CLUSTER SAMPLING

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. The Linear Model with Cluster Effects
2. Cluster-Robust Inference with Large Group Sizes
3. Cluster Samples with Unit-Specific Panel Data
4. Estimation with a Small Number of Groups
5. Clustering and Stratification

1. The Linear Model with Cluster Effects

- For each group or cluster g, let $\{(y_{gm}, \mathbf{x}_g, \mathbf{z}_{gm}) : m = 1, \ldots, M_g\}$ be the observable data, where M_g is the number of units in cluster or group g, y_{gm} is a scalar response, $\mathbf{x}_g$ is a $1 \times K$ vector containing explanatory variables that vary only at the cluster or group level, and $\mathbf{z}_{gm}$ 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 any modeling, we should consider the data as a cluster sample to ensure valid inference.

- The linear model with an additive error is

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + v_{gm} \tag{1.1}$$

for $m = 1,\ldots,M_g$, $g = 1,\ldots,G$.

- The observations are independent across g.

• Key questions:

(1) Are we primarily interested in $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$?

(2) Does v_{gm} contain a common group effect, as in

$$v_{gm} = c_g + u_{gm}, m = 1,\ldots,M_g, \tag{1.2}$$

where c_g is an unobserved group (cluster) effect and u_{gm} is the idiosyncratic component?

(3) Are the regressors $(\mathbf{x}_g, \mathbf{z}_{gm})$ 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 the group sizes, M_g, fixed is well developed [White (1984), Arellano (1987)]. How should one use these methods? If

$$E(v_{gm}|\mathbf{x}_g, \mathbf{z}_{gm}) = 0 \tag{1.3}$$

then pooled OLS estimator of y_{gm} on $1, \mathbf{x}_g, \mathbf{z}_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G,$ is consistent for $\boldsymbol{\lambda} \equiv (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}')'$ (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 $Var(v_{gm}|\mathbf{x}_g, \mathbf{z}_{gm})$, or both. Write $\mathbf{W}_g$ 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

$$\left(\sum_{g=1}^{G} \mathbf{W}_g' \mathbf{W}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{W}_g' \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \mathbf{W}_g\right) \left(\sum_{g=1}^{G} \mathbf{W}_g' \mathbf{W}_g\right)^{-1}, \tag{1.4}$$

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

• In State, used "cluster" option with standard regression command:

```
reg y x1 ... xK z1 ... zL, cluster(clusterid)
```

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

• Structure is identical to panel data case, and so is asymptotics (because $G \to \infty$ plays the role of $N \to \infty$. The fixed M_g setting is like fixed T in panel data case.)

• Cluster samples are usuall "unbalanced," that is, the M_g vary across g.

- Generalized Least Squares: Strengthen the exogeneity assumption to

$$E(v_{gm}|\mathbf{x}_g, \mathbf{Z}_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{1.5}$$

where $\mathbf{Z}_g$ 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 $\mathbf{z}_{gm}$ includes only unit-specific variables, (1.5) rules out "peer effects." But one can include measures of peers in $\mathbf{z}_{gm}$ – 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 $\mathbf{v}_g = (v_{g1}, v_{g2}, \ldots, v_{g,M_g})'$ has the “random effects” form,

$$Var(\mathbf{v}_g) = \sigma_c^2 \mathbf{j}_{M_g}' \mathbf{j}_{M_g} + \sigma_u^2 \mathbf{I}_{M_g}, \tag{1.6}$$

where $\mathbf{j}_{M_g}$ is the $M_g \times 1$ vector of ones and $\mathbf{I}_{M_g}$ is the $M_g \times M_g$ identity matrix.

• The usual assumptions include the "system homoskedasticity" assumption,

$$Var(\mathbf{v}_g|\mathbf{x}_g, \mathbf{Z}_g) = Var(\mathbf{v}_g). \tag{1.7}$$

• The random effects estimator $\hat{\boldsymbol{\lambda}}_{RE}$ is asymptotically more efficient than pooled OLS under (1.5), (1.6), and (1.7) as $G \rightarrow \infty$ 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 $Var(\mathbf{v}_g|\mathbf{x}_g, \mathbf{Z}_g)$ even in the cluster sampling case.
- The motivation for using the usual RE estimator when $Var(\mathbf{v}_g|\mathbf{x}_g, \mathbf{Z}_g)$ 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,

$$y_{gm} = \alpha + \mathbf{x}_g \boldsymbol{\beta} + \mathbf{z}_{gm} \boldsymbol{\gamma}_g + v_{gm}. \tag{1.8}$$

By estimating a standard random effects model that assumes common slopes $\boldsymbol{\gamma}$, we effectively include $\mathbf{z}_{gm}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ in the idiosyncratic error:

$$y_{gm} = \alpha + \mathbf{x}_g \boldsymbol{\beta} + \mathbf{z}_{gm} \boldsymbol{\gamma} + c_g + [u_{gm} + \mathbf{z}_{gm}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})]$$

- The usual RE transformation does not remove the correlation across errors due to $\mathbf{z}_{gm}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$, and the conditional correlation depends on $\mathbf{Z}_g$ in general.

