CLUSTER SAMPLES AND CLUSTERING

Jeff Wooldridge
Michigan State University
LABOUR Lectures, EIEF
October 18-19, 2011

1. The Linear Model with Cluster Effects
2. Cluster-Robust Inference with Large Group Sizes
3. Cluster Samples with Unit-Specific Panel Data
4. Clustering and Stratification
5. Two-Way Clustering
1. The Linear Model with Cluster Effects

- For each group or cluster $g$, let $\{(y_{gm}, x_g, z_{gm}) : m = 1, \ldots, M_g\}$ be the observable data, where $M_g$ is the number of units in cluster or group $g$, $y_{gm}$ is a scalar response, $x_g$ is a $1 \times K$ vector containing explanatory variables that vary only at the cluster or group level, and $z_{gm}$ is a $1 \times L$ vector of covariates that vary within (as well as across) groups.
• Without a cluster identifier, a cluster sample looks like a cross section data set. Statistically, the key difference is that the sample of clusters has been drawn from a “large” population of clusters.
• The clusters are assumed to be independent of each other, but outcomes within a cluster should be allowed to be correlated.
• An example is randomly drawing fourth-grade classrooms from a large population of classrooms (say, in the state of Michigan). Each class is a cluster and the students within a class are the invididual units. Or we draw industries and then we have firms within an industry. Or we draw hospitals and then we have patients within a hospital.
• If higher-level explanatory variables are included in a model, we should consider the data as a cluster sample at the higher level to ensure valid inference.
The linear model with an additive error is

\[ y_{gm} = \alpha + x_g \beta + z_{gm} \gamma + \nu_{gm} \]  

(1.1)

for \( m = 1, \ldots, M_g, g = 1, \ldots, G \).

• The observations are independent across \( g \) (group or cluster).
Key questions:
(1) Are we primarily interested in $\mathbf{\beta}$ (group-level coefficients) or $\mathbf{\gamma}$ (individual-level coefficients)?
(2) Does $\nu_{gm}$ contain a common group effect, as in

$$\nu_{gm} = c_g + u_{gm}, m = 1, \ldots, M_g,$$

where $c_g$ is an unobserved group (cluster) effect and $u_{gm}$ is the idiosyncratic component? (Act as if it does.)
(3) Are the regressors \((x_g, z_{gm})\) appropriately exogenous?

(4) How big are the group sizes \((M_g)\) and number of groups \((G)\)? For now, we are assuming “large” \(G\) and “small” \(M_g\), but we cannot give specific values.
• The theory with $G \to \infty$ and fixed group sizes, $M_g$, is well developed [White (1984), Arellano (1987)].

• How should one use the theory? If we assume

$$E(v_{gm}|x_g, z_{gm}) = 0$$

then pooled OLS estimator of $y_{gm}$ on

$1, x_g, z_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G$, is consistent for $\lambda \equiv (\alpha, \beta', \gamma')'$

(as $G \to \infty$ with $M_g$ fixed) and $\sqrt{G}$-asymptotically normal.
• Robust variance matrix is needed to account for correlation within clusters or heteroskedasticity in \( Var(v_{gm}|x_g, z_{gm}) \), or both. Write \( W_g \) as the \( M_g \times (1 + K + L) \) matrix of all regressors for group \( g \). Then the \( (1 + K + L) \times (1 + K + L) \) variance matrix estimator is

\[
\left( \sum_{g=1}^{G} W_g' W_g \right)^{-1} \left( \sum_{g=1}^{G} W_g' \hat{v}_g \hat{v}_g' W_g \right) \left( \sum_{g=1}^{G} W_g' W_g \right)^{-1},
\]

(1.4)

where \( \hat{v}_g \) is the \( M_g \times 1 \) vector of pooled OLS residuals for group \( g \). This “sandwich” estimator is now computed routinely using “cluster” options.

• Sometimes an adjustment is made, such as multiplying by \( G/(G - 1) \).
In Stata, used “cluster” option with standard regression command:
\[ \text{reg } y \ x1 \ldots \ xK \ z1 \ldots \ zL, \text{ cluster(clusterid)} \]

• These standard errors are, as in the panel data case, robust to unknown heteroskedasticity, too.

• Structure of asymptotic variance is identical to panel data case (because \( G \to \infty \) plays the role of \( N \to \infty \) and \( M_g \) fixed is like fixed \( T \) in panel data case).

• Cluster samples are usually “unbalanced,” that is, the \( M_g \) vary across \( g \).
• Generalized Least Squares: Strengthen the exogeneity assumption to
\[ E(v_{gm}|x_g, Z_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \]  
(1.5)
where \( Z_g \) is the \( M_g \times L \) matrix of unit-specific covariates. Condition (1.5) is “strict exogeneity” for cluster samples (without a time dimension).

• If \( z_{gm} \) includes only unit-specific variables, (1.5) rules out “peer effects.” But one can include measures of peers in \( z_{gm} \) – for example, the fraction of friends living in poverty or living with only one parent.
• Full RE approach: the $M_g \times M_g$ variance-covariance matrix of $v_g = (v_{g1}, v_{g2}, \ldots, v_{gM_g})'$ has the “random effects” form,

$$Var(v_g) = \sigma_c^2 j_{M_g}' j_{M_g} + \sigma_u^2 I_{M_g},$$  \hspace{1cm} (1.6)

where $j_{M_g}$ is the $M_g \times 1$ vector of ones and $I_{M_g}$ is the $M_g \times M_g$ identity matrix.
• The usual assumptions include the “system homoskedasticity” assumption,

\[ \text{Var}(v_g | x_g, Z_g) = \text{Var}(v_g). \]  

(1.7)

• The random effects estimator \( \hat{\lambda}_{RE} \) is asymptotically more efficient than pooled OLS under (1.5), (1.6), and (1.7) as \( G \to \infty \) with the \( M_g \) fixed. The RE estimates and test statistics for cluster samples are computed routinely by popular software packages (sometimes by making it look like a panel data set).
- An important point is often overlooked: one can, and in many cases should, make RE inference completely robust to an unknown form of $\text{Var}(v_g|x_g, Z_g)$ even in the cluster sampling case.
- The motivation for using the usual RE estimator when $\text{Var}(v_g|x_g, Z_g)$ does not have the RE structure is the same as that for GEE: the RE estimator may be more efficient than POLS.
Example: Random coefficient model,

\[ y_{gm} = \alpha + x_g \beta + z_{gm} \gamma_g + v_{gm}. \]  \hspace{1cm} (1.8)

By estimating a standard random effects model that assumes common slopes \(\gamma\), we effectively include \(z_{gm}(\gamma_g - \gamma)\) in the idiosyncratic error:

\[ y_{gm} = \alpha + x_g \beta + z_{gm} \gamma + c_g + [u_{gm} + z_{gm}(\gamma_g - \gamma)] \]

The usual RE transformation does not remove the correlation across errors due to \(z_{gm}(\gamma_g - \gamma)\), and the conditional correlation depends on \(Z_g\) in general.
• If only $\gamma$ is of interest, fixed effects is attractive. Namely, apply pooled OLS to the equation with group means removed:

$$y_{gm} - \bar{y}_g = (z_{gm} - \bar{z}_g)\gamma + u_{gm} - \bar{u}_g. \quad (1.9)$$

