8. Panel data II: random effects and clustered data

LPO 8852: Regression II

Sean P. Corcoran

Panel data: fixed effects models

In Lecture 7, we used panel data to address omitted variables bias due to unobserved heterogeneity (u_i):

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_i + e_{it}$$

i is a group or individual with multiple observations t, and $Cov(x_{it}, u_i) \neq 0$. (NOTE: switching notation here—u_i was c_i in FE lecture)

Estimation methods:

- Fixed effects "within" regression (LSDV; `xtreg, fe`; or `areg`)
- First-difference or long-difference

Key assumption: *strict exogeneity*, no within- or cross-period correlation between e_{it} and x_{it}.

Panel data: fixed effects models

Advantages:

- Unobserved u_i can be correlated with the explanatory variables
- β_1 is estimated using *within-group* (i) variation in x, y

Disadvantages:

- Cannot estimate slope coefficients for time-invariant x
- Fixed effects "remove" a lot of the variation in y
- The "within" model is less efficient (higher standard errors)
- There may be more measurement error (and attenuation bias) when relying on within-group *changes* vs. levels
- Group intercepts use up a lot of degrees of freedom

Random effects

The fixed effects model allows u_i to be correlated with x_{it}. An alternative conception of u_i is as a *random* effect, uncorrelated with x_{it}.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \underbrace{u_i + e_{it}}_{v_{it}}$$

Think of v_{it} as a *composite* error consisting of a between-group component (u_i) common to all observations within the group and a within-group component (e_{it}). It is assumed u_i and e_{it} are independent of one another and:

$$u_i \sim N(0, \sigma_u^2)$$
$$e_{it} \sim N(0, \sigma_e^2)$$

Sometimes called a "random intercepts" model.

Random effects

If u_i is uncorrelated with x_{it}, then the composite error term v_{it} is uncorrelated with x_{it}. (We already assumed e_{it} is uncorrelated with x_{it}). This means the OLS estimator for β_1 will be unbiased and consistent.

Note: estimation of this model does *not* involve estimating the u_i's as parameters as in the LSDV model.

Random effects

The composite error term v_{it} is not, however, i.i.d.:

$$Corr(v_{it}, v_{is}) = \rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \text{ for } s \neq t$$

The common error for observations in group i (u_i) results in correlation between the composite error in period t (v_{it}) and in period s (v_{is}).

This means OLS is consistent but not efficient, and that traditional standard error formulas assuming i.i.d. errors are incorrect. The ratio above (ρ) is called the **intra-class correlation** (more on this later).

Estimation using GLS (details later): `xtreg, re`.

Success for All example

- Success for All is a whole-school literacy intervention.
- Borman et al. (2005) conducted a randomized evaluation of SFA in 2001-02 and 2002-03 (21 treatment schools and 20 control).
- A *cluster-randomized* design with randomization at the school level.
- The data used by Murnane & Willett (*ch7_sfa.dta*) include grade 1 only. The outcome of interest is *wattack*, the student's score on a "Word-Attack" test.

Next slide: an "unconditional" model with no x_{it} estimates variance components σ_u^2 and σ_e^2 and the intra-class correlation ρ.

Random effects with xtreg

```
. xtreg wattack, re i(schid)

Random-effects GLS regression                   Number of obs     =      2,334
Group variable: schid                           Number of groups  =         41

R-sq:                                           Obs per group:
     within  = 0.0000                                         min =         10
     between = 0.0000                                         avg =       56.9
     overall = 0.0000                                         max =        134

                                                Wald chi2(0)      =          .
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =          .

------------------------------------------------------------------------------
     wattack |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   477.5356   1.447118   329.99   0.000     474.6994    480.3719
-------------+----------------------------------------------------------------
     sigma_u |  8.8705267
     sigma_e |  17.725757
         rho |  .20027618   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

This example: Success for All impact evaluation (from Murnane & Willett). $\sigma_u^2 = 8.87^2 = 78.7$ and $\sigma_e^2 = 17.73^2 = 314.35$. $\rho = 0.200$.

loneway

loneway (one-way ANOVA) is another handy command for estimating variance components and ICC. (Note the difference in σ_u and ρ from xtreg, re. With unbalanced panels, these will differ slightly).