- If only $\boldsymbol{\gamma}$ is of interest, fixed effects is attractive. Namely, apply pooled OLS to the equation with group means removed:

$$y_{gm} - \bar{y}_g = (\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)\boldsymbol{\gamma} + u_{gm} - \bar{u}_g. \tag{1.9}$$

- FE allows arbitrary correlation between c_g and $\{\mathbf{z}_{gm} : m = 1, \ldots, M_g\}$.

• Can be important to allow $Var(\mathbf{u}_g|\mathbf{Z}_g)$ 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 $(\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ appears in the error term:

$$y_{gm} - \bar{y}_g = (\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)\boldsymbol{\gamma} + (u_{gm} - \bar{u}_g) + (\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$$

- A fully robust variance matrix estimator of $\hat{\boldsymbol{\gamma}}_{FE}$ is

$$\left(\sum_{g=1}^{G}\ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}\left(\sum_{g=1}^{G}\ddot{\mathbf{Z}}_g'\hat{\ddot{\mathbf{u}}}_g\hat{\ddot{\mathbf{u}}}_g'\ddot{\mathbf{Z}}_g\right)\left(\sum_{g=1}^{G}\ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}, \tag{1.10}$$

where $\ddot{\mathbf{Z}}_g$ is the matrix of within-group deviations from means and $\hat{\ddot{\mathbf{u}}}_g$ 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

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \bar{\mathbf{z}}_g\boldsymbol{\xi} + e_{gm}, \tag{1.11}$$

which allows inclusion of $\mathbf{x}_g$ and a simple test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$. Again, fully robust inference.

- POLS and RE of (1.11) both give the FE estimate of $\boldsymbol{\gamma}$.
- 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                          15 Mar 2009 11:25
 size:       155,232 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
distid          float  %9.0g                  district identifier
schid           int    %9.0g                  school identifier
lunch           float  %9.0g                  percent eligible, free lunch
enroll          int    %9.0g                  school enrollment
staff           float  %9.0g                  staff per 1000 students
exppp           int    %9.0g                  expenditures per pupil
avgsal          float  %9.0g                  average teacher salary, $
avgben          int    %9.0g                  average teacher non-salary
                                                benefits, $
math4           float  %9.0g                  percent passing 4th grade math
                                                test
story4          float  %9.0g                  percent passing 4th grade
                                                reading test
bs              float  %9.0g                  avgben/avgsal
lavgsal         float  %9.0g                  log(avgsal)
lenroll         float  %9.0g                  log(enroll)
lstaff          float  %9.0g                  log(staff)
-------------------------------------------------------------------------------
Sorted by:  distid  schid
```

```
. reg lavgsal bs lstaff lenroll lunch

      Source |       SS       df       MS              Number of obs =    1848
-------------+------------------------------           F(  4,  1843) =  429.78
       Model |  48.3485452     4  12.0871363           Prob > F      =  0.0000
    Residual |  51.8328336  1843  .028124164           R-squared     =  0.4826
-------------+------------------------------           Adj R-squared =  0.4815
       Total |  100.181379  1847  .054240054           Root MSE      =   .1677

------------------------------------------------------------------------------
     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
     lavgsal |      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
       _cons |   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
     lavgsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bs |  -.5034597   .3277449    -1.54   0.125    -1.147282    .1403623
       _cons |   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                          Obs per group: min =         1
       between = 0.3852                                         avg =       3.4
       overall = 0.4671                                         max =       162

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

------------------------------------------------------------------------------
     lavgsal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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                   Number of obs      =      1848
Group variable: distid                          Number of groups   =       537

R-sq:  within  = 0.5453                         Obs per group: min =         1
       between = 0.3852                                        avg =       3.4
       overall = 0.4671                                        max =       162

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

                               (Std. Err. adjusted for 537 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
     lavgsal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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)
------------------------------------------------------------------------------
```

```
. 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

                                                F(4,1307)          =    397.05
corr(u_i, Xb)  = 0.1433                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     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
```

```
. 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                         Obs per group: min =         1
       between = 0.3544                                        avg =       3.4
       overall = 0.4567                                        max =       162

                                                F(4,536)           =     57.84
corr(u_i, Xb)  = 0.1433                         Prob > F           =    0.0000

                               (Std. Err. adjusted for 537 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
     lavgsal |      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)
------------------------------------------------------------------------------
             |               Robust
     lavgsal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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)
------------------------------------------------------------------------------
```

```
. * 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)
------------------------------------------------------------------------------
             |               Robust
     lavgsal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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)
------------------------------------------------------------------------------
```

```
. 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, $G \rightarrow \infty$ asymptotic analysis not realistic? Apply results of Hansen (2007, *Journal of Econometrics*).
- Hansen (2007, Theorem 2) 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, v_{gm} within each group (but independence across groups).

