^{N} \hat{u}_i^2 x_i' x_i' \right) (X'X)^{-1}.
  \]

- Similarly for OLS with independent clustered errors
  \[
  \text{Avar}[\hat{\beta}] = (E[XX'])^{-1} \left( \sum_{g=1}^{G} E[X_g u_g u_g' X_g] \right) (E[XX'])^{-1}
  \]
  can be consistently estimated as \( G \to \infty \) by the cluster-robust variance estimate (CRVE)
  \[
  \hat{V}_{\text{CR}}[\hat{\beta}] = (X'X)^{-1} \left( \sum_{g=1}^{G} X_g \hat{u}_g \hat{u}_g' X_g \right) (X'X)^{-1}.
  \]

- \text{regress uses } \hat{u}_g = c \hat{u}_g = c(y_g - X_g \hat{\beta}) \text{ where}
  \[
  c = \frac{G}{G-1} \frac{N-1}{N-K} \approx \frac{G}{G-1}.
  \]
The CRVE was

- proposed by White (1984) for balanced case
- proposed by Liang and Zeger (1986) for grouped data
- proposed by Arellano (1987) for FE estimator for short panels (group on individual)
- Hansen (2007a) and Carter, Schnepel and Steigerwald (2013) also allow $N_g \rightarrow \infty$.
- popularized by incorporation in Stata as the cluster option (Rogers (1993)).
- also allows for heteroskedasticity so is cluster- and heteroskedastic-robust.

Stata with cluster identifier `id_clu`

- `regress y x, vce(cluster id_clu)`
- `xtreg y x, pa corr(ind) vce(robust)`
  - after `xtset id_clu`
  - from version 12.1 on Stata interprets `vce(robust)` as cluster-robust for all `xt` commands.
Some Limitations

- Theory requires number of clusters $G \to \infty$
  - problem if e.g. the cluster is state and there are few states
  - this is considered in the subsequent of slides.

- The rank of $\hat{V}_{CR}[\hat{\beta}]$ is at most the minimum of $k$ (the dimension of $\beta$) and $G - 1$
  - if $k > G - 1$ it is possible to test at most $G - 1$ restrictions
  - then output gives no overall $F$-test but individual $t$-tests are okay.
5. Feasible GLS with Cluster-Robust Inference

- Potential efficiency gains for feasible GLS compared to OLS.
- Specify a model for \( \Omega_g = \mathbb{E}[u_g u_g' | X_g] \), such as within-cluster equicorrelation.
- Given a consistent estimate \( \hat{\Omega} \) of \( \Omega \), the feasible GLS estimator of \( \beta \) is

\[
\hat{\beta}_{FGLS} = \left( \sum_{g=1}^{G} X_g' \hat{\Omega}_g^{-1} X_g \right)^{-1} \sum_{g=1}^{G} X_g' \hat{\Omega}_g^{-1} y_g.
\]

- Default \( \hat{V}[^{\beta}_{FGLS}] = (X' \hat{\Omega}^{-1} X)^{-1} \) requires correct \( \Omega \).
- To guard against misspecified \( \Omega_g \) uses cluster-robust

\[
\hat{V}_{CR}[^{\beta}_{FGLS}] = (X' \hat{\Omega}^{-1} X)^{-1} \left( \sum_{g=1}^{G} X_g' \hat{\Omega}_g^{-1} \hat{u}_g \hat{u}_g' \hat{\Omega}_g^{-1} X_g \right) (X' \hat{\Omega}^{-1} X)^{-1}
\]

- where \( \hat{u}_g = y_g - X_g \hat{\beta}_{FGLS} \) and \( \hat{\Omega} = \text{Diag}[\hat{\Omega}_g] \)
- assumes \( u_g \) and \( u_h \) are uncorrelated, for \( g \neq h \)
- and needs \( G \rightarrow \infty \).
FGLS Example 1

- Example 2 - Time series correlation for panel data and DiD
  - AR(1) error \( u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \) and \( \varepsilon_{it} \) i.i.d.
  - Implies Corr\([u_{i,t}, u_{i,t-k}] = \rho^k\).
  - Default VE: `xtreg y x, pa corr(ar 1)`
  - Cluster-robust VE: `xtreg y x, pa corr(ar 1) vce(robust)`

- Stata allows a range of correlation structures
  - exchangeable; independent; AR(p); MA(q); Kiefer(1980)

- Puzzle - why is FGLS not used more?
  - Easily done in Stata with robust VCE if \( G \to \infty \)
  - Unless FE’s present and \( N_g \) small (see later).
FGLS Example 2