```
. loneway wattack schid

          One-way Analysis of Variance for wattack: word attack posttest

                                        Number of obs =      2,334
                                            R-squared =     0.2185

    Source                SS         df      MS            F     Prob > F

Between schid         201450.43      40   5036.2607     16.03     0.0000
Within schid          720466.21   2,293   314.20244

Total                 921916.63   2,333   395.16358

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]

            0.20993     0.04402      0.12366     0.29621

         Estimated SD of schid effect          9.137203
         Estimated SD within schid             17.72576
         Est. reliability of a schid mean       0.93761
              (evaluated at n=56.56)
```


Random effects with xtreg

```
. xtreg wattack sfa ppvt, re i(schid)

Random-effects GLS regression                   Number of obs     =      2,334
Group variable: schid                           Number of groups  =         41

R-sq:                                           Obs per group:
     within  = 0.1101                                         min =         10
     between = 0.3960                                         avg =       56.9
     overall = 0.1820                                         max =        134

                                                Wald chi2(2)      =     308.21
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     wattack |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sfa |   3.440921   2.297268     1.50   0.134    -1.061642    7.943485
        ppvt |   .4851754   .0278075    17.45   0.000     .4306737    .5396771
       _cons |   432.0475   2.972263   145.36   0.000      426.222     437.873
-------------+----------------------------------------------------------------
     sigma_u |  6.9082397
     sigma_e |  16.725172
         rho |  .14574141   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

This regression: includes the treatment indicator (*sfa*) and one covariate (*ppvt*). Note changes in σ_u and σ_e, ρ. The residual variability is reduced with the inclusion of x's.

Random effects

Class size and passing rates in TX (see previous panel data lecture):

```
. xtreg avgpassing avgclass, re i(campus)

Random-effects GLS regression                   Number of obs     =     16,062
Group variable: campus                          Number of groups  =      4,326

R-sq:                                           Obs per group:
     within  = 0.0018                                         min =          1
     between = 0.0098                                         avg =        3.7
     overall = 0.0060                                         max =          4

                                                Wald chi2(1)      =       2.74
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0978

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.0442893   .0267548    -1.66   0.098    -.0967277    .0081491
       _cons |   76.21828   .5503649   138.49   0.000     75.13959    77.29698
-------------+----------------------------------------------------------------
     sigma_u |  12.391941
     sigma_e |  6.4870883
         rho |  .78490199   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```


Random effects

Compare to fixed effects: very different slope coefficient estimate.

```
. xtreg avgpassing avgclass, fe i(campus)

Fixed-effects (within) regression               Number of obs     =     16,062
Group variable: campus                          Number of groups  =      4,326

R-sq:                                           Obs per group:
     within  = 0.0018                                         min =          1
     between = 0.0098                                         avg =        3.7
     overall = 0.0060                                         max =          4

                                                F(1,11735)        =      21.30
corr(u_i, Xb)  = -0.1189                        Prob > F          =     0.0000

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.1339024   .0290105    -4.62   0.000    -.1907678   -.0770371
       _cons |   78.09211   .5590819   139.68   0.000     76.99621      79.188
-------------+----------------------------------------------------------------
     sigma_u |  12.997022
     sigma_e |  6.4870883
         rho |  .80056238   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(4325, 11735) = 13.83                Prob > F = 0.0000
```

Random vs. fixed effects

- The RE model is biased and inconsistent if the FE assumptions are more appropriate (correlation between x_{it} and u_i).
- If the RE assumptions hold (no correlation between x_{it} and u_i), both RE and FE are *consistent*. They should give "similar" answers in large samples, but the FE model will be *inefficient* (larger standard errors).
- A sufficiently large difference in point estimates suggests the FE assumption is probably correct and RE is inconsistent.
- The **Hausman test** is a formal test of this.

Hausman test

First use `estimates store` to save your `fe` and `re` estimates. Name them FE and RE, for example.