- 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 $\mathbf{z}_{gm}$. 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

$$y_{gmt} = \eta_t + \mathbf{w}_g\boldsymbol{\alpha} + \mathbf{x}_{gm}\boldsymbol{\beta} + \mathbf{z}_{gmt}\boldsymbol{\delta} + h_g + c_{gm} + u_{gmt},$$

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

$$v_{gmt} = h_g + c_{gm} + u_{gmt}.$$

- Generally, the model can include variables that change at any level.
- Some elements of $\mathbf{z}_{gmt}$ 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 $\mathbf{w}_g$, or variables that change across g and t, need to recognize h_g.

- If assume the error v_{gmt} is uncorrelated with $(\mathbf{w}_g, \mathbf{x}_{gm}, \mathbf{z}_{gmt})$, pooled OLS is simple and attractive. Consistent as $G \rightarrow \infty$ 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:

```
reg y w1 ... wJ x1 ... xK z1 ... zL,
    cluster(industryid)
```

- Compare with inference robust only to serial correlation:

```
reg y w1 ... wJ x1 ... xK z1 ... zL,
    cluster(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:

```
xtset firmid year
xtreg y w1 ... wJ x1 ... xK z1 ... zL, re
      cluster(industryid)
```

- Again, compare with inference robust only to neglected serial correlation:

```
xtreg y w1 ... wJ x1 ... xK z1 ... zL, re
      cluster(firmid)
```

- Formal analysis. Write the equation for each cluster as

$$\mathbf{y}_g = \mathbf{R}_g \boldsymbol{\theta} + \mathbf{v}_g$$

where a row of $\mathbf{R}_g$ is $(1, d2, \ldots, dT, \mathbf{w}_g, \mathbf{x}_{gm}, \mathbf{z}_{gmt})$ (which includes a full set of period dummies) and $\boldsymbol{\theta}$ is the vector of all regression parameters. For cluster g, $\mathbf{y}_g$ contains $M_g T$ elements (T periods for each unit m).

- In particular,

$$\mathbf{y}_g = \begin{pmatrix} \mathbf{y}_{g1} \\ \mathbf{y}_{g2} \\ \vdots \\ \mathbf{y}_{g,M_g} \end{pmatrix}, \quad \mathbf{y}_{gm} = \begin{pmatrix} y_{gm1} \\ y_{gm2} \\ \vdots \\ y_{gmT} \end{pmatrix}$$

so that each $\mathbf{y}_{gm}$ is $T \times 1$; $\mathbf{v}_g$ has an identical structure. Now, we can obtain $\mathbf{\Omega}_g = Var(\mathbf{v}_g)$ under various assumptions and apply feasible GLS.

- RE at the unit level is obtained by choosing $\mathbf{\Omega}_g = \mathbf{I}_{M_g} \otimes \mathbf{\Lambda},$ where $\mathbf{\Lambda}$ is the $T \times T$ matrix with the RE structure. If there is within-cluster correlation, this is not the correct form of $Var(\mathbf{v}_g)$, 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 σ_h^2, σ_c^2, and σ_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 $\mathbf{\Omega}_g$ as follows:

$$\begin{aligned} Var(\mathbf{v}_{gm}) &= (\sigma_h^2 + \sigma_c^2)\mathbf{j}_T\mathbf{j}_T' + \sigma_u^2\mathbf{I}_T \\ Cov(\mathbf{v}_{gm}, \mathbf{v}_{gr}) &= \sigma_h^2\mathbf{j}_T\mathbf{j}_T', r \neq m \end{aligned}$$

$$\mathbf{\Omega}_g = \begin{pmatrix} (\sigma_h^2 + \sigma_c^2)\mathbf{j}_T\mathbf{j}_T' + \sigma_u^2\mathbf{I}_T & \cdots & \sigma_h^2\mathbf{j}_T\mathbf{j}_T' \\ \vdots & \ddots & \vdots \\ \sigma_h^2\mathbf{j}_T\mathbf{j}_T' & \cdots & (\sigma_h^2 + \sigma_c^2)\mathbf{j}_T\mathbf{j}_T' + \sigma_u^2\mathbf{I}_T \end{pmatrix}$$

- The robust asymptotic variance of $\hat{\boldsymbol{\theta}}$ is estimated as

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \left(\sum_{g=1}^{G} \mathbf{R}_g' \hat{\boldsymbol{\Omega}}_g^{-1} \mathbf{R}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{R}_g' \hat{\boldsymbol{\Omega}}_g^{-1} \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \hat{\boldsymbol{\Omega}}_g^{-1} \mathbf{R}_g\right)^{-1} \cdot \left(\sum_{g=1}^{G} \mathbf{R}_g' \hat{\boldsymbol{\Omega}}_g^{-1} \mathbf{R}_g\right)^{-1},$$

where $\hat{\mathbf{v}}_g = \mathbf{y}_g - \mathbf{R}_g \hat{\boldsymbol{\theta}}$.

- Unfortunately, routines intended for estimating HLMs (or mixed models) assume that the structure imposed on $\mathbf{\Omega}_g$ is correct, and that $Var(\mathbf{v}_g|\mathbf{R}_g) = Var(\mathbf{v}_g)$. 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 $\mathbf{w}_{gt} = \mathbf{w}_g$ for all t, $\mathbf{w}_g$ is also eliminated. But the unit-specific effects, c_{mg}, are still part of the error term. If we are mainly interested in $\boldsymbol{\delta}$, the coefficients on the time-varying variables $\mathbf{z}_{gmt}$, 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

$$\begin{aligned} y_{gmt} &= \eta_t + \mathbf{w}_g \boldsymbol{\alpha} + \mathbf{x}_{gm} \boldsymbol{\beta} + \mathbf{z}_{gmt} \mathbf{d}_{mg} + h_g + c_{mg} + u_{gmt} \\ &= \eta_t + \mathbf{w}_{gt} \boldsymbol{\alpha} + \mathbf{x}_{gm} \boldsymbol{\beta} + \mathbf{z}_{gmt} \boldsymbol{\delta} + h_g + c_{mg} + u_{gmt} + \mathbf{z}_{gmt} \mathbf{e}_{gm}, \end{aligned}$$

where $\mathbf{d}_{gm} = \boldsymbol{\delta} + \mathbf{e}_{gm}$ is a set of unit-specific intercepts on the individual, time-varying covariates $\mathbf{z}_{gmt}$.

- The time-demeaned equation within individual m in cluster g is

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

- FE is still consistent if $E(\mathbf{d}_{mg}|\mathbf{z}_{gmt} - \bar{\mathbf{z}}_{gm}) = E(\mathbf{d}_{mg})$, $m = 1,\ldots,M_g$, $t = 1,\ldots,T$, and all g, and so cluster-robust inference, which is automatically robust to serial correlation and heteroskedsticity, 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                          13 Mar 2009 11:30
 size:       893,750 (99.8% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
distid          float  %9.0g                  district identifier
schid           int    %9.0g                  school identifier
lunch           float  %9.0g                  % eligible for free lunch
enrol           int    %9.0g                  number of students
exppp           int    %9.0g                  expenditure per pupil
math4           float  %9.0g                  % satisfactory, 4th grade math
                                                test
year            int    %9.0g                  1992=school yr 1991-2
cpi             float  %9.0g                  consumer price index
rexppp          float  %9.0g                  (exppp/cpi)*1.695: 1997 $
lrexpp          float  %9.0g                  log(rexpp)
lenrol          float  %9.0g                  log(enrol)
avgrexp         float  %9.0g                  (rexppp + rexppp_1)/2
lavgrexp        float  %9.0g                  log(avgrexp)
tobs            byte   %9.0g                  number of time periods
-------------------------------------------------------------------------------
Sorted by:  schid  year
```

```
. * egen tobs = sum(1), by(schid)

. tab tobs if y98

  number of |
       time |
    periods |      Freq.     Percent        Cum.
------------+-----------------------------------
          3 |        487       29.28       29.28
          4 |        254       15.27       44.56
          5 |        922       55.44      100.00
------------+-----------------------------------
      Total |      1,663      100.00
```

```
. xtreg math4 lavgrexp lunch lenrol y95-y98, fe

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

R-sq:  within  = 0.3602                         Obs per group: min =         3
       between = 0.0292                                        avg =       4.2
       overall = 0.1514                                        max =         5

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   6.288376   2.098685     3.00   0.003     2.174117    10.40264
       lunch |  -.0215072   .0312185    -0.69   0.491     -.082708    .0396935
      lenrol |  -2.038461   1.791604    -1.14   0.255    -5.550718    1.473797
         y95 |    11.6192   .5545233    20.95   0.000     10.53212    12.70629
         y96 |   13.05561   .6630948    19.69   0.000     11.75568    14.35554
         y97 |   10.14771   .7024067    14.45   0.000     8.770713    11.52471
         y98 |   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)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   6.288376   2.431317     2.59   0.010     1.519651     11.0571
       lunch |  -.0215072   .0390732    -0.55   0.582    -.0981445      .05513
      lenrol |  -2.038461   1.789094    -1.14   0.255    -5.547545    1.470623
         y95 |    11.6192   .5358469    21.68   0.000     10.56821     12.6702
         y96 |   13.05561   .6910815    18.89   0.000     11.70014    14.41108
         y97 |   10.14771   .7326314    13.85   0.000     8.710745    11.58468
         y98 |   23.41404   .7669553    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)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   6.288376   3.132334     2.01   0.045     .1331271    12.44363
       lunch |  -.0215072   .0399206    -0.54   0.590    -.0999539    .0569395
      lenrol |  -2.038461   2.098607    -0.97   0.332    -6.162365    2.085443
         y95 |    11.6192   .7210398    16.11   0.000     10.20231     13.0361
         y96 |   13.05561   .9326851    14.00   0.000     11.22282     14.8884
         y97 |   10.14771   .9576417    10.60   0.000      8.26588    12.02954
         y98 |   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)
------------------------------------------------------------------------------
```