- Example 1 - Random effects
  - $y_{ig} = x_{ig}' \beta + \alpha_g + \varepsilon_{ig}$ where $\alpha_g$ and $\varepsilon_{ig}$ are i.i.d. errors.
  - Within-cluster errors are equicorrelated or exchangeable
  - Default VE: `xtreg y x, pa corr(exch)`
  - Cluster-robust VE: `xtreg y x, pa corr(exch) vce(robust)`

- Richer variation of random effects is hierarchical linear model or mixed model
  - $y_{ig} = x_{ig}' \beta_g + u_{ig}$
  - $\beta_g = W_g \gamma + v_j$ where $u_{ig}$ and $v_g$ are errors
  - In Stata use `mixed`.
  - Special case is RE MLE
    - * `xtmixed y x || id_clu: , covar(unstr) mle`
AR(2) Error - Default Standard Errors

- More efficient Feasible GLS assuming an error correlation model over time.
  - Here specify AR(2): \( u_{it} = \rho_1 u_{i,t-1} + \rho_2 u_{i,t-2} + \epsilon_{it} \)

```
xtreg lwage $xlist, pa corr(ar 2) nolog
```

GEE population-averaged model

<table>
<thead>
<tr>
<th>Number of obs</th>
<th>4165</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group and time vars:</td>
<td>id t</td>
</tr>
<tr>
<td>Number of groups</td>
<td>595</td>
</tr>
<tr>
<td>Link:</td>
<td>identity</td>
</tr>
<tr>
<td>Obs per group:</td>
<td>min = 7</td>
</tr>
<tr>
<td>Family:</td>
<td>Gaussian</td>
</tr>
<tr>
<td>avg = 7.0</td>
<td></td>
</tr>
<tr>
<td>Correlation:</td>
<td>AR(2)</td>
</tr>
<tr>
<td>max = 7</td>
<td></td>
</tr>
<tr>
<td>Scale parameter:</td>
<td>.1966639</td>
</tr>
<tr>
<td>Wald chi2(4)</td>
<td>837.36</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

| lwage | Coef. | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|-------|-------|-----------|-----|-----|---------------------|
| exp   | 0.0718915 | 0.0042152 | 17.06 | 0.000 | 0.0636298 - 0.0801531 |
| exp2  | -0.0008966 | 0.000093 | -9.64 | 0.000 | -0.0010788 - 0.0007143 |
| wks   | 0.0002964 | 0.0006389 | 0.46  | 0.643 | -0.0009558 0.0015485 |
| ed    | 0.0905069 | 0.0059771 | 15.14 | 0.000 | 0.0787921 0.1022217 |
| _cons | 4.526381 | 0.0965424 | 46.88 | 0.000 | 4.337162 4.715601 |

. xtreg lwage $xlist, pa corr(ar 2) nolog
```

GEE population-averaged model

<table>
<thead>
<tr>
<th>Number of obs</th>
<th>4165</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group and time vars:</td>
<td>id t</td>
</tr>
<tr>
<td>Number of groups</td>
<td>595</td>
</tr>
<tr>
<td>Link:</td>
<td>identity</td>
</tr>
<tr>
<td>Obs per group:</td>
<td>min = 7</td>
</tr>
<tr>
<td>Family:</td>
<td>Gaussian</td>
</tr>
<tr>
<td>avg = 7.0</td>
<td></td>
</tr>
<tr>
<td>Correlation:</td>
<td>AR(2)</td>
</tr>
<tr>
<td>max = 7</td>
<td></td>
</tr>
<tr>
<td>Scale parameter:</td>
<td>.1966639</td>
</tr>
<tr>
<td>Wald chi2(4)</td>
<td>837.36</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
AR(2) Error - Cluster Robust Standard Errors

- Repeat previous estimator but with cluster-robust standard errors.
  - Robust se’s similar except for wks