• FE allows arbitrary correlation between $c_g$ and

$$\{z_{gm} : m = 1, \ldots, M_g\}.$$
• Can be important to allow $Var(u_g|Z_g)$ to have arbitrary form, including within-group correlation and heteroskedasticity. Using the argument for the panel data case, FE can consistently estimate the average effect in the random coefficient case. But $(z_{gm} - \bar{z}_g)(\gamma_g - \gamma)$ appears in the error term:

$$y_{gm} - \bar{y}_g = (z_{gm} - \bar{z}_g)\gamma + (u_{gm} - \bar{u}_g) + (z_{gm} - \bar{z}_g)(\gamma_g - \gamma)$$
• A fully robust variance matrix estimator of $\hat{\gamma}_{FE}$ is

$$
\left( \sum_{g=1}^{G} \bar{Z}_g \bar{Z}_g \right)^{-1} \left( \sum_{g=1}^{G} \bar{Z}_g \hat{u}_g \hat{u}_g' \bar{Z}_g \right) \left( \sum_{g=1}^{G} \bar{Z}_g \bar{Z}_g \right)^{-1},
$$

where $\bar{Z}_g$ is the matrix of within-group deviations from means and $\hat{u}_g$ is the $M_g \times 1$ vector of fixed effects residuals. This estimator is justified with large-$G$ asymptotics.
• Can also use pooled OLS or RE on

\[ y_{gm} = \alpha + x_g \beta + z_{gm} \gamma + \bar{z}_g \xi + e_{gm}, \]  

(1.11)

which allows inclusion of \( x_g \) and a simple test of \( H_0 : \xi = 0 \). Again, fully robust inference.

• POLS and RE of (1.11) both give the FE estimate of \( \gamma \).

• Example: Estimating the Salary-Benefits Tradeoff for Elementary School Teachers in Michigan.

• Clusters are school districts. Units are schools within a district.
. des

Contains data from C:\mitbook1_2e\statafiles\benefits.dta
obs: 1,848
vars: 18
size: 155,232 (99.9% of memory free)

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>display</th>
<th>value</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>distid</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>district identifier</td>
</tr>
<tr>
<td>schid</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>school identifier</td>
</tr>
<tr>
<td>lunch</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>percent eligible, free lunch</td>
</tr>
<tr>
<td>enroll</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>school enrollment</td>
</tr>
<tr>
<td>staff</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>staff per 1000 students</td>
</tr>
<tr>
<td>expppp</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>expenditures per pupil</td>
</tr>
<tr>
<td>avgsal</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>average teacher salary, $</td>
</tr>
<tr>
<td>avgben</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>average teacher non-salary benefits, $</td>
</tr>
<tr>
<td>math4</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>percent passing 4th grade math test</td>
</tr>
<tr>
<td>story4</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>percent passing 4th grade reading test</td>
</tr>
<tr>
<td>bs</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>avgben/avgsal</td>
</tr>
<tr>
<td>lavgsal</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>log(avgsal)</td>
</tr>
<tr>
<td>lenroll</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>log(enroll)</td>
</tr>
<tr>
<td>lstaff</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>log(staff)</td>
</tr>
</tbody>
</table>

Sorted by: distid schid
. reg lavgsal bs lstaff lenroll lunch

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 1848</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>48.3485452</td>
<td>4</td>
<td>12.0871363</td>
<td>F( 4, 1843) = 429.78</td>
</tr>
<tr>
<td>Residual</td>
<td>51.8328336</td>
<td>1843</td>
<td>0.028124164</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>100.181379</td>
<td>1847</td>
<td>0.054240054</td>
<td>R-squared = 0.4826</td>
</tr>
</tbody>
</table>

| ____________________________ | ____________________________ | ____________________________ | ____________________________ | ____________________________ |
| lavgsal | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
| bs | -.1774396 | .1219691 | -1.45 | 0.146 | -.4166518 .0617725 |
| lstaff | -.6907025 | .0184598 | -37.42 | 0.000 | -.7269068 -.6544981 |
| lenroll | -.0292406 | .0084997 | -3.44 | 0.001 | -.0459107 -.0125705 |
| lunch | -.0008471 | .0001625 | -5.21 | 0.000 | -.0011658 -.0005284 |
| _cons | 13.72361 | .1121095 | 122.41 | 0.000 | 13.50374 13.94349 |
. reg lavgsal bs lstaff lenroll lunch, cluster(distid)  
(Std. Err. adjusted for 537 clusters in distid)

| Robust |     Coef. | Std. Err. |   t   |    P>|t| | [95% Conf. Interval] |
|--------|-----------|-----------|-------|---------|----------------------|
| bs     | -.1774396 | .2596214  | -0.68 | 0.495   | -.6874398 .3325605   |
| lstaff | -.6907025 | .0352962  | -19.57| 0.000   | -.7600383 -.6213666  |
| lenroll| -.0292406 | .0257414  | -1.14 | 0.256   | -.079807 .0213258    |
| lunch  | -.0008471 | .0005709  | -1.48 | 0.138   | -.0019686 .0002744   |
|        | 13.72361  | .2562909  | 53.55 | 0.000   | 13.22016 14.22707    |

. reg lavgsal bs, cluster(distid)  
Linear regression  
Number of obs = 1848  
F(1, 536) = 2.36  
Prob > F = 0.1251  
R-squared = 0.0049  
Root MSE = .23238  
(Std. Err. adjusted for 537 clusters in distid)

| Robust |     Coef. | Std. Err. |   t   |    P>|t| | [95% Conf. Interval] |
|--------|-----------|-----------|-------|---------|----------------------|
| bs     | -.5034597 | .3277449  | -1.54 | 0.125   | -1.147282 .1403623   |
|        | 10.64757  | .1056538  | 100.78| 0.000   | 10.44003 10.85512    |
. xtreg lavgsal bs lstaff lenroll lunch, re

Random-effects GLS regression
Number of obs = 1848
Group variable: distid
Number of groups = 537

R-sq: within = 0.5453
    between = 0.3852
    overall = 0.4671

Obs per group:
    min = 1
    avg = 3.4
    max = 162

Random effects u_i ~ Gaussian
Wald chi2(4) = 1890.56
corr(u_i, X) = 0 (assumed)
Prob > chi2 = 0.0000

|       | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|-------|--------|-----------|-------|------|---------------------|
| lavgsal |       |           |       |      |                     |
| bs     | -.3812698 | .1118678 | -3.41 | 0.001 | -.6005267           -.162013 |
| lstaff | -.6174177  | .0153587  | -40.20 | 0.000 | -.6475202           -.5873151 |
| lenroll| -.0249189  | .0075532  | -3.30  | 0.001 | -.0397228           -.0101149 |
| lunch  | .0002995   | .0001794  | 1.67   | 0.095 | -.0000521           .0006511 |
| _cons  | 13.36682   | .0975734  | 136.99 | 0.000 | 13.17558            13.55806 |