```
xtreg avgpassing avglcass, fe i(campus)
estimates store FE
xtreg avgpassing avgclass, re i(campus)
estimates store RE
hausman FE RE
```

Hausman test

Null hypothesis: RE assumptions hold, both estimators consistent but RE is efficient. Alternative: RE assumptions do *not* hold and the RE estimator is inconsistent. In the TX example we can reject H_0:

```
. hausman FE RE

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
    avgclass |   -.1339024    -.0442893       -.0896131        .0112156
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       63.84
                Prob>chi2 =      0.0000
```


Review of GLS

In a linear regression with known heteroskedasticity, we can transform the original data and apply OLS to the transformed data. E.g.:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

with $\mathrm{Var}(u_i) = k_i \sigma_u^2$. The GLS transformation divides the data by $\sqrt{k_i}$. Observations with greater variance get *less* weight. The transformed model satisfies homoskedasticity.

GLS estimation of random effects models

The random effects model with one covariate is:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \underbrace{u_i + e_{it}}_{v_{it}}$$

GLS estimation again involves a transformation. Let:

$$\theta = 1 - \sqrt{\frac{\sigma_e^2}{\sigma_e^2 + T\sigma_u^2}}$$

(and note the term under the square root looks like but is different from the ICC). T is the number of observations per group, assuming a balanced panel.

GLS estimation of random effects models

The transformations of y_{it} and x_{it} are:

$$y_{it} - \theta\bar{y}_i$$
$$x_{it} - \theta\bar{x}_i$$

and OLS is estimated on the transformed model:

$$y_{it} - \theta\bar{y}_i = \beta_0(1-\theta) + \beta_1(x_{it} - \theta\bar{x}_i) + (v_{it} - \theta\bar{v}_i)$$

The transformed y_{it} and x_{it} are *quasi-demeaned*. If $\theta = 1$, we have the demeaned (within) model.

GLS estimation of random effects models

θ is not known so it must first be estimated with consistent estimators for σ_e^2 and σ_u^2. Then, $\hat{\theta}$ is used in OLS estimation ("feasible GLS").

$$\hat{\theta} = 1 - \sqrt{\frac{\hat{\sigma}_e^2}{\hat{\sigma}_e^2 + T\hat{\sigma}_u^2}}$$

Consistent estimators for σ_u^2 and σ_e^2 can be obtained using pooled OLS or fixed effects residuals.

GLS estimation of random effects models

One method for estimating σ_u^2 and σ_e^2: note that

$$v_{it} = u_i + e_{it}$$

$$v_{it} v_{is} = (u_i + e_{it})(u_i + e_{is})$$

$$E(v_{it} v_{is}) = \underbrace{E(u_i^2)}_{\sigma_u^2} + \underbrace{E(u_i e_{is})}_{0} + \underbrace{E(u_i e_{it})}_{0} + \underbrace{E(e_{it} e_{is})}_{0}$$

Get the composite residuals $\hat{v}_{it}$ using pooled OLS. The square of the RMSE in this regression estimates σ_v^2. The within-group covariance in $\hat{v}_{it}$ (the sample analog of $E(v_{it} v_{is})$ above) provides a consistent estimate of σ_u^2. Then, $\hat{\sigma}_e^2 = \hat{\sigma}_v^2 - \hat{\sigma}_u^2$. See problem set.

GLS estimation of random effects models

$$y_{it} - \theta\bar{y}_i = \beta_0(1-\theta) + \beta_1(x_{it} - \theta\bar{x}_i) + (v_{it} - \theta\bar{v}_i)$$

$$\theta = 1 - \sqrt{\frac{\sigma_e^2}{\sigma_e^2 + T\sigma_u^2}}$$

Notice the transformation subtracts a *fraction* of the within-group mean, where the fraction depends on σ_e^2, σ_u^2, and T.

- When $\theta = 0$, the model reduces to pooled OLS
- When $\theta = 1$, the model reduces to fixed effects (within)
- So, the value of θ is indicative of which model RE is closer to

θ gets closer to 1 as between-group variation σ_u^2 grows relative to within-group variation σ_e^2, and as the number of time periods T grows.

GLS estimation of random effects models

Can request $\hat{\theta}$ in `xtreg, re`:

```
. xtreg avgpassing avgclass, re i(campus) theta

Random-effects GLS regression                   Number of obs     =     16,062
Group variable: campus                          Number of groups  =      4,326

R-sq:                                           Obs per group:
     within  = 0.0018                                         min =          1
     between = 0.0098                                         avg =        3.7
     overall = 0.0060                                         max =          4

                                                Wald chi2(1)      =       2.74
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0978

------------------- theta --------------------
  min      5%       median        95%      max
0.5362   0.6529     0.7468     0.7468   0.7468

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.0442893   .0267548    -1.66   0.098    -.0967277    .0081491
       _cons |   76.21828   .5503649   138.49   0.000     75.13959    77.29698
-------------+----------------------------------------------------------------
     sigma_u |  12.391941
     sigma_e |  6.4870883
         rho |  .78490199   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

This uses the original unbalanced panel, so $\hat{\theta}$ varies with group size.