```
. xtreg lwage exp exp2 wks ed, pa corr(ar 2) vce(robust) nolog
```

|             | Coef.   | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|-------------|---------|-----------|-------|------|---------------------|
| lwage       |         |           |       |      |                     |
| exp         | 0.0718915 | 0.003999   | 17.98 | 0.000 | 0.0640535 - 0.0797294 |
| exp2        | -0.0008966 | 0.0000933  | -9.61 | 0.000 | -0.0010794 - 0.0007137 |
| wks         | 0.0002964 | 0.0010553  | 0.28  | 0.779 | -0.001772 - 0.0023647 |
| ed          | 0.0905069 | 0.0060161  | 15.04 | 0.000 | 0.0787156 - 0.102982 |
| _cons       | 4.526381  | 0.1056897  | 42.83 | 0.000 | 4.319233 - 4.733529  |

- Same as xtgee lwage $xlist, pa corr(ar 2) vce(robust) nolog
6. Random Effects Estimator

- Random effects estimator is FGLS estimator for the RE model

\[ y_{it} = \alpha_i + x_{it}' \beta + \varepsilon_{it} \]

\[ \alpha_i \sim \text{i.i.d.}[\alpha, \sigma_\alpha^2] \]

\[ \varepsilon_{it} \sim \text{i.i.d.}[0, \sigma_\varepsilon^2] \]

- The RE model implies equicorrelated (or exchangeable) errors

  - \( \text{Var}[\alpha_i + \varepsilon_{it}] = \text{Var}[\alpha_i] + \text{Var}[\varepsilon_{it}] = \sigma_\alpha^2 + \sigma_\varepsilon^2 \)
  
  - For \( s \neq t \), \( \text{Cov}[\alpha_i + \varepsilon_{it}, \alpha_j + \varepsilon_{is}] = \text{Cov}[\alpha_i, \alpha_j] = \sigma_\alpha^2 \)
  
  - So \( \text{Corr}[\alpha_i + \varepsilon_{it}, \alpha_j + \varepsilon_{is}] = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2) \) for all \( s \neq t \).

- FGLS can be shown to equal OLS in the transformed model

\[ (y_{it} - \hat{\theta}_i \bar{y}_i) = (x_{it} - \hat{\theta}_i \bar{x}_i)' \beta + \text{error}, \]

where \( \hat{\theta}_i \) is a consistent estimate of \( \theta_i = 1 - \sqrt{\sigma_\varepsilon^2 / (T_i \sigma_\alpha^2 + \sigma_\varepsilon^2)} \).
Random effects estimates with cluster-robust standard errors:

\[
\begin{align*}
\rho &= 0.81505521 \quad \text{(fraction of variance due to } u_i) \\
\sigma_e &= 0.15220316 \\
\sigma_u &= 0.31951859 \\
_cons &= 3.829366 \quad 0.1333931 \quad 28.71 \quad 0.000 \quad 3.567921 \quad 4.090812 \\
ed &= 0.1117099 \quad 0.0083954 \quad 13.31 \quad 0.000 \quad 0.0952552 \quad 0.1281647 \\
wks &= 0.0009658 \quad 0.0009259 \quad 1.04 \quad 0.297 \quad -0.000849 \quad 0.0027806 \\
exp2 &= -0.0007726 \quad 0.0000896 \quad -8.62 \quad 0.000 \quad -0.0009481 \quad -0.000597 \\
exp &= 0.0888609 \quad 0.0039992 \quad 22.22 \quad 0.000 \quad 0.0810227 \quad 0.0966992 \\
lwage &= 3.829366 \quad 0.1333931 \quad 28.71 \quad 0.000 \quad 3.567921 \quad 4.090812 \\
sigma_u &= 0.31951859 \\
sigma_e &= 0.15220316 \\
rho &= 0.81505521 (fraction of variance due to } u_i) \\
\end{align*}
\]

\[
\text{xtreg lwage exp exp2 wks ed, re vce(robust) theta}
\]

Option theta gives \( \hat{\theta} = 0.82 = 1 - \sqrt{0.152^2 / (7 \times 0.319^2 + 0.152^2)} \).
7. Fixed Effects Estimator

- Mean-differencing eliminates $\alpha_i$

  $y_{it} = \alpha_i + x'_{it} \beta + \varepsilon_{it}$

  $\Rightarrow \bar{y}_i = \alpha_i + \bar{x}'_i \beta + \bar{\varepsilon}_i$

  $\Rightarrow (y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$

- The within or fixed effects estimator is OLS of $(y_{it} - \bar{y}_i)$ on $(x_{it} - \bar{x}_i)$
  - Efficiency loss as use only within variation
  - Coefficient of any time-invariant regressor is not identified $(x_{it} = \bar{x}_i)$
  - Use cluster-robust standard errors
  - Stata command `xtreg, fe`
### 7. Fixed Effects Estimator

**Within or FE estimates: Default SE's**

- **Variable ed is not identified because time-invariant regressor in this dataset!**