 sigma_u |  .12627558
 sigma_e |  .09996638
   rho   |  .61473634 (fraction of variance due to u_i)
```
.xtreg lavgsal bs lstaff lenroll lunch, re cluster(distid)

Random-effects GLS regression
Group variable: distid
Number of obs = 1848
Number of groups = 537

R-sq: within = 0.5453
      between = 0.3852
      overall = 0.4671

Obs per group: min = 1
               avg = 3.4
               max = 162

Random effects u_i ~Gaussian
      corr(u_i, X) = 0 (assumed)

Wald chi2(4) = 316.91
Prob > chi2 = 0.0000

(Std. Err. adjusted for 537 clusters in distid)

|                  | Coef.    | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|------------------|----------|-----------|-------|------|---------------------|
|                  | lavgsal  |           |       |      |                     |
|                  | bs       | -.3812698 | .1504893 | -2.53 | 0.011 | -.6762235 -.0863162 |
|                  | lstaff   | -.6174177 | .0363789 | -16.97 | 0.000 | -.688719 -.5461163 |
|                  | lenroll  | -.0249189 | .0115371 | -2.16 | 0.031 | -.0475312 -.0023065 |
|                  | lunch    | .0002995  | .0001963 | -1.53 | 0.127 | -.0000852 .0006841 |
|                  | _cons    | 13.36682  | .1968713 | 67.90 | 0.000 | 12.98096 13.75268  |

|                  | sigma_u  | .12627558 |
|                  | sigma_e  | .09996638 |
|                  | rho      | .61473634  (fraction of variance due to u_i) |
```

24
. xtreg lavgsal bs lstaff lenroll lunch, fe

Fixed-effects (within) regression       Number of obs      =       1848
Group variable: distid                  Number of groups   =       537

R-sq: within = 0.5486                   Obs per group: min =           1
          between = 0.3544                 avg =     3.4
          overall = 0.4567                max =    162

corr(u_i, Xb) = 0.1433

------------------------------------------------------------------------------
lavgsal | Coef.  Std. Err.     t    P>|t|     [95% Conf. Interval]
------------------------------------------------------------------------------
     bs  | -.4948449   .133039  -3.72  0.000  -.7558382   -.2338515
     lstaff | -.6218901   .0167565 -37.11  0.000  -.6547627  -.5890175
     lenroll | -.0515063   .0094004  -5.48  0.000  -.0699478  -.0330648
     lunch |  .0005138   .0002088   2.46  0.014   .0001042   .0009234
     _cons |  13.61783   .1133406  120.15  0.000   13.39548   13.84018
------------------------------------------------------------------------------
sigma_u |  .15491886
sigma_e |  .09996638
          rho      |  .70602068  (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:  F(536, 1307) =   7.24       Prob > F = 0.0000

25
xtreg lavgsal bs lstaff lenroll lunch, fe cluster(distid)

Fixed-effects (within) regression
Number of obs = 1848
Group variable: distid Number of groups = 537

R-sq: within = 0.5486
between = 0.3544
overall = 0.4567

Obs per group: min = 1
avg = 3.4
max = 162

F(4,536) = 57.84
Prob > F = 0.0000

(Std. Err. adjusted for 537 clusters in distid)

|     | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----|-------|-----------|-------|------|----------------------|
| bs  | -.4948449 | .1937316 | -2.55 | 0.011 | -.8754112 -.1142785 |
| lstaff | -.6218901 | .0431812 | -14.40 | 0.000 | -.7067152 -.5370649 |
| lenroll | -.0515063 | .0130887 | -3.94 | 0.000 | -.0772178 -.0257948 |
| lunch | .0005138 | .0002127 | 2.42  | 0.016 | .0000959 .0009317  |
| _cons | 13.61783 | .2413169 | 56.43 | 0.000 | 13.14379 14.09187 |

.sigma_u | .15491886
.sigma_e | .09996638
.rho | .70602068 (fraction of variance due to u_i)
. xtreg lavgsal bs lstaff lenroll lunch, re cluster(distid) theta

Random-effects GLS regression                         Number of obs  =  1848
Group variable: distid                               Number of groups =  537

Random effects u_i ~Gaussian                       Wald chi2(4)   =  316.91
corr(u_i, X) = 0 (assumed)                           Prob > chi2    =  0.0000

----------------------- theta -----------------------
         min      5%    median      95%      max
0.3793    0.3793    0.3793    0.7572    0.9379

(Std. Err. adjusted for 537 clusters in distid)

|                  | Coef.     | Std. Err. | z       | P>|z|    | [95% Conf. Interval] |
|------------------|-----------|-----------|---------|--------|----------------------|
| lavgsal bs       | -0.3812698| 0.1504893 | -2.53   | 0.011  | -0.6762235          | -.0863162          |
| lstaff           | -0.6174177| 0.0363789 | -16.97  | 0.000  | -0.688719           | -.5461163          |
| lenroll          | -0.0249189| 0.0115371 | -2.16   | 0.031  | -0.0475312          | -.0023065          |
| lunch            | 0.0002995 | 0.0001963 | 1.53    | 0.127  | -0.0000852          | .0006841           |
| _cons            | 13.36682  | 1.968713  | 67.90   | 0.000  | 12.98096            | 13.75268           |

sigma_u | .12627558
sigma_e | .09996638
rho     | .61473634 (fraction of variance due to u_i)
. * Create within-district means of all covariates.

. egen bsbar = mean(bs), by(distid)
. egen lstaffbar = mean(lstaff), by(distid)
. egen lenrollbar = mean(lenroll), by(distid)
. egen lunchbar = mean(lunch), by(distid)
. xtreg lavgsal bs lstaff lenroll lunch bsbar lstaffbar lenrollbar lunchbar, 
    re cluster(distid)

Random-effects GLS regression                              Number of obs =       1848
Group variable: distid                                      Number of groups =     537

(Std. Err. adjusted for 537 clusters in distid)

| Coef.  | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|--------|-----------|-------|-------|----------------------|
| lavgsal|          |       |       |                      |
| bs     | -.4948449 | .1939422 | -2.55 | 0.011               | -.8749646   -.1147252 |
| lstaff | -.6218901 | .0432281 | -14.39| 0.000               | -.7066157   -.5371645 |
| lenroll| -.0515063 | .013103  | -3.93 | 0.000               | -.0771876   -.025825  |
| lunch  | .0005138  | .000213  | 2.41  | 0.016               | .0000964    .0009312 |
| bsbar  | .2998553  | .3031961 | 0.99  | 0.323               | -.2943981   .8941088  |
| lstaffbar| -.0255493| .0651932 | -0.39 | 0.695               | -.1533256   .1022269  |
| lenrollbar| .0657285| .020655  | 3.18  | 0.001               | .0252455    .1062116  |
| lunchbar| -.0007259| .0004378 | -1.66 | 0.097               | -.0015839   .0001322  |
| _cons  | 13.22003  | .2556139 | 51.72 | 0.000               | 12.71904    13.72103   |

|         |          |       |       |                      |
| sigma_u | .12627558|       |       |                      |
| sigma_e | .09996638|       |       |                      |
| rho     | .61473633|       |       | (fraction of variance due to u_i) |

29
. test bsbar lstaffbar lenrollbar lunchbar

( 1) bsbar = 0
( 2) lstaffbar = 0
( 3) lenrollbar = 0
( 4) lunchbar = 0

    chi2( 4) =  20.70
Prob > chi2 =  0.0004
2. Cluster-Robust Inference with Large Group Sizes

• What if one applies robust inference when the fixed $M_g$, $G \to \infty$ asymptotic analysis is not realistic? If the $M_g$ are “large” along with $G$, valid inference is still possible.