GLS estimation of random effects models

Can request $\hat{\theta}$ in `xtreg, re`:

```
. xtreg avgpassing avgclass, re i(campus) theta

Random-effects GLS regression                   Number of obs      =     14,796
Group variable: campus                          Number of groups   =      3,699

R-sq:                                           Obs per group:
     within  = 0.0020                                         min =          4
     between = 0.0138                                         avg =        4.0
     overall = 0.0061                                         max =          4

                                                Wald chi2(1)       =       2.97
corr(u_i, X)     = 0 (assumed)                  Prob > chi2        =     0.0848
theta            = .73287384

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.0484254   .0280999    -1.72   0.085    -.1035003    .0066494
       _cons |   76.51251   .5742248   133.24   0.000     75.38705    77.63797
-------------+----------------------------------------------------------------
     sigma_u |  11.706021
     sigma_e |  6.4897977
         rho |  .76490175   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

This uses the <u>balanced</u> panel, so $\hat{\theta}$ is constant.

GLS estimation of random effects models

It is useful to consider the error term in the quasi-demeaned model:

$$v_{it} - \theta \bar{v}_i = (1 - \theta) u_i + (e_{it} - \theta \bar{e}_i)$$

Suppose the RE assumption that u_i is uncorrelated with x_{it} does *not* hold. As $\theta \rightarrow 1$, the u_i component of the error term diminishes in importance, the RE estimator tends toward the FE estimator, and any bias associated with RE tends to zero.

MLE estimation of random effects models

Random effects models can also be estimated using maximum likelihood in which case all parameters of the model (β's, σ's) are estimated jointly:

```
. xtreg avgpassing avgclass, mle i(campus)

Fitting constant-only model:
Iteration 0:   log likelihood = -53584.523
Iteration 1:   log likelihood = -53584.523

Fitting full model:
Iteration 0:   log likelihood = -53674.187
Iteration 1:   log likelihood = -53583.763
Iteration 2:   log likelihood = -53582.969
Iteration 3:   log likelihood = -53582.969

Random-effects ML regression                    Number of obs     =     14,796
Group variable: campus                          Number of groups  =      3,699

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =          4
                                                              avg =        4.0
                                                              max =          4

                                                LR chi2(1)        =       3.11
Log likelihood  = -53582.969                    Prob > chi2       =     0.0780

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.0496391   .0281539    -1.76   0.078    -.1048197    .0055415
       _cons |   76.53576   .5755876   132.97   0.000     75.40763    77.66389
-------------+----------------------------------------------------------------
    /sigma_u |    11.8066   .1481004                      11.51987    12.10047
    /sigma_e |   6.492198   .0436102                      6.407283    6.578237
         rho |   .7678329   .0051631                      .7575916    .7778289
------------------------------------------------------------------------------
LR test of sigma_u=0: chibar2(01) = 1.2e+04           Prob >= chibar2 = 0.000
```


Getting estimates of u_i

As with `xtreg, fe`, one can obtain the $\hat{u}_i$ estimates of the group random effects. Unlike `fe`, these are not coefficient estimates but rather estimated from residuals. The random effects $\hat{u}_i$ can be calculated in two ways:

- Maximum likelihood (following `xtreg, mle`)
- Empirical Bayes / shrinkage approach: the Best Linear Unbiased Predictors (BLUPs)

Shrinkage approach: multiply $\hat{u}_i$ by a shrinkage factor $\hat{R}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{T_i}}$

where T_i is the number of observations in group i. Examples on next 3 slides.

Getting estimates of u_i: MLE

```
. xtreg avgpassing avgclass, re mle i(campus)

Fitting constant-only model:
Iteration 0:   log likelihood = -53584.523
Iteration 1:   log likelihood = -53584.523

Fitting full model:
Iteration 0:   log likelihood = -53674.187
Iteration 1:   log likelihood = -53583.763
Iteration 2:   log likelihood = -53582.969
Iteration 3:   log likelihood = -53582.969

Random-effects ML regression                    Number of obs     =     14,796
Group variable: campus                          Number of groups  =      3,699