• Can allow the slopes to depend on observed covariates and then use various GLS approaches. An equation for unit m at time t in cluster g is

$$y_{gmt} = \mathbf{z}_{gmt}\mathbf{d}_{gm} + v_{gmt}$$

and then decompose the idiosyncratic error, v_{gmt}, as

$$v_{gmt} = \eta_t + c_{gm} + u_{gmt},$$

where the η_t are aggregate time effects. Absorb the group effect, h_{gt}, into u_{gmt}, and allow c_{gm} and u_{gmt} do be correlated within group.

- For each (g, m) define

$$\bar{\mathbf{r}}_{gm} = (\mathbf{w}_g, \bar{\mathbf{x}}_g, \mathbf{x}_{gm}, \bar{\mathbf{z}}_{gm}),$$

where $\bar{\mathbf{x}}_g = M_g^{-1} \sum_{p=1}^{M_g} \mathbf{x}_{gp}$ and $\bar{\mathbf{z}}_{gm} = T^{-1} \sum_{s=1}^{T} \mathbf{z}_{gms}$. In other words, $\bar{\mathbf{r}}_{gm}$ includes the group-level covariates along with group averages of the unit-specific covariates, the unit-specific covariates, and the time averages of the covariates that change over time.

• Assume

$$
\begin{aligned}
c_{gm} &= \alpha + \bar{\mathbf{r}}_{gm}\boldsymbol{\gamma} + a_{gm} \\
\mathbf{d}_{gm} &= \boldsymbol{\delta} + \boldsymbol{\Pi}(\bar{\mathbf{r}}_{gm} - \boldsymbol{\mu}_{\bar{\mathbf{r}}})' + \mathbf{e}_{gm}
\end{aligned}
$$

insert these in the equation, and use basic algebra:

$$
y_{gmt} = \zeta_t + \bar{\mathbf{r}}_{gm}\boldsymbol{\gamma} + \mathbf{z}_{gmt}\boldsymbol{\delta} + [(\bar{\mathbf{r}}_{gm} - \boldsymbol{\mu}_{\bar{\mathbf{r}}}) \otimes \mathbf{z}_{gmt}]\boldsymbol{\pi} + a_{gm} + \mathbf{z}_{gmt}\mathbf{e}_{gm} + u_{gmt},
$$

where $\boldsymbol{\pi} = \mathrm{vec}(\boldsymbol{\Pi})$.

• Important to center $\bar{\mathbf{r}}_{gm}$ about its average before forming the interactions to make $\boldsymbol{\delta}$ the APE.

• Now can apply various GLS methods to this equation, using cluster-robust inference at the g level.

• Similar discussion holds in the context of instrumental variables. Suppose we start with the model

$$y_{gmt} = \eta_t + \mathbf{r}_{gmt}\boldsymbol{\theta} + v_{gmt}$$

where $\mathbf{r}_{gmt}$ contains all covariates and v_{gmt} is the composite error. If we have exogenous variables, say $\mathbf{q}_{gmt}$, such that $E(\mathbf{q}'_{gmt}v_{gmt}) = \mathbf{0}$ and the rank condition holds, then pooled 2SLS is attractive for its simplicity.

- It does not matter whether elements of $\mathbf{r}_{gmt}$ or $\mathbf{q}_{gmt}$ contain elements that change only across g, across g and m, across g and t, or across g, m, and t, provided the rank condition holds. Without further assumptions, the 2SLS variance matrix estimator, and inference generally, should be robust to arbitrary serial correlation and cluster correlation at the most aggregated level. For example, if g indexes counties and m indexes manufacturing plants operating within a county, then we should cluster at the county level.

- May have policy and instruments change only at the county level over time, along with exogenous explanatory variables that change at the plant level (either constant or over time). In evaluating whether the rank condition holds – say, for a single endogenous variable w_{gmt} – one can use a pooled OLS regression w_{gmt} on 1, $d2_t, \ldots, dT_t$, $\mathbf{q}_{gmt}$ (assuming that $\mathbf{q}_{gmt}$ contains all exogenous variables).
- Such a test should be made robust to arbitrary cluster and serial correlation to be convincing.
- The test works even if w_{gmt} does not change across m (or even t for that matter), and the same with $\mathbf{q}_{gmt}$.

- Again, cluster robust inference is valid with large G provided it is made fully robust.
- In the previous scenario, if we apply, say, fixed effects 2SLS, where we eliminate a time-constant, plant-level effect, then we need the variables of interest to at least change over time (if not across m); the same is true of the instruments.
- If we have instruments that change only by g, the FE2SLS estimator – whether we remove a county-level or plant-level effect – does not identify $\boldsymbol{\theta}$.