• Hansen (2007, Theorem 2, *Journal of Econometrics*) shows that with $G$ and $M_g$ both getting large the usual inference based on the robust “sandwich” estimator is valid with arbitrary correlation among the errors, $\nu_{gm}$, within each group. (Independence across groups is maintained.)
• For example, if we have a sample of $G = 100$ schools and roughly $M_g = 100$ students per school cluster-robust inference for pooled OLS should produce inference of roughly the correct size.
• Unfortunately, in the presence of cluster effects with a small number of groups \((G)\) and large group sizes \((M_g)\), cluster-robust inference with pooled OLS falls outside Hansen’s theoretical findings. We should not expect good properties of the cluster-robust inference with small groups and large group sizes.
• Example: Suppose $G = 10$ hospitals have been sampled with several hundred patients per hospital. If the explanatory variable of interest varies only at the hospital level, tempting to use pooled OLS with cluster-robust inference. But we have no theoretical justification for doing so, and reasons to expect it will not work well.
• If the explanatory variables of interest vary within group, FE is attractive. First, allows $c_g$ to be arbitrarily correlated with the $z_{gm}$. Second, with large $M_g$, can treat the $c_g$ as parameters to estimate – because we can estimate them precisely – and then assume that the observations are independent across $m$ (as well as $g$). This means that the usual inference is valid, perhaps with adjustment for heteroskedasticity.
• For panel data applications, Hansen’s (2007) results, particularly Theorem 3, imply that cluster-robust inference for the fixed effects estimator should work well when the cross section \((N)\) and time series \((T)\) dimensions are similar and not too small. If full time effects are allowed in addition to unit-specific fixed effects – as they often should – then the asymptotics must be with \(N\) and \(T\) both getting large.
• Any serial dependence in the idiosyncratic errors is assumed to be weakly dependent. Simulations in Bertrand, Duflo, and Mullainathan (2004) and Hansen (2007) verify that the robust cluster-robust variance matrix works well when $N$ and $T$ are about 50 and the idiosyncratic errors follow a stable AR(1) model.
3. Cluster Samples with Unit-Specific Panel Data

- Often, cluster samples come with a time component, so that there are two potential sources of correlation across observations: across time within the same individual and across individuals within the same group.
- Assume here that there is a natural nesting. Each unit belongs to a cluster and the cluster identification does not change over time.
- For example, we might have annual panel data at the firm level, and each firm belongs to the same industry (cluster) for all years. Or, we have panel data for schools that each belong to a district.
• Special case of **hierarchical linear model (HLM)** setup or **mixed models** or **multilevel models**.

• Now we have three data subscripts on at least some variables that we observe. For example, the response variable is $y_{gmt}$, where $g$ indexes the group or cluster, $m$ is the unit within the group, and $t$ is the time index.

• Assume we have a balanced panel with the time periods running from $t = 1, \ldots, T$. (Unbalanced case not difficult, assuming exogenous selection.) Within cluster $g$ there are $M_g$ units, and we have sampled $G$ clusters. (In the HLM literature, $g$ is usually called the *first level* and $m$ the *second level.*)
• We assume that we have many groups, $G$, and relatively few members of the group. Asymptotics: fixed $M_g$ and $T$ fixed with $G$ getting large. For example, if we can sample, say, several hundred school districts, with a few to maybe a few dozen schools per district, over a handful of years, then we have a data set that can be analyzed in the current framework.
A standard linear model with constant slopes can be written, for \( t = 1, \ldots, T, m = 1, \ldots, M_g \), and a random draw \( g \) from the population of clusters as

\[
y_{gmt} = \eta_t + w_g \alpha + x_{gm} \beta + z_{gmt} \delta + h_g + c_{gm} + u_{gmt},
\]

where, say, \( h_g \) is the industry or district effect, \( c_{gm} \) is the firm effect or school effect (firm or school \( m \) in industry or district \( g \)), and \( u_{gmt} \) is the idiosyncratic effect. In other words, the composite error is

\[
\nu_{gmt} = h_g + c_{gm} + u_{gmt}.
\]
• Generally, the model can include variables that change at any level.
• Some elements of $z_{gmt}$ might change only across $g$ and $t$, and not by unit. This is an important special case for policy analysis where the policy applies at the group level but changes over time.
• With the presence of $w_g$, or variables that change across $g$ and $t$, need to recognize $h_g$. 
• If assume the error $v_{gmt}$ is uncorrelated with $(w_g, x_{gm}, z_{gmt})$, pooled OLS is simple and attractive. Consistent as $G \to \infty$ for any cluster or serial correlation pattern.

• The most general inference for pooled OLS – still maintaining independence across clusters – is to allow any kind of serial correlation across units or time, or both, within a cluster.
• In Stata:
  \[\text{reg } y \ w_1 \ ... \ w_J \ x_1 \ ... \ x_K \ z_1 \ ... \ z_L,\]
  \[\text{cluster}(\text{industryid})\]

• Compare with inference robust only to serial correlation:
  \[\text{reg } y \ w_1 \ ... \ w_J \ x_1 \ ... \ x_K \ z_1 \ ... \ z_L,\]
  \[\text{cluster}(\text{firmid})\]

• In the context of cluster sampling with panel data, the latter is no
  longer “fully robust” because it ignores possible within-cluster
  correlation.
• Can apply a generalized least squares analysis that makes assumptions about the components of the composite error. Typically, assume components are pairwise uncorrelated, the $c_{gm}$ are uncorrelated within cluster (with common variance), and the $u_{gmt}$ are uncorrelated within cluster and across time (with common variance).

• Resulting feasible GLS estimator is an extension of the usual random effects estimator for panel data.

• Because of the large-$G$ setting, the estimator is consistent and asymptotically normal whether or not the actual variance structure we use in estimation is the proper one.
• To guard against heteroskedasticity in any of the errors and serial correlation in the \( \{u_{gmt}\} \), one should use fully robust inference that does not rely on the form of the unconditional variance matrix (which may also differ from the conditional variance matrix).

• Simpler strategy: apply random effects at the individual level, effectively ignoring the clusters in estimation. In other words, treat the data as a standard panel data set in estimation and apply usual RE. To account for the cluster sampling in inference, one computes a fully robust variance matrix estimator for the usual random effects estimator.
• In Stata:
  
  \texttt{xtset firmid year}
  
  \texttt{xtreg y w1 \ldots wJ x1 \ldots xK z1 \ldots zL, re}
  
  \texttt{cluster(industryid)}

• Again, compare with inference robust only to neglected serial correlation:
  