```stata
. xtreg lwage $xlist, fe
note: ed omitted because of collinearity
```

**Fixed-effects (within) regression**

- **Group variable: id**
- **Number of obs** = 4165
- **Number of groups** = 595
- **R-sq: within = 0.6566**
- **between = 0.0276**
- **overall = 0.0476**
- **F(3,3567) = 2273.74**
- **Prob > F = 0.0000**

**corr(u_i, Xb) = -0.9107**

| lwage  | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------|--------|-----------|-------|------|---------------------|
| exp    | 0.1137879 | 0.0024689  | 46.09 | 0.000 | 0.1089473 - 0.1186284 |
| exp2   | -0.0004244 | 0.0000546  | -7.77 | 0.000 | -0.0005315 - 0.0003173 |
| wks    | 0.0008359  | 0.0005997  | -1.39 | 0.163 | -0.0003399 - 0.0020116 |
| ed     | 0       |           |       |      |                      |
| _cons  | 4.596396  | 0.0389061  | 118.14| 0.000 | 4.520116 - 4.672677  |

**sigma_u = 1.0362039**

- **sigma_e = 0.15220316**
- **rho = 0.97888036** (fraction of variance due to u_i)

**F test that all u_i=0: F(594, 3567) = 56.52**

- **Prob > F = 0.0000**

---

A. Colin Cameron  Univ. of Calif. - Davis  
Clustered Data: Part 1  
Aug 28 - Sep 1, 2017  
41 / 64
Within or FE estimates: Cluster Robust se’s

- standard errors are 50% larger even after inclusion of fixed effect!

```
.xtreg lwage exp exp2 wks ed, fe vce(robust)
```

Fixed-effects (within) regression

| Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|-------|-----------|---|-----|---------------------|
| lwage |           |   |     |                     |
| exp   | 0.1137879 | 0.0040289 | 28.24 | 0.000 | 0.1058753 - 0.1217004 |
| exp2  | -0.0004244 | 0.0000822 | -5.16 | 0.000 | -0.0005858 - 0.0002629 |
| wks   | 0.0008359 | 0.0008697 | 0.96  | 0.337 | -0.0008721 - 0.0025439 |
| ed    | (dropped) |           |       |       |                     |
| _cons | 4.596396  | 0.0600887 | 76.49 | 0.000 | 4.478384 - 4.714408 |

(Std. Err. adjusted for 595 clusters in id)
Least Squares Dummy Variable Model for FE

- Several ways to compute FE estimator aside from `xtreg, fe`.
- Least squares dummy variables:
  - \(d_{ji,t}\) for \(j = 1, \ldots, N\) are \(N\) dummies equal to 1 if \(i = j\)
  - Estimate directly using `regress` or use `areg`

\[
y_{it} = \sum_{j=1}^{N} \alpha_i d_{ji,t} + x'_{it} \beta + \epsilon_{it}
\]

- Implementation in Stata
  - * FE model fitted as LSDV using `areg`
    \[
    \text{areg lwage exp exp2 wks ed, absorb(id) vce(cluster id)}
    \]
  - * FE model fitted using LSDV using `regress`
    \[
    \text{set matsize 800}
    \text{quietly xi: regress lwage exp exp2 wks ed i.id, vce(cluster id)}
    \text{estimates table, keep(exp exp2 wks ed _cons) b se b(%12.7f)}
    \]
Within or FE estimates obtained using `areg`

- cluster-robust se’s are approximately $\sqrt{7/6}$ too large

```
. * FE model fitted as LSDV using areg with cluster-robust standard errors
. areg lwage exp exp2 wks ed, absorb(id) vce(cluster id)
```

```
Linear regression, absorbing indicators

| Coef. | Std. Err. | t     | P>|t| | 95% Conf. Interval |
|-------|-----------|-------|------|-------------------|
| lwage |           |       |      |                   |
| exp   | .1137879  | .0043514 | 26.15 | 0.000 | .1052418 .1223339 |
| exp2  | -.0004244 | .0000888 | -4.78 | 0.000 | -.0005988 -.00025 |
| wks   | .0008359  | .0009393 | 0.89  | 0.374 | -.0010089 .0026806 |
| ed    | 0 (omitted) |       |       |       |                   |
| _cons | 4.596396  | .0648993 | 70.82 | 0.000 | 4.468936 4.723856 |