4. Estimation with a Small Number of Groups

- When G is small and each M_g is large, we might have a different sampling scheme: large random samples are drawn from different segments of a population. Except for the relative dimensions of G and M_g, the resulting data set is essentially indistinguishable from a data set obtained by sampling entire clusters.
- The problem of proper inference when M_g is large relative to G – the “Moulton (1990) problem” – has been recently studied by Donald and Lang (2007).

- DL treat the problem as a small number of random draws from a large number of groups (because they assume independence).
- Simplest case: a single regressor that varies only by group:

$$\begin{aligned} y_{gm} &= \alpha + \beta x_g + c_g + u_{gm} \\ &= \delta_g + \beta x_g + u_{gm}. \end{aligned}$$

In second equation, common slope, β, but intercept, δ_g, that varies across g.

- DL focus on first equation, where c_g is assumed to be independent of x_g with zero mean.

- Note: Because c_g is assumed independent of x_g, the DL criticism of standard pooled methods is not one of endogeneity. It is one of inference.
- DL highlight the problems of applying standard inference leaving c_g as part of the error term, $v_{gm} = c_g + u_{gm}$.
- Pooled OLS inference applied to

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm}$$

can be badly biased because it ignores the cluster correlation. Hansen's results do not apply. (And we cannot use fixed effects estimation here.)

- DL propose studying the regression in averages:

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G.$$

- Add some strong assumptions: $M_g = M$ for all g, $c_g | x_g \sim Normal(0, \sigma_c^2)$ and $u_{gm} | x_g, c_g \sim Normal(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim Normal(0, \sigma_c^2 + \sigma_u^2/M)$. Then the model in averages satisfies the classical linear model assumptions (we assume independent sampling across g).

- So, we can just use the "between" regression

$$\bar{y}_g \text{ on } 1, x_g, g = 1, \ldots, G.$$

- The estimates of α and β are identical to pooled OLS across g and m

when $M_g = M$ for all g.

• Conditional on the x_g, $\hat{\beta}$ inherits its distribution from $\{\bar{v}_g : g = 1, \ldots, G\}$, the within-group averages of the composite errors.

• We can use inference based on the t_{G-2} distribution to test hypotheses about β, provided $G > 2$.

• If G is small, the requirements for a significant t statistic using the t_{G-2} distribution are much more stringent then if we use the $t_{M_1+M_2+\ldots+M_G-2}$ distribution – which is what we would be doing if we use the usual pooled OLS statistics.

• Using the averages in an OLS regression is *not* the same as using cluster-robust standard errors for pooled OLS. Those are not justified and, anyway, we would use the wrong df in the t distribution.

• We can apply the DL method without normality of the u_{gm} if the group sizes are large because $Var(\bar{v}_g) = \sigma_c^2 + \sigma_u^2/M_g$ so that $\bar{u}_g$ is a negligible part of $\bar{v}_g$. But we still need to assume c_g is normally distributed.

• If $\mathbf{z}_{gm}$ appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \bar{\mathbf{z}}_g\boldsymbol{\gamma} + \bar{v}_g, g = 1,\ldots,G,$$

provided $G > K + L + 1$.

- Inference can be carried out using the $t_{G-K-L-1}$ distribution.
- Regressions on averages are reasonably common, at least as a check on results using disaggregated data, but usually with larger G then just a handful.
- If $G = 2$ in the DL setting, we cannot do inference (there are zero degrees of freedom).
- Suppose x_g is binary, indicating treatment and control ($g = 2$ is the treatment, $g = 1$ is the control). The DL estimate of β is the usual one: $\hat{\beta} = \bar{y}_2 - \bar{y}_1$. But we cannot compute a standard error for $\hat{\beta}$.

- So according the the DL framework the traditional comparison-of-means approach to policy analysis cannot be used. Should we just give up when $G = 2$?
- In a sense the problem is an artifact of saying there are three group-level parameters. If we write

$$y_{gm} = \delta_g + \beta x_g + u_{gm}$$

where $x_1 = 0$ and $x_2 = 1$, then $E(y_{1m}) = \delta_1$ and $E(y_{2m}) = \delta_2 + \beta$. There are only two means but three parameters.

- The usual approach simply defines $\mu_1 = E(y_{1m})$, $\mu_2 = E(y_{2m})$, and then uses random samples from each group to estimate the means. Any "cluster effect" is contained in the means.
- Remember, in the DL framework, the cluster effect is independent of x_g, so the DL criticism is not about systematic bias.

- Applies to simple difference-in-differences settings. Let $y_{gm} = w_{gm2} - w_{gm1}$ be the change in a variable w from period one to two. So, we have a before period and an after period, and suppose a treated group ($x_2 = 1$) and a control group ($x_1 = 0$). So $G = 2$.
- The estimator of β is the DD estimator:

$$\hat{\beta} = \overline{\Delta w}_2 - \overline{\Delta w}_1$$

where $\overline{\Delta w}_2$ is the average of changes for the treament group and $\overline{\Delta w}_1$ is the average change for the control.