  \texttt{xtreg y w1 \ldots wJ x1 \ldots xK z1 \ldots zL, re}
  
  \texttt{cluster(firmid)}
Formal analysis. Write the equation for each cluster as

$$y_g = R_g \theta + v_g$$

where a row of $R_g$ is $(1, d2, \ldots, dT, w_g, x_{gm}, z_{gmt})$ (which includes a full set of period dummies) and $\theta$ is the vector of all regression parameters. For cluster $g$, $y_g$ contains $M g T$ elements ($T$ periods for each unit $m$).
• In particular,

\[
y_g = \begin{pmatrix}
y_{g1} \\
y_{g2} \\
\vdots \\
y_{gM_g}
\end{pmatrix}, \quad y_{gm} = \begin{pmatrix}
y_{gm1} \\
y_{gm2} \\
\vdots \\
y_{gmT}
\end{pmatrix}
\]

so that each \( y_{gm} \) is \( T \times 1 \); \( \mathbf{v}_g \) has an identical structure. Now, we can obtain \( \Omega_g = \text{Var}(\mathbf{v}_g) \) under various assumptions and apply feasible GLS.
• RE at the unit level is obtained by choosing $\Omega_g = I_{Mg} \otimes \Lambda$, where $\Lambda$ is the $T \times T$ matrix with the RE structure. If there is within-cluster correlation, this is not the correct form of $Var(v_g)$, and that is why robust inference is generally needed after RE estimation.
• For the case that $v_{gmt} = h_g + c_{gm} + u_{gmt}$ where the terms have variances $\sigma_h^2$, $\sigma_c^2$, and $\sigma_u^2$, respectively, they are pairwise uncorrelated, $c_{gm}$ and $c_{gr}$ are uncorrelated for $r \neq m$, and $\{u_{gmt} : t = 1, \ldots, T\}$ is serially uncorrelated, we can obtain $\Omega_g$ as follows:

\begin{align*}
    Var(v_{gm}) &= (\sigma_h^2 + \sigma_c^2)j_Tj_T' + \sigma_u^2I_T \\
    Cov(v_{gm}, v_{gr}) &= \sigma_h^2j_Tj_T', \quad r \neq m
\end{align*}

\[
    \Omega_g = \begin{pmatrix}
        (\sigma_h^2 + \sigma_c^2)j_Tj_T' + \sigma_u^2I_T & \cdots & \sigma_h^2j_Tj_T' \\
        \vdots & \ddots & \vdots \\
        \sigma_h^2j_Tj_T' & \cdots & (\sigma_h^2 + \sigma_c^2)j_Tj_T' + \sigma_u^2I_T
    \end{pmatrix}
\]
• The robust asymptotic variance of $\hat{\theta}$ is estimated as

$$\widehat{A\text{var}}(\hat{\theta}) = \left( \sum_{g=1}^{G} R_g \hat{\Omega}_g^{-1} R_g \right)^{-1} \left( \sum_{g=1}^{G} R_g \hat{\Omega}_g^{-1} \hat{v}_g \hat{v}_g' \hat{\Omega}_g^{-1} R_g \right)^{-1} \cdot \left( \sum_{g=1}^{G} R_g \hat{\Omega}_g^{-1} R_g \right)^{-1},$$

where $\hat{v}_g = y_g - R_g \hat{\theta}$. 
Unfortunately, routines intended for estimating HLMs (or mixed models) assume that the structure imposed on $\Omega_g$ is correct, and that $Var(v_g|R_g) = Var(v_g)$. The resulting inference could be misleading, especially if serial correlation in $\{u_{gmt}\}$ is not allowed.

In Stata, the command is `xtmixed`. 
• Because of the nested data structure, we have available different versions of fixed effects estimators. Subtracting cluster averages from all observations within a cluster eliminates $h_g$; when $w_{gt} = w_g$ for all $t$, $w_g$ is also eliminated. But the unit-specific effects, $c_{mg}$, are still part of the error term. If we are mainly interested in $\delta$, the coefficients on the time-varying variables $z_{gmt}$, then removing $c_{gm}$ (along with $h_g$) is attractive. In other words, use a standard fixed effects analysis at the individual level.
• If the units are allowed to change groups over time – such as children changing schools – then we would replace $h_{g}$ with $h_{gt}$, and then subtracting off individual-specific means would not remove the time-varying cluster effects.
• Even if we use unit “fixed effects” – that is, we demean the data at the unit level – we might still use inference robust to clustering at the aggregate level. Suppose the model is

\[ y_{gmt} = \eta_t + w_g \alpha + x_{gm} \beta + z_{gmt} d_{mg} + h_g + c_{mg} + u_{gmt} \]

\[ = \eta_t + w_{gt} \alpha + x_{gm} \beta + z_{gmt} \delta + h_g + c_{mg} + u_{gmt} + z_{gmt} e_{gm}, \]

where \( d_{gm} = \delta + e_{gm} \) is a set of unit-specific intercepts on the individual, time-varying covariates \( z_{gmt} \).
• The time-demeaned equation within individual $m$ in cluster $g$ is

$$y_{gmt} - \bar{y}_{gm} = \zeta_t + (z_{gmt} - \bar{z}_{gm})\delta + (u_{gmt} - \bar{u}_{gm}) + (z_{gmt} - \bar{z}_{gm})e_{gm}.$$  

• FE is still consistent if $E(d_{mg}|z_{gmt} - \bar{z}_{gm}) = E(d_{mg}), m = 1, \ldots, M_g, t = 1, \ldots, T,$ and all $g$, and so cluster-robust inference, which is automatically robust to serial correlation and heteroskedasticity, makes perfectly good sense.
**Example:** Effects of Funding on Student Performance

```
. use meap94_98

. des

Contains data from meap94_98.dta
obs: 7,150
vars: 26
size: 893,750 (99.8% of memory free)

```
<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>display</th>
<th>value</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>distid</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>district identifier</td>
</tr>
<tr>
<td>schid</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>school identifier</td>
</tr>
<tr>
<td>lunch</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>% eligible for free lunch</td>
</tr>
<tr>
<td>enrol</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>number of students</td>
</tr>
<tr>
<td>exppp</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>expenditure per pupil</td>
</tr>
<tr>
<td>math4</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>% satisfactory, 4th grade math test</td>
</tr>
<tr>
<td>year</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td>1992=school yr 1991-2</td>
</tr>
<tr>
<td>cpi</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>consumer price index</td>
</tr>
<tr>
<td>rexppp</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>(exppp/cpi)*1.695: 1997 $</td>
</tr>
<tr>
<td>lrexpp</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>log(rexpp)</td>
</tr>
<tr>
<td>lenrol</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>log(enrol)</td>
</tr>
<tr>
<td>avgrexp</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>(rexppp + rexppp_1)/2</td>
</tr>
<tr>
<td>lavgrexp</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>log(avgrexp)</td>
</tr>
<tr>
<td>tobs</td>
<td>byte</td>
<td>%9.0g</td>
<td></td>
<td>number of time periods</td>
</tr>
</tbody>
</table>

Sorted by: schid  year
. * egen tobs = sum(1), by(schid)

. tab tobs if y98

<table>
<thead>
<tr>
<th>number of time periods</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>487</td>
<td>29.28</td>
<td>29.28</td>
</tr>
<tr>
<td>4</td>
<td>254</td>
<td>15.27</td>
<td>44.56</td>
</tr>
<tr>
<td>5</td>
<td>922</td>
<td>55.44</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>1,663</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
. xtreg math4 lavgrexp lunch lenrol y95-y98, fe