(Std. Err. adjusted for 595 clusters in id)
```

Root MSE = 0.1522
Adj R-squared = 0.8912
R-squared = 0.9068
Prob > F = 0.0000
F( 3, 594) = 908.44
Number of obs = 4,165

Note: ed omitted because of collinearity
7. Fixed Effects Estimator

Least Squares Dummy Variable Model

- **IMPORTANT:** if the LSDV approach is used with `regress` or `areg` then
  - cluster-robust standard errors in Stata are overstated
  - due to wrong degrees of freedom correction
  - especially if there are few observations per cluster.

- **Stata regress and areg uses** $c = \frac{G}{G-1} \frac{N-1}{N-Kall}$
  - Here $Kall = K + G$ (the $\beta$'s plus the dummies)
  - For $Ng = 2$ so $N = 2G$: $c = \frac{G}{G-1} \frac{2G-1}{2G-K-G} \approx 1 \times 2 = 2$
  - Where `xtreg` knows to use $c = \frac{G}{G-1} \frac{G}{G-1} \approx 1$
  - Lesson: with $G$ small use `xtreg`, `fe`
  - Otherwise if use `areg` or `regress` i. get se's that are approximately $\sqrt{\frac{G}{G-1}}$ times too big.
  - In panel setting $G = T$. 

A. Colin Cameron  Univ. of Calif. - Davis ...  Clumped Data: Part 1  Aug 28 - Sep 1, 2017  45 / 64
Aside: Mundlak/Chamberlain Model

- Mundlak and Chamberlain suppose the fixed effects

\[ \alpha_i = \bar{x}_i \pi + \text{error}. \]

- So OLS regress \( y_{it} \) on intercept \( x_{it} \) and \( \bar{x}_i \)
- Yields same \( \beta \) estimate as the FE estimator.

* FE model fitted by add mean of x as a regressor

global xlist exp exp2 wks ed

sort id

foreach x of varlist $xlist {
    by id: egen mean'x' = mean('x')
}

regress lwage exp exp2 wks ed mean*, vce(robust)
Aside: Between Estimator

- OLS of $\bar{y}_i$ on intercept and $\bar{x}_i$
  - `xtreg`, `be` has no heteroskedastic robust option but can bootstrap.
8. Fixed versus Random Effects Estimators

- RE has advantages: estimates all parameters & may be 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.
- Do not use \texttt{hausman} command – it requires that RE estimator is fully efficient (see next slide).
- Instead do one of the following
  - 1. Do a panel bootstrap of the Hausman test.
    - Test $H_0 : \gamma = 0$ in the auxiliary OLS regression
      $$
      (y_{it} - \hat{\theta} \bar{y}_i) = (1 - \hat{\theta}) \alpha + (x_{it} - \hat{\theta} x_i)' \beta + (x_{1it} - \bar{x}_{1i})' \gamma + \nu_{it},
      $$
    - where $x_{1i} \subset x_i$ denotes time-varying regressors only.
    - Use cluster-robust standard errors for this test.
    - Stata add-on \texttt{xtoverid} after \texttt{xtreg,re} does this.
Hausman Test Theory

- Hausman test compares to estimators $\hat{\theta}$ and $\tilde{\theta}$
  - Test $H_0: \text{plim}(\hat{\theta} - \tilde{\theta}) = 0$ against $H_a: \text{plim}(\hat{\theta} - \tilde{\theta}) \neq 0$.
  - e.g. OLS versus 2SLS with possible endogenous regressor
  - e.g. RE versus FE with possible fixed effect.

- Under $H_0$, as usual $(\hat{\theta} - \tilde{\theta}) \sim N[0, V[\hat{\theta} - \tilde{\theta}])$.

- So form $\chi^2$ statistic: $H = (\hat{\theta} - \tilde{\theta})' \left[ V[\hat{\theta} - \tilde{\theta}] \right]^{-1} (\hat{\theta} - \tilde{\theta})$
  - reject $H_0$ if $H > \chi^2$ critical value.

- Problem: To implement we need estimate of $V[\hat{\theta} - \tilde{\theta}]$.

- Hausman (1978) assumed $\hat{\theta}$ is fully efficient under $H_0$
  - then $\text{Cov}[\hat{\theta}, \tilde{\theta}] = \text{Var}[\hat{\theta}]$
  - implying $V[\hat{\theta} - \tilde{\theta}] = V[\hat{\theta}] + V[\tilde{\theta}] - 2 \times V[\hat{\theta}] = V[\tilde{\theta}] - V[\hat{\theta}]$
  - but we rarely have $\hat{\theta}$ fully efficient.
**Hausman Test Wrong**

* Wrong Hausman test assuming RE estimator is fully efficient under null hypothesis