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =          4
                                                              avg =        4.0
                                                              max =          4

                                                LR chi2(1)        =       3.11
Log likelihood  = -53582.969                    Prob > chi2       =     0.0780

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.0496391   .0281539    -1.76   0.078    -.1048197    .0055415
       _cons |   76.53576   .5755876   132.97   0.000     75.40763    77.66389
-------------+----------------------------------------------------------------
    /sigma_u |    11.8066   .1481004                      11.51987    12.10047
    /sigma_e |   6.492198   .0436102                      6.407283    6.578237
         rho |   .7678329   .0051631                      .7575916    .7778289
------------------------------------------------------------------------------
LR test of sigma_u=0: chibar2(01) = 1.2e+04            Prob >= chibar2 = 0.000

. predict uhat1, u

. sum uhat1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       uhat1 |     14,796    8.39e-09    12.24512  -47.43509   23.42125
```


Getting estimates of u_i: BLUP

```
. xtreg avgpassing avgclass, re mle i(campus)

Fitting constant-only model:
Iteration 0:   log likelihood = -53584.523
Iteration 1:   log likelihood = -53584.523

Fitting full model:
Iteration 0:   log likelihood = -53674.187
Iteration 1:   log likelihood = -53583.763
Iteration 2:   log likelihood = -53582.969
Iteration 3:   log likelihood = -53582.969

Random-effects ML regression                    Number of obs     =     14,796
Group variable: campus                          Number of groups  =      3,699

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =          4
                                                              avg =        4.0
                                                              max =          4

                                                LR chi2(1)        =       3.11
Log likelihood  = -53582.969                    Prob > chi2       =     0.0780

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.0496391   .0281539    -1.76   0.078    -.1048197    .0055415
       _cons |   76.53576   .5755876   132.97   0.000     75.40763    77.66389
-------------+----------------------------------------------------------------
    /sigma_u |    11.8066   .1481004                      11.51987    12.10047
    /sigma_e |   6.492198   .0436102                      6.407283    6.578237
         rho |   .7678329   .0051631                      .7575916    .7778289
------------------------------------------------------------------------------
LR test of sigma_u=0: chibar2(01) = 1.2e+04            Prob >= chibar2 = 0.000

. gen shrink = _b[/sigma_u]^2 / (_b[/sigma_u]^2 + (_b[/sigma_e]^2)/4)

. gen uhat1s = uhat1*shrink

. summ uhat1s shrink

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      uhat1s |     14,796    1.16e-08    11.38455  -44.10139   21.77522
      shrink |     14,796    .9297209           0   .9297209   .9297209
```

Getting estimates of u_i: BLUP using `xtmixed`

```
. xtmixed avgpassing avgclass || campus: , mle

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -53582.969
Iteration 1:   log likelihood = -53582.969

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =     14,796
Group variable: campus                          Number of groups  =      3,699

                                                Obs per group:
                                                              min =          4
                                                              avg =        4.0
                                                              max =          4

                                                Wald chi2(1)      =       3.13
Log likelihood = -53582.969                     Prob > chi2       =     0.0770

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    avgclass |  -.0496392   .0280727    -1.77   0.077    -.1046606    .0053823
       _cons |   76.53576   .5741313   133.31   0.000     75.41048    77.66103
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
campus: Identity             |
                   sd(_cons) |    11.8066   .1481006      11.51987    12.10047
-----------------------------+------------------------------------------------
                sd(Residual) |   6.492197   .0436102      6.407283    6.578236
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 11666.05      Prob >= chibar2 = 0.0000