Fixed-effects (within) regression

Group variable: schid

Number of obs = 7150
Number of groups = 1683

R-sq: within = 0.3602
between = 0.0292
overall = 0.1514

|     | Coef.     | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|-----|-----------|-----------|------|------|---------------------|
| math4 | 6.288376  | 2.098685  | 3.00 | 0.003 | 2.174117 10.40264  |
| lavgrexp | -2.038461 | 1.791604  | -1.14| 0.255 | -5.550718 1.473797 |
| lunch | -.0215072 | .0312185  | -0.69| 0.491 | -.082708 .0396935 |
| lenrol | 11.6192   | .5545233  | 20.95| 0.000 | 10.53212 12.70629 |
| y95 | 13.05561  | .6630948  | 19.69| 0.000 | 11.75568 14.35554 |
| y96 | 10.14771  | .7024067  | 14.45| 0.000 | 8.770713 11.52471 |
| y97 | 23.41404  | .7187237  | 32.58| 0.000 | 22.00506 24.82303 |
|_cons | 11.84422  | 22.81097  | 0.52 | 0.604 | -32.87436 56.5628 |

sigma_u | 15.84958
sigma_e | 11.325028
rho | .66200804  (fraction of variance due to u_i)

F test that all u_i=0:  F(1682, 5460) = 4.82  Prob > F = 0.0000
. xtreg math4 lavgrexp lunch lenrol y95-y98, fe cluster(schid)

Fixed-effects (within) regression
Number of obs = 7150
Group variable: schid Number of groups = 1683

(Std. Err. adjusted for 1683 clusters in schid)

|          | Coef.    | Robust Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|----------|----------|------------------|-------|------|----------------------|
| lavgrexp| 6.288376 | 2.431317         | 2.59  | 0.010| 1.519651 11.0571     |
| lunch   | -0.0215072 | 0.0390732 | -0.55 | 0.582| -.0981445 .05513     |
| lenrol  | -2.038461 | 1.789094        | -1.14 | 0.255| -5.547545 1.470623   |
| y95     | 11.6192  | 5.358469        | 21.68 | 0.000| 10.56821 12.6702     |
| y96     | 13.05561 | 6.910815        | 18.89 | 0.000| 11.70014 14.41108    |
| y97     | 10.14771 | 7.326314        | 13.85 | 0.000| 8.710745 11.58468    |
| y98     | 23.41404 | 7.699553        | 30.53 | 0.000| 21.90975 24.91833    |
| _cons   | 11.84422 | 25.16643        | 0.47  | 0.638| -37.51659 61.20503   |

| sigma_u | 15.84958 |
| sigma_e | 11.325028 |
| rho     | .66200804 (fraction of variance due to u_i) |
```
. xtreg math4 lavgrexp lunch lenrol y95-y98, fe cluster(distid)

(Std. Err. adjusted for 467 clusters in distid)

|            | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------|--------|-----------|-------|-----|----------------------|
| math4      | 6.288376 | 3.132334  | 2.01  | 0.045 | .1331271 12.44363    |
| lavgrexp   | -0.0215072 | 0.0399206 | -0.54 | 0.590 | -.0999539 .0569395  |
| lunch      | -2.038461 | 2.098607  | -0.97 | 0.332 | -6.162365 2.085443  |
| lenrol     | 11.6192  | .7210398  | 16.11 | 0.000 | 10.20231 13.0361     |
| y95        | 13.05561 | .9326851  | 14.00 | 0.000 | 11.22282 14.8884     |
| y96        | 23.41404 | 1.027313  | 22.79 | 0.000 | 21.3953 25.43278     |
| _cons      | 11.84422 | 32.68429  | 0.36  | 0.717 | -52.38262 76.07107   |

sigma_u    | 15.84958 |
sigma_e    | 11.325028 |
rho        | .66200804 (fraction of variance due to u_i) |

62
```
4. Clustering and Stratification

Survey data often characterized by clustering and VP sampling. Suppose that $g$ represents the primary sampling unit (say, city) and individuals or families (indexed by $m$) are sampled within each PSU with probability $p_{gm}$. If $\hat{\beta}$ is the pooled OLS estimator across PSUs and individuals, its variance is estimated as
If the probabilities are estimated using retention frequencies, estimate is conservative, as before.
• Multi-stage sampling schemes introduce even more complications. Let there be $S$ strata (e.g., states in the U.S.), exhaustive and mutually exclusive. Within stratum $s$, there are $C_s$ clusters (e.g., neighborhoods).

• Large-sample approximations: the number of clusters sampled, $N_s$, gets large. This allows for arbitrary correlation (say, across households) within cluster.
• Within stratum $s$ and cluster $c$, let there be $M_{sc}$ total units (household or individuals). Therefore, the total number of units in the population is

$$M = \sum_{s=1}^{S} \sum_{c=1}^{C_s} M_{sc}.$$
• Let $z$ be a variable whose mean we want to estimate. List all population values as \( \{z_{scm}^o : m = 1, \ldots, M_{sc}, c = 1, \ldots, C_s, s = 1, \ldots, S\} \), so the population mean is

\[
\mu = M^{-1} \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z_{scm}^o.
\]

Define the total in the population as

\[
\tau = \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z_{scm}^o = M\mu.
\]
Totals within each cluster and then stratum are, respectively,

\[ \tau_{sc} = \sum_{m=1}^{M_{sc}} z_{scm}^0 \]

\[ \tau_s = \sum_{c=1}^{C_s} \tau_{sc} \]

- Sampling scheme:

(i) For each stratum \( s \), randomly draw \( N_s \) clusters, with replacement. 
(Fine for “large” \( C_s \).)

(ii) For each cluster \( c \) drawn in step (i), randomly sample \( K_{sc} \) households with replacement.
• For each pair \((s, c)\), define

\[
\hat{\mu}_{sc} = K_{sc}^{-1} \sum_{m=1}^{K_{sc}} z_{scm}.
\]

Because this is a random sample within \((s, c)\),

\[
E(\hat{\mu}_{sc}) = \mu_{sc} = M_{sc}^{-1} \sum_{m=1}^{M_{sc}} z_{scm}^o.
\]
To continue up to the cluster level we need to estimate the total, \( \tau_{sc} = M_{sc} \mu_{sc} \). So, \( \hat{\tau}_{sc} = M_{sc} \hat{\mu}_{sc} \) is an unbiased estimator of \( \tau_{sc} \) for all \( \{(s, c) : c = 1, \ldots, C_s, s = 1, \ldots, S\} \). (We can think of computing this estimate even if we eventually do not use some clusters.)
• Next, consider randomly drawing $N_s$ clusters from stratum $s$. Can show that an unbiased estimator of the total $\tau_s$ for stratum $s$ is

$$C_s \cdot N_s^{-1} \sum_{c=1}^{N_s} \hat{\tau}_{sc}.$$ 

• Finally, the total in the population is estimated as

$$\sum_{s=1}^{S} \left( C_s \cdot N_s^{-1} \sum_{c=1}^{N_s} \hat{\tau}_{sc} \right) \equiv \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} z_{scm}$$

where the weight for stratum-cluster pair $(s, c)$ is

$$\omega_{sc} \equiv \frac{C_s}{N_s} \cdot \frac{M_{sc}}{K_{sc}}.$$
• Note how $\omega_{sc} = (C_s/N_s)(M_{sc}/K_{sc})$ accounts for under- or over-sampled clusters within strata and under- or over-sampled units within clusters.