```
. hausman FE_def RE_def, sigmamore
```

<table>
<thead>
<tr>
<th></th>
<th>FE_def</th>
<th>RE_def</th>
<th>Difference</th>
<th>sqrt(diag(V_b-V_B))</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp</td>
<td>.1137879</td>
<td>.0888609</td>
<td>.0249269</td>
<td>.0012778</td>
</tr>
<tr>
<td>exp2</td>
<td>-.0004244</td>
<td>-.0007726</td>
<td>.0003482</td>
<td>.0000285</td>
</tr>
<tr>
<td>wks</td>
<td>.0008359</td>
<td>.0009658</td>
<td>-.0001299</td>
<td>.0001108</td>
</tr>
</tbody>
</table>

\[
b = \text{consistent under Ho and Ha; obtained from } \text{xtreg}
\]

\[
B = \text{inconsistent under Ha, efficient under Ho; obtained from } \text{xtreg}
\]

**Test:**

\[
\text{Ho: difference in coefficients not systematic}
\]

\[
\chi^2(3) = (b-B)'[(V_b-V_B)^{(-1)}](b-B)
\]

\[
= 1513.02
\]

\[
\text{Prob}>\chi^2 = 0.0000
\]
Hausman Test Correct

- Following is manual

```
. global xlist exp exp2 wks

. foreach x of varlist $xlist {
   2.   by id: egen mean`x' = mean(`x')
   3. }

. quietly regress lwage exp exp2 wks meanexp meanexp2 meanwks, vce(cluster id)

. test meanexp meanexp2 meanwks

( 1)  meanexp = 0
( 2)  meanexp2 = 0
( 3)  meanwks = 0

F(  3,  594) =  630.59
Prob > F =  0.0000
```

- Get exactly same result using simpler

```
quietly regress lwage exp exp2 wks ///
meanexp meanexp2 meanwks, vce(cluster id)

```

- Can also use Stata add-on xtoverid after xtreg, re
9. First Difference Estimator

- First-differencing eliminates $\alpha_i$

$$y_{it} = \alpha_i + x_{it}'\beta + \varepsilon_{it}$$

$$\Rightarrow y_{i,t-1} = \alpha_i + x_{i,t-1}'\beta + \varepsilon_{i,t-1}$$

$$\Rightarrow (y_{it} - y_{i,t-1}) = (x_{it} - x_{i,t-1})'\beta + (\varepsilon_{it} - \varepsilon_{i,t-1})$$

- First differences estimator
  - OLS regression of $\Delta y_{it}$ on $\Delta x_{it}$, i.e. use first differences.
  - Coefficient of any time-invariant regressor is not identified ($x_{it} = x_{i,t-1}$).

- Not used much for basic FE model
  - FE estimator is fully efficient if $\varepsilon_{it}$ is iid $(0, \sigma_{\varepsilon}^2)$
  - FD estimator is fully efficient if $\varepsilon_{it} = \varepsilon_{i,t-1} + \nu_{it}$ where $\nu_{it}$ is iid $(0, \sigma_{\nu}^2)$
  - FE=FD if $T = 2$ as then $y_{i2} - \bar{y}_i = y_{i2} - \frac{y_{i1} + y_{i2}}{2} = (y_{i2} - y_{i1})/2$. 
First Difference Estimator (continued)

- No direct Stata command.
- Can regress $D.(\text{lwage } \$xlist), \text{vce(cluster id)}$
- More comparable to FE is regress with noconstant
  \begin{verbatim}
  regress D.(lwage $xlist), noconstant vce(cluster id)
  \end{verbatim}
  - FE is also noconstant, but then adds back in $\bar{y}$.
- FD is used in models with fixed effects and lagged dependent variable
  - e.g. $y_{it} = \alpha_i + \rho y_{i,t-1} + x'_{it} \beta + u_{it}$
  - then within estimator is inconsistent if short panel
  - instead do IV on FD model with lagged $y'_{i,t}$ as instruments
  - this is Arellano-Bond estimator \texttt{xtabond} (also add-on \texttt{xtabond2}).
First differences estimator

```
. regress d.lwage d.exp d.exp2 d.wks d.ed, noconstant vce(cluster id)
```

Note: _delete_ omitted because of collinearity

Linear regression

| D.lwage | Coef.   | Robust Std. Err. | t     | P>|t| | 95% Conf. Interval |
|---------|---------|------------------|-------|------|-------------------|
| exp D1. | .1170654| .0040974         | 28.57 | 0.000| [.1090182, .1251126] |
| exp2 D1.| -.0005321| .0000808        | -6.58 | 0.000| [-.0006908, -.0003734] |
| wks D1. | -.0002683| .0011783        | -0.23 | 0.820| [-.0025824, .0020459] |
| ed D1.  | (omitted)|                 |       |      |                   |
```

Number of obs = 3570
F( 3, 594) = 1035.19
Prob > F = 0.0000
R-squared = 0.2209
Root MSE = 0.18156

(Std. Err. adjusted for 595 clusters in id)
10. 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)
### Coefficients vary considerably across OLS, RE, FE and RE estimators.

- FE and RE similar as $\hat{\theta} = 0.82 \approx 1$.