. predict uhat2, reffects

. sum uhat2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       uhat2 |     14,796   -6.21e-10    11.38455  -44.10139   21.77523
```


Getting estimates of u_i

The shrinkage factor is smaller for groups with fewer observations (T_i). Their $\hat{u}_i$ is "shrunk" more toward the overall mean group effect of 0.

- RE estimates generally smaller than FE estimates in absolute value
- True for both MLE and EB estimates of the RE, but especially the EB
- The rank order of the $\hat{u}_i$ is usually preserved whether one assumes RE or FE

Random vs. fixed effects

When and where random effects are appropriate:

- As a rule, if the FE assumption holds the RE model is inappropriate. See the Texas class size example, where the Hausman test rejected RE.
- RE is appropriate with grouped or clustered data. See the Success for All example: assignment to treatment was random at the school level, so we need not be concerned about correlation between treatment and the error term. However, the errors are not i.i.d.

See Rabe-Hesketh and Skrondal MLM text for more guidance on RE vs. FE decision.

xttest0

The command xttest0 (following xtreg) provides a formal test for the presence of random effects. H_0 in this case is that the variance across panel units is zero, and thus RE is unnecessary.

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        wattack[schid,t] = Xb + u[schid] + e[schid,t]

        Estimated results:
                        |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                 wattack |   395.1636       19.87872
                       e |   279.7314       16.72517
                       u |   47.72378        6.90824

        Test:   Var(u) = 0
                             chibar2(01) =   1266.18
                          Prob > chibar2 =    0.0000
```

Clustered data

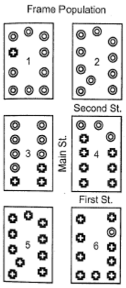
Short-panel data can be thought of as type of "clustered" data, where the individuals (groups) are the clusters with multiple observations t.

When a sample is drawn using clusters, the traditional standard error formula presuming i.i.d. draws will be incorrect.

- Consider the following example using a multi-stage sampling design.
- First the clusters are randomly sampled (the "primary sampling unit").
- Then units within the cluster are randomly sampled.

We will simulate sample means calculated from a simple random sample, and via a clustered sample. We wish to estimate the percent poor in the population. ⊚ poor household and ✜ = rich.

Example: SRS



Samples of Two Distinct Blocks
(Proportion "⊚" Households in Realization)

Block Numbers	Proportion "⊚"
1,2	19/20 = .95 *
2,3	17/20 = .85 *
1,3	16/20 = .80 *
2,4	13/20 = .65
1,4	12/20 = .60
2,6	11/20 = .55
1,6	10/20 = .50
2,5	10/20 = .50
3,4	10/20 = .50
1,5	9/20 = .45
3,6	8/20 = .40
3,5	7/20 = .35
4,6	4/20 = .20 *
4,5	3/20 = .15 *
5,6	1/20 = .05 *

Figure 4.4 A bird's-eye view of a population of 30 "✜" and 30 "⊚" households clustered into six city blocks, from which two blocks are selected.

Circle = poor
Cross = rich

In population, 50% poor households

From Groves et al.

$N = 60$ and $s = 0.504$

Example: SRS

- Consider the mean of a simple random sample of $n = 20$, $\sum_{i=1}^{n} x_i/20$.
- We know this estimator will have a sampling distribution with mean μ and standard error of $\sigma/\sqrt{20}$ which we can estimate with $s/\sqrt{20}$
- Technically, we are sampling a large share of the population in this example (20/60) and need to adjust the standard error downward with the *finite population correction factor*.
- The *fpc* is approximately 1 when the population size N is large.

$$fpc = \sqrt{\frac{N-n}{n}} = \sqrt{\frac{60-20}{60}} = 0.816$$

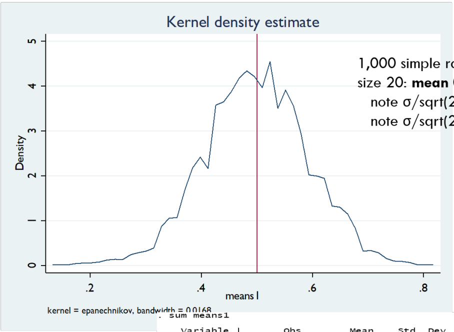
Example: SRS

Applying the *fpc*, the standard error of $\bar{x}$ under a SRS will be:

$$\sqrt{\frac{N-n}{n}} \times \frac{s}{\sqrt{n}} = 0.816 \times \frac{0.504}{\sqrt{20}} = 0.092$$

For the following picture, draw 1,000 SRS and compute $\bar{x}$ for each. Plot the sampling distribution and compute its mean and standard deviation (i.e., the standard error of $\bar{x}$).