- Card and Krueger (1994) minimum wage example: $G = 2$ so, according to DL, cannot put a confidence interval around the estimated change in employment.
- If we go back to

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm}$$

when $x_1 = 0, x_2 = 1$, one can argue that c_g should just be part of the estimated mean for group g. It is assumed assignment is exogenous.
- In the traditional view, we are estimating $\mu_1 = \alpha + c_1$ and $\mu_2 = \alpha + \beta + c_2$ and so the estimated policy effect is $\beta + (c_2 - c_1)$.

• Even when DL approach applies, should we use it? Suppose $G = 4$ with two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$). DL involves the OLS regression $\bar{y}_g$ on $1, x_g$, $g = 1, \ldots, 4$; inference is based on the t_2 distribution. Can show

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2,$$

which shows $\hat{\beta}$ is approximately normal (for most underlying population distributions) even with moderate group sizes M_g.

• In effect, the DL approach rejects usual inference based on means from large samples because it may not be the case that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. Why not allow heterogeneous means?

• Could just define the treatment effect as, say,

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2,$$

and then plug in the unbiased, consistent, asymptotically normal estimators of the μ_g under random sampling within each g.

• The expression $\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2$ hints at a different way to view the small G, large M_g setup. We estimated two parameters, α and β, given four moments that we can estimate with the data.

• The OLS estimates of α and β can be interpreted as minimum distance estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. In the general MD notation, $\boldsymbol{\pi} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and

$$\mathbf{h}(\boldsymbol{\theta}) = \begin{pmatrix} \alpha \\ \alpha \\ \alpha + \beta \\ \alpha + \beta \end{pmatrix}.$$

• Can show that if we use the 4×4 identity matrix as the weight matrix, we get $\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2$ and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$.

- In the general setting, with large group sizes M_g, and whether or not G is especially large, we can put the problem into an MD framework, as done by Loeb and Bound (1996), who had $G = 36$ cohort-division groups and many observations per group.
- Idea is to think of a set of G linear models at the invididual (m) level with group-specific intercepts (and possibly slopes).

- For each group g, write

$$y_{gm} = \delta_g + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + u_{gm}$$
$$E(u_{gm}) = 0, E(\mathbf{z}'_{gm}u_{gm}) = \mathbf{0}.$$

Within-group OLS estimators of δ_g and $\boldsymbol{\gamma}_g$ are $\sqrt{M_g}$-asymptotically normal under random sampling within group.

- The presence of aggregate features $\mathbf{x}_g$ can be viewed as putting restrictions on the intercepts:

$$\delta_g = \alpha + \mathbf{x}_g \boldsymbol{\beta}, g = 1, \ldots, G.$$

- With K attributes ($\mathbf{x}_g$ is $1 \times K$) we must have $G \geq K + 1$ to determine α and $\boldsymbol{\beta}$.
- In the first stage, obtain $\hat{\delta}_g$, either by group-specific regressions or pooling to impose some common slope elements in $\boldsymbol{\gamma}_g$.
- If we impose some restrictions on the $\boldsymbol{\gamma}_g$, such as $\boldsymbol{\gamma}_g = \boldsymbol{\gamma}$ for all g, the $\hat{\delta}_g$ are (asymptotically) correlated.

• Let $\hat{\mathbf{V}}$ be the $G \times G$ estimated (asymptotic) variance of the $G \times 1$ vector $\hat{\boldsymbol{\delta}}$. Let $\mathbf{X}$ be the $G \times (K+1)$ matrix with rows $(1, \mathbf{x}_g)$. The MD estimator is

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\delta}}$$

The asymptotics are as each group size gets large, and $\hat{\boldsymbol{\theta}}$ has an asymptotic normal distribution; its estimated asymptotic variance is $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$.

• Estimator looks like "GLS," but inference is with G (number of rows in $\mathbf{X}$) fixed and M_g growing.

• When separate group regressions are used for each g, the $\hat{\delta}_g$ are independent and $\hat{\mathbf{V}}$ is diagonal, and $\hat{\boldsymbol{\theta}}$ looks like a weighted least squares estimator. That is, treat the $\{(\hat{\delta}_g, \mathbf{x}_g) : g = 1, \ldots, G\}$ as the data and use WLS of $\hat{\delta}_g$ on $1, \mathbf{x}_g$ using weights $1/[se(\hat{\delta}_g)]^2$.

• Can test the $G - (K + 1)$ overidentification restrictions using the SSR from the "weighted least squares" as approximately χ^2_{G-K-1}.

• What happens if the overidentifying restrictions reject?

(1) Can search for more features to include in $\mathbf{x}_g$. If $G = K + 1$, no restrictions to test.