### 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!
Standard Errors Comparison

- Compares default to panel-robust standard errors for RE and FE.

<table>
<thead>
<tr>
<th>Variable</th>
<th>RE_def</th>
<th>RE</th>
<th>FE_def</th>
<th>FE</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp</td>
<td>0.0889</td>
<td>0.0889</td>
<td>0.1138</td>
<td>0.1138</td>
</tr>
<tr>
<td></td>
<td>0.0028</td>
<td>0.0040</td>
<td>0.0025</td>
<td>0.0040</td>
</tr>
<tr>
<td>exp2</td>
<td>-0.0008</td>
<td>-0.0008</td>
<td>-0.0004</td>
<td>-0.0004</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>wks</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0008</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>0.0007</td>
<td>0.0009</td>
<td>0.0006</td>
<td>0.0009</td>
</tr>
<tr>
<td>ed</td>
<td>0.1117</td>
<td>0.1117</td>
<td>(omitted)</td>
<td>(omitted)</td>
</tr>
<tr>
<td></td>
<td>0.0061</td>
<td>0.0084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>3.8294</td>
<td>3.8294</td>
<td>4.5964</td>
<td>4.5964</td>
</tr>
<tr>
<td></td>
<td>0.0936</td>
<td>0.1334</td>
<td>0.0389</td>
<td>0.0601</td>
</tr>
<tr>
<td>N</td>
<td>4165</td>
<td>4165</td>
<td>4165</td>
<td>4165</td>
</tr>
</tbody>
</table>

Legend: b/se
11. Panel Bootstrap

- Do pairs bootstrap where resample \((y, x)\) over individuals \(i\) rather than observations \((i, t)\).
- Do \(B\) iterations of this step. On the \(b^{th}\) iteration:
  - form a sample of \(G\) clusters \(\{(y_1^*, X_1^*), \ldots, (y_G^*, X_G^*)\}\) by resampling with replacement \(G\) times from the original sample
  - obtain estimate \(\hat{\beta}_b\), \(b = 1, \ldots, B\).
- Then \(\hat{V}[\hat{\beta}] = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})'\), \(\bar{\beta} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b\).
- In Stata use cluster(id) option of bootstrap
  - but need to replace xtset id t with simply xtset id
  - and if FE then also add idcluster(newid) option
- This pairs panel (or clustered) bootstrap
  - yields essentially the same results as usual cluster-robust standard errors
  - is a bootstrap without asymptotic refinement.
- NOTE: Stata 14 introduced new random number generator giving different bootstrap results.
Panel Bootstrap OLS Estimator

. * OLS panel bootstrap using xtreg, pa and vce(boot)  
. xtreg lwage exp exp2 wks ed, pa corr(ind) vce(boot, reps(400) //>
> seed(10101) nodots)

GEE population-averaged model  
Number of obs = 4,165
Group variable: id  
Number of groups = 595
Link: identity  
Obs per group:
Family: Gaussian  
min = 7
Correlation: independent  
avg = 7.0
max = 7
Scale parameter: .1525603  
Wald chi2(4) = 306.46
Prob > chi2 = 0.0000
Pearson chi2(4165): 635.41 Deviance = 635.41
Dispersion (Pearson): .1525603 Dispersion = .1525603

(Replications based on 595 clusters in id)

| lwage | Observed Coef. | Bootstrap Std. Err. | z | P>|z| | Normal-based [95% Conf. Interval] |
|-------|----------------|---------------------|---|------|-----------------------------|
| exp   | .044675        | .004895             | 9.13 | 0.000 | .035081 - .0542691 |
| exp2  | -.0007156      | .0001166            | -6.14 | 0.000 | -.0009442 - -.0004871 |
| wks   | .005827        | .0018287            | 3.19 | 0.001 | .0022429 - .0094111 |
| ed    | .0760407       | .0052064            | 14.61 | 0.000 | .0658364 - .086245 |
| _cons | 4.907961       | .1362406            | 36.02 | 0.000 | 4.640934 - 5.174987 |

Note: the following gives exactly the same

  xtreg lwage exp exp2 wks ed, ///
  pa corr(ind) vce(boot, reps(400) seed(10101))
Panel Bootstrap RE Estimator

Random-effects GLS regression
Group variable: id
Number of obs = 4,165
Number of groups = 595