• Appears in the literature on complex survey sampling, sometimes without $M_{sc}/K_{sc}$ when each cluster is sampled as a complete unit, and so $M_{sc}/K_{sc} = 1$.

• To estimate the mean $\mu$, just divide by $M$, the total number of units sampled.

$$\hat{\mu} = M^{-1} \left( \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} z_{scm} \right).$$
• To study regression, specify the problem as

$$\min_{\beta} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} (y_{scm} - x_{scm} \beta)^2.$$ 

The asymptotic variance combines clustering with weighting to account for the multi-stage sampling. Following Bhattacharya (2005, Journal of Econometrics), an appropriate asymptotic variance estimate has a sandwich form,

$$\left( \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} x'_{scm} x_{scm} \right)^{-1} \hat{\Sigma} \left( \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} x'_{scm} x_{scm} \right)^{-1}.$$
\( \hat{\mathbf{B}} \) is somewhat complicated:

\[
\hat{\mathbf{B}} = \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc}^2 \hat{u}_{scm}^2 \mathbf{x}_{scm}' \mathbf{x}_{scm} \\
+ \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \sum_{r \neq m}^{K_{sc}} \omega_{sc}^2 \hat{u}_{scm} \hat{u}_{scr} \mathbf{x}_{scm}' \mathbf{x}_{scr} \\
- \sum_{s=1}^{S} N_s^{-1} \left( \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}_{scm}' \hat{u}_{scm} \right) \left( \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}_{scm}' \hat{u}_{scm} \right)' 
\]
• The first part of $\hat{\Phi}$ is obtained using the White “heteroskedasticity”-robust form. The second piece accounts for the clustering. The third piece reduces the variance by accounting for the nonzero means of the “score” within strata.
• Suppose that the population is stratified by region, taking on values 1 through 8, and the primary sampling unit is zip code. Within each zip code we obtain a sample of families, possibly using VP sampling.

• Stata command:

```stata
svyset zipcode [pweight = sampwght],
strata(region)
```

• Now we can use a set of econometric commands. For example,

```stata
svy: reg y x1 ... xK
```
. use http://www.stata-press.com/data/r10/nhanes2f

. svyset psuid [pweight = finalwgt], strata(stratid)
Pweight: finalwgt
VCE: linearized
Single unit: missing
Strata 1: stratid
SU 1: psuid
FPC 1: <zero>

. tab health

1=excellent | Freq. Percent Cum.  
-------------+----------+---------+----------+----------
    poor | 729     7.05    7.05   
   fair  | 1,670   16.16   23.21   
average | 2,938   28.43   51.64   
good   | 2,591   25.07   76.71   
excellent | 2,407   23.29  100.00   
-------------+----------+---------+----------+----------
     Total | 10,335  100.00  100.00  100.00  

. sum lead

Variable |   Obs   Mean     Std. Dev.   Min   Max
----------|--------|---------|----------------|------|-----
     lead | 4942   14.32032   6.167695   2     80


. svy: oprobit health lead female black age weight
(running oprobit on estimation sample)

Survey: Ordered probit regression

|                | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----------------|-------|-----------|------|-----|----------------------|
| health         |       |           |      |     |                      |
| lead           | -.0059646 | .0045114 | -1.32 | 0.196 | -.0151656 to .0032364 |
| female         | -.1529889 | .057348  | -2.67 | 0.012 | -.2699508 to -.036027  |
| black          | -.535801  | .0622171 | -8.61 | 0.000 | -.6626937 to -.4089084 |
| age            | -.0236837 | .0011995 | -19.75| 0.000 | -.02613 to -.0212373  |
| weight         | -.0035402 | .0010954 | -3.23 | 0.003 | -.0057743 to -.0013061 |
| /cut1          | -3.278321 | .1711369 | -19.16| 0.000 | -3.627357 to -2.929285 |
| /cut2          | -2.496875 | .1571842 | -15.89| 0.000 | -2.817454 to -2.176296 |
| /cut3          | -1.611873 | .1511986 | -10.66| 0.000 | -1.920244 to -1.303501 |
| /cut4          | -.8415657 | .1488381 | -5.65 | 0.000 | -1.145123 to -.5380083 |
. oprobit health lead female black age weight

Iteration 0:  log likelihood = -7526.7772
Iteration 1:  log likelihood = -7133.9477
Iteration 2:  log likelihood = -7133.6805

Ordered probit regression

|         | Coef.    | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|---------|----------|-----------|-------|-----|-----------------------|
| health  |          |           |       |     |                       |
| lead    | -.0011088| .0026942  | -0.41 | 0.681| -.0063893 -.0041718  |
| female  | -.1039273| .0352721  | -2.95 | 0.003| -.1730594 -.0347952  |
| black   | -.4942909| .0502051  | -9.85 | 0.000| -.592691 -.3958908   |
| age     | -.0237787| .0009147  | -26.00| 0.000| -.0255715 -.0219859  |
| weight  | -.0027245| .0010558  | -2.58 | 0.010| -.0047938 -.0006551  |
| cut1    | -3.072779 | .1087758  |       |     | -3.285975 -2.859582  |
| cut2    | -2.249324 | .1057841  |       |     | -2.456657 -2.041991  |
| cut3    | -1.396732 | .1038044  |       |     | -1.600185 -1.19328   |
| cut4    | -0.6615336| .1028773  |       |     | -.8631693 -.4598978  |
5. Two-Way Clustering

- Recent interest in two-way clustering – usually across time within a firm and across firms within a given time period.
- Often the underlying model is set up as follows:

\[ y_{it} = x_{it}\beta + g_t + c_i + u_{it} \]

where \( g_t \) and \( c_i \) are both viewed as random (as is \( u_{it} \), of course).
- The presence of \( c_i \), as usual, induces correlation across time. But \( \{u_{it}\} \) can also be serially correlated across time.
• If the firm effects $c_i$ and idiosyncratic errors $\{u_{iit}\}$ are uncorrelated across $i$ – as in many data generating mechanisms – then all cross-sectional correlation is due to the presence of $g_t$, the aggregate time effects.

• Eliminating $g_t$ by the within transformation across firms in the same time period solves the problem. In practice, include time dummies along with firm dummies. Or, include time dummies and use standard FE software, as in

\[
\text{xtset firmid year} \\
\text{xi: xtreg y x1 ... xK i.year, fe cluster(firmid)}
\]
• Could there be left over cross-sectional correlation even if we use time dummies? Yes, for example if

\[ y_{it} = x_{it}b_t + g_t + c_i + u_{it} \]

but we act as if \( b_t = \beta \). [Then \( x_{it}(b_t - \beta) \) is part of the error term.] But how important is this in practice?

• Peterson (2009, Review of Financial Studies) and Gow, Ormazabal, and Taylor (2010, Accounting Review) study various standard errors, including two-way clustering, for simulated and actual data. The two-way clustering appears to work well even for \( T \) as small as 10 and \( N = 200 \).
• But some empirical examples in GOT are misleading. With very large $N$ and relatively small $T$ (15 to 30), they compare two-way clustering with clustering for cross-sectional correlation. The obvious approach is to include year dummies (which they do in some cases) and cluster for serial correlation. Firm effects may or may not be needed for consistency of the parameters, but including them would reduce the time series correlation. GOT do not include them.