(2) Think about whether a rejection is important. In the program evaluation applications, rejection generally occurs if group means within the control groups or within the treatment groups differ. For example, in the $G = 4$ case with $x_1 = x_2 = 0$ and $x_3 = x_4 = 1$, the test will reject if $\mu_1 \neq \mu_2$ or $\mu_3 \neq \mu_4$. But why should we care? We might want to allow heterogeneous policy effects and define the parameter of interest as

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2.$$

(3) Apply the DL approach on the group-specific intercepts. That is, write

$$\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g, g = 1, \ldots, G$$

and assume that this equation satisfies the classical linear model assumptions.

- With large group sizes, we can act as if

$$\hat{\delta}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g, g = 1, \ldots, G$$

because $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$ and we can ignore the $O_p(M_g^{-1/2})$ part. But we must assume c_g is homoskedastic, normally distributed, and independent of $\mathbf{x}_g$.

- Note how we only need $G > K + 1$ because the $\mathbf{z}_{gm}$ have been accounted for in the first stage in obtaining the $\hat{\delta}_g$. But we are ignoring the estimation error in the $\hat{\delta}_g$.

5. 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

$$
\left(\sum_{g=1}^{G} \sum_{m=1}^{M_g} \mathbf{x}'_{gm} \mathbf{x}_{gm} / p_{gm} \right)^{-1}
\cdot \left[\sum_{g=1}^{G} \sum_{m=1}^{M_g} \sum_{r=1}^{M_g} \hat{u}_{gm} \hat{u}_{gr} \mathbf{x}'_{gm} \mathbf{x}_{gr} / (p_{gm} p_{gr}) \right]
\cdot \left(\sum_{g=1}^{G} \sum_{m=1}^{M_g} \mathbf{x}'_{gm} \mathbf{x}_{gm} / p_{gm} \right)^{-1}.
$$

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 $\{z^o_{scm} : m = 1, \ldots, M_{sc}, c = 1, \ldots, C_s, s = 1, \ldots, S\}$, so the population mean is

$$\mu = M^{-1} \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z^o_{scm}.$$

Define the total in the population as

$$\tau = \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z^o_{scm} = M\mu.$$

Totals within each cluster and then stratum are, respectively,

$$\tau_{sc} = \sum_{m=1}^{M_{sc}} z_{scm}^{o}$$

$$\tau_{s} = \sum_{c=1}^{C_s} \tau_{sc}$$

• 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 the total, $\tau_{sc} = M_{sc}\mu_{sc}$. So, $\hat{\tau}_{sc} = M_{sc}\hat{\mu}_{sc}$ is an unbiased estimator of τ_{sc} for all $\{(s, c) : c = 1, \ldots, C_s, s = 1, \ldots, S\}$ (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 τ_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 μ, 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 (and many other estimation methods), specify the problem as

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

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

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

where $\hat{\mathbf{B}}$ is somewhat complicated:

$$
\begin{aligned}
\hat{\mathbf{B}} = & \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \omega_{sc}^2 \hat{u}_{scm}^2 \mathbf{x}_{scm}' \mathbf{x}_{scm} \\
& + \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\sum_{r\neq m}^{K_{sc}} \omega_{sc}^2 \hat{u}_{scm} \hat{u}_{scr} \mathbf{x}_{scm}' \mathbf{x}_{scr}
\end{aligned}
$$

$$
-\sum_{s=1}^{S} N_s^{-1} \left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}_{scm}' \hat{u}_{scm} \right) \left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}_{scm}' \hat{u}_{scm} \right)'
$$

- The first part of $\hat{\mathbf{B}}$ 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:

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

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

```
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 |
      ,..., |
     5=poor |      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

. 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

Number of strata   =        31                  Number of obs      =      4940
Number of PSUs     =        62                  Population size    =  56316764
                                                Design df          =        31
                                                F(   5,     27)    =     78.49
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lead |  -.0059646   .0045114    -1.32   0.196    -.0151656    .0032364
      female |  -.1529889    .057348    -2.67   0.012    -.2699508    -.036027
       black |   -.535801   .0622171    -8.61   0.000    -.6626937   -.4089084
         age |  -.0236837   .0011995   -19.75   0.000      -.02613   -.0212373
      weight |  -.0035402   .0010954    -3.23   0.003    -.0057743   -.0013061
-------------+----------------------------------------------------------------
       /cut1 |  -3.278321   .1711369   -19.16   0.000    -3.627357   -2.929285
       /cut2 |  -2.496875   .1571842   -15.89   0.000    -2.817454   -2.176296
       /cut3 |  -1.611873   .1511986   -10.66   0.000    -1.920244   -1.303501
       /cut4 |  -.8415657   .1488381    -5.65   0.000    -1.145123   -.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                         Number of obs   =       4940
                                                  LR chi2(5)      =     786.19
                                                  Prob > chi2     =     0.0000
Log likelihood = -7133.6805                       Pseudo R2       =     0.0522

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        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 |  -.6615336   .1028773                     -.8631693   -.4598978
------------------------------------------------------------------------------
```