Obs per group:
   min =  7
   avg =  7.0
   max =  7

Wald chi2(4) = 683.66
Prob > chi2 = 0.0000

(Replications based on 595 clusters in id)

|        | Observed Coef. | Bootstrap Std. Err. | z     | P>|z|   | Normal-based [95% Conf. Interval] |
|--------|----------------|---------------------|-------|-------|----------------------------------|
|        | lwage          |                     |       |       |                                  |
| exp    | .0888609       | .0044227            | 20.09 | 0.000 | .0801926 - .0975293              |
| exp2   | -.0007726      | .0000935            | -8.26 | 0.000 | -.0009559 - -.0005892            |
| wks    | .0009658       | .0009027            | 1.07  | 0.285 | -.0008035 - .0027351             |
| ed     | .1117099       | .009053             | 12.34 | 0.000 | .0939664 - .1294535              |
| _cons  | 3.829366       | .156578             | 24.46 | 0.000 | 3.522479 - 4.136254              |
| sigma_u| .31951859      |                     |       |       |                                  |
| sigma_e| .15220316      |                     |       |       |                                  |
| rho    | .81505521      | (fraction of variance due to u_i) |       |       |                                  |

Same as xtreg lwage exp exp2 wks ed, ///
re vce(boot, reps(400) seed(10101) nodots)
Panel Bootstrap FE Estimator - add idcluster()

. * FE panel bootstrap using bootstrap: with cluster(id) and idcluster
. * Need to add idcluster for bootstrap of FE
. bootstrap _b, reps(400) seed(10101) cluster(id) idcluster(newid) ///
>   nodots: xtreg lwage exp exp2 wks ed, fe

Fixed-effects (within) regression
Group variable: id

R-sq:
within = 0.6566
between = 0.0276
overall = 0.0476

corr(u_i, Xb) = -0.9107

Obs per group:
min = 7
avg = 7.0
max = 7

Wald chi2(3) = 3475.91
Prob > chi2 = 0.0000

(Replications based on 595 clusters in id)

| lwage  | Observed Coef. | Bootstrap Std. Err. | z   | P>|z| | Normal-based [95% Conf. Interval] |
|--------|----------------|---------------------|-----|------|----------------------------------|
| exp    | .1137879       | .0042744            | 26.62 | 0.000 | .1054102  .1221655 |
| exp2   | -.0004244      | .0000855            | -4.96 | 0.000 | -.000592  -.0002568 |
| wks    | .0008359       | .0008405            | 0.99 | 0.320 | -.0008115 .0024833 |
| ed     | 0              | (omitted)           |     |      |                                  |
| _cons  | 4.596396       | .0733031            | 62.70 | 0.000 | 4.452725  4.740068 |

| sigma_u | 1.0362039 |
| sigma_e | .15220316 |
| rho     | .97888036 (fraction of variance due to u_i) |

● Same as xtreg lwage exp exp2 wks ed, ///
   fe vce(boot, reps(400) seed(10101) nodots)
Panel Jackknife

- An alternative re-sampling scheme is a leave-one-cluster-out jackknife.
- Let $\hat{\beta}_{-i}$ denote the estimator of $\beta$ when the $i^{th}$ cluster (here $i^{th}$ if $N$ individuals) is deleted.

$$
\hat{\sigma}_{jack;\text{boot}}^2[\beta] = \frac{N - 1}{N} \sum_{i=1}^{N} (\hat{\beta}_{-i} - \hat{\beta})(\hat{\beta}_{-i} - \hat{\beta})',
$$

where $\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_{-i}$.
- For Stata xt commands this is option vce(jackknife).
12. Clustered Cross-section Data

- **xt** commands are not just for panel data
  - can apply several of them to clustered cross-section data.
- Consider data on individual *i* in village *j* with **clustering on village**.
- A **cluster-specific model** (here village-specific) specifies
  \[ y_{ji} = \alpha_i + x_{ji}'\beta + \varepsilon_{ji}. \]
- Here clustering is on village (not individual) and the repeated measures are over individuals (not time).
- Assuming **equicorrelated errors** can be more reasonable here than with panel data (where correlation dampens over time).
  - So perhaps less need for `vce(robust)` after `xtreg`.
- This is done in the subsequent set of slides.
13. Summary of Stata Panel Commands

- Linear panel estimators for short panels with exogenous regressors

  **Panel summary**
  - `xtset`; `xtdescribe`; `xtsum`; `xtdata`;
  - `xtline`; `xttab`; `xttran`

  **Pooled OLS**
  - `regress`

  **Feasible GLS**
  - `xtreg`, `pa`
  - `xtgee`, `family(gaussian)`

  **Random effects**
  - `xtreg`, `re`; `xtregar`, `re`

  **Fixed effects**
  - `xtreg`, `fe`; `xtregar`, `fe`

  **Random slopes**
  - `mixed`; `quadchk`

  **First differences**
  - `regress` (with differenced data)