• Thompson (2011, Journal of Financial Economics) shows that the two-way clustering (or double clustering) is valid provided $N$ and $T$ are both “large.” The aggregate shocks must dissipate over time.
• In simulations, two-way clustering seems to work well for $N = 50$ and $T = 25$, but there is little justification for clustering for cross-sectional correlation if, say, $N = 1,000$ and $T = 5$.
• In such scenarios, the two-way clustering actually produces standard errors that are too small if there is no time series nor cross-sectional correlation.
• With large $N$ and small $T$, it seems including time effects and clustering for serial correlation is the only theoretically justified procedure.
• Example: Airfare equation:

. * First, two-way clustering without firm effects.

. xi: cluster2 lfare concen ldist ldistsq i.year, fcluster(id) tcluster(year)
   i.year _Iyear_1997-2000 (naturally coded; _Iyear_1997 omitted)

Linear regression with 2D clustered SEs

Number of obs = 4596
F( 6, 4589) = 558.39
Prob > F = 0.0000
R-squared = 0.4062
Number of clusters (id) = 1149
Number of clusters (year) = 4
Root MSE = 0.3365

|               | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|---------------|--------|-----------|-------|------|----------------------|
| concen        | .3601203 | .0560493  | 6.43  | 0.000 | .2502368 .4700039    |
| ldist         | -.9016004 | .235178   | -3.83 | 0.000 | -1.362662 -.4405384  |
| ldistsq       | .1030196 | .0174188  | 5.91  | 0.000 | .0688704 .1371688    |
| _Iyear_1998   | .0211244 | .          | .     | .    | .                    |
| _Iyear_1999   | .0378496 | .          | .     | .    | .                    |
| _Iyear_2000   | .09987 | .          | .     | .    | .                    |
| _cons         | 6.209258 | .7956274  | 7.80  | 0.000 | 4.649445 7.76907     |

SE clustered by id and year
* Now cluster only within time period:

```
.xi: reg lfare concen ldist ldistsq i.year, cluster(year)
i.year _Iyear_1997-2000 (naturally coded; _Iyear_1997 omitted)
```

Linear regression

|          | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----------|--------|-----------|-------|------|----------------------|
| lfare    | .3601203 | .0269237  | 13.38 | .001 | .274437 .4458037     |
| concen   | -.9016004 | .0337267  | -26.73| .000 | -1.008934 .794267    |
| ldist    | .1030196 | .0024454  | 42.13 | .000 | .0952372 .110802     |
| ldistsq  | .0211244 | .0002405  | 87.84 | .000 | .020359 .0218897     |
| _Iyear_1998 | .0378496 | .0002066  | 183.19| .000 | .0371921 .0385071    |
| _Iyear_1999 | .09987 | .0002954  | 338.11| .000 | .0989299 .10081      |
| _Iyear_2000 | 6.209258 | .1539302  | 40.34 | .000 | 5.719383 6.699132    |

(Std. Err. adjusted for 4 clusters in year)

* These standard errors are much too small, illustrating the point made by
* Gow, Ormazabal, and Taylor. But these are hardly the natural standard
* errors to use.
. * Now cluster only within route to account for the substantial serial
. * correlation. Remember, the route effect is left in the error term.

. xi: reg lfare concen ldist ldistsq i.year, cluster(id)
i.year _Iyear_1997-2000 (naturally coded; _Iyear_1997 omitted)

Linear regression

|                  | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------------|--------|-----------|-------|------|----------------------|
| lfare            | .3601203 | .058556   | 6.15  | 0.000 | .2452315  .4750092 |
| concen           | -.9016004 | .2719464  | -3.32 | 0.001 | -1.435168 -.3680328 |
| ldist            | .1030196  | .0201602  | 5.11  | 0.000 | .0634647  .1425745 |
| ldistsq          | .0211244  | .0041744  | 5.09  | 0.000 | .0129871  .0292617 |
| _Iyear_1998      | .0378496  | .0051795  | 7.31  | 0.000 | .0276872  .048012 |
| _Iyear_1999      | .09987    | .0056469  | 17.69 | 0.000 | .0887906  .1109493 |
| _Iyear_2000      | 6.209258  | .9117551  | 6.81  | 0.000 | 4.420364  7.998151 |

* These are much closer to the two-way cluster standard errors; in fact,
* somewhat larger. And these have justification with large N and small T.
. What if we also use firm FEs (on a reduced sample)?

```
. xi: xtreg lfare concen 1dist ldistsq i.year, fe cluster(id)
   i.year _Iyear_1997-2000 (naturally coded; _Iyear_1997 omitted)

Fixed-effects (within) regression                       Number of obs   =    400
Group variable: id                                     Number of groups =    100
                                                      F(4,99)        =   11.52
F(4,99)                                                corr(u_i, Xb) =  -0.3263
Prob > F                                                (Std. Err. adjusted for 100 clusters in id)

------------------------------------------------------------------------------
                     Robust            Robust
  lfare |     Coef.  Std. Err.     t   P>|t|    [95% Conf. Interval]
-------------|-----------------------------------------
   concen |   0.558547   0.209726  2.66 0.009    0.142406    0.974688
(dropped)                             
   ldist  |   (dropped)                          
   ldistsq |   (dropped)                          
   _Iyear_1998 |  -0.004301  0.018578  -0.23 0.817    -0.041163    0.032562
   _Iyear_1999 |   0.032450  0.020025   1.62 0.108    -0.007288    0.072179
   _Iyear_2000 |   0.087841  0.020673   4.25 0.000     0.046821     0.128860
   _cons   |    4.67532   0.140864  33.19 0.000     4.395817     4.954826
------------------------------------------------------------------------------
   sigma_u |    0.370745
   sigma_e |    0.118667
     rho   |    0.907071 (fraction of variance due to u_i)
```

88
. xi: cluster2 lfare concen ldist ldistsq i.id i.year, fcluster(id)
   tcluster(year)
i.id  _Iid_1-100  (naturally coded; _Iid_1 omitted)
i.year  _Iyear_1997-2000  (naturally coded; _Iyear_1997 omitted)

Linear regression with 2D clustered SEs

Number of obs = 400
F(103, 296) = 229.75
Prob > F = 0.0000
R-squared = 0.9211
Number of clusters (id) = 100
Root MSE = 0.1187

|             | Coef.     | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-------------|-----------|-----------|-------|-------|----------------------|
| concen      | .5585469  | .2756093  | 2.03  | 0.044 | .0161449      1.100949 |
| ldist       | (dropped) |           |       |       |                     |
| ldistsq     | .0535631  | .0007484  | 71.57 | 0.000 | .0520902      .0550361 |
| _Iid_2      | -.2053579 | .0484883  | -4.24 | 0.000 | -.3007833   -.1099324 |
| _Iid_3      | .3550181  | .        |       |       |                     |
| ... (output suppressed) |

SE clustered by id and year

* Now the two-way standard error is quite a bit larger. But we have no theory
* telling us it is valid.
• The “cluster2” command is tied to structures such as

\[ y_{it} = x_{it} \beta + g_t + c_i + u_{it} \]

It does not allow situations where, say, shocks to firm \( h \) in year \( t - 1 \) are correlated with shocks to firm \( i \neq h \) in year \( t \).

• Thompson (2011) allows for some dependence across time for different units, but must specify the maximum lag.