Example: SRS



1,000 simple random samples of size 20: **mean 0.499** and **sd 0.092**
note σ/sqrt(20)= 0.112
note σ/sqrt(20)*FPC= 0.092

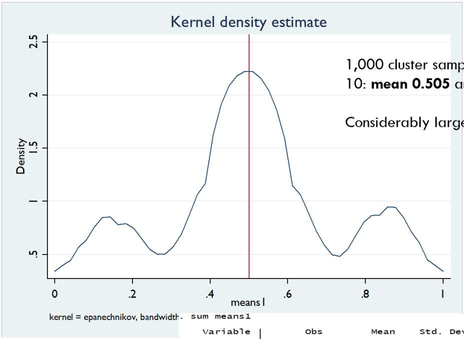
. sum means1

Variable	Obs	Mean	Std. Dev.	Min	Max
means1	1000	.49855	.0921352	.15	.8

Example: cluster sampling

- Now consider the two-stage cluster sample in which 2 blocks are selected (the PSU) and then 10 households from each block.
- Draw 1,000 two-stage cluster samples and compute $\bar{x}$ for each. Plot the sampling distribution and compute its mean and standard deviation (i.e., the standard error of $\bar{x}$).

Example: cluster sampling



1,000 cluster samples with 2 clusters of 10: **mean 0.505** and **sd 0.243**

Considerably larger variance!

```
. sum means1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
means1	1000	.5048	.2425506	.05	.95

Example: cluster sampling

- If for any given sample we had calculated the standard error of $\bar{x}$ as $s/\sqrt{n}$, we would have greatly *understated* the true standard error, and thus *overstated* our estimator's precision!
- The ratio of the variance of $\bar{x}$ under the two sampling designs (cluster vs. SRS) is called the **design effect** or *deff* (d^2).
- The *deft* is the square root of the design effect. Can be used as a rule of thumb to "scale up" or "inflate" a standard error calculated under the assumption of SRS. In the above example:

$$d = \sqrt{d^2} = 0.243/0.092 = 2.64$$

Clustered data and sampling variability

Why the increase in sampling variability under cluster sampling? In the population, there is variation *between* and *within* clusters.

- Holding total variability constant, the greater the variation *between* clusters, the less the variability *within* clusters.
- The greater the share of variability that is between-cluster, the more you "lose" by a cluster sample design.

Imagine a population with perfect homogeneity within clusters. A sample of 1 from a cluster provides just as much information as a larger sample from that cluster. The "effective sample size" shrinks from $N \times n_c$ to N (the number of groups, or clusters). In the above example, from 20 to 2. (n_c is the number of observations per cluster).

Intra-class correlation, revisited

$$ICC = \rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

The ICC is all of the following:

- $Corr(v_{it}, v_{is})$: the extent to which observations within a group (i) are correlated (see earlier definition).
- The fraction of overall variation that is between groups.
- A measure of the amount of "clustering" in the variable of interest. With perfect homogeneity within groups, $\rho = 1$.

ICC and the design effect

Recall the *design effect* (d^2) is the ratio of sampling variation under two designs: cluster sampling and SRS. We saw that statistics under cluster sampling have larger standard errors compared to SRS. The extent to which the standard errors will be larger is related to the amount of clustering (ρ, or ICC):

$$d^2 = 1 + (n_c - 1)\rho$$

.

The larger is ρ, the larger the design effect.

- If $\rho = 0$ then $d^2 = 1$. No clustering, standard errors same as SRS.
- If $\rho = 1$ then $d^2 = n_c$. Extreme clustering, standard errors are larger by a factor of $\sqrt{n_c}$ ($n_c =$ the number of observations per cluster).

Clustered data and regressions

Panel (or cross-sectional) data used in a regression may not have been the result of a cluster sample, but nonetheless has similar "nested" features.

- Classrooms within schools
- Children within classrooms
- Individual students at multiple points in time
- Households within a village
- Members of a household
- State by year difference-in-difference studies

Clustered data and regressions

Clustered data can present problems for inference in a regression, especially for explanatory variables that vary only at the group (i) level. Example: in the Tennessee STAR class size experiment students within schools were randomized to a small or large class (x_i).

Even under random assignment to x_i there is likely to be within-group (cluster) correlation in v_{it}, as in the random effects model:

$$y_{it} = \beta_0 + \beta_1 x_i + \underbrace{u_i + e_{it}}_{v_{it}}$$

Clustered data and regressions

Intuitively, we are trying to estimate the relationship between y_i and x_i in a context in which:

- Observations are clustered within groups i
- x_i does not vary across units within the same group

In this context, more observations from a cluster provide little additional information on y or x! "New" information would require new clusters.

Clustered data and regressions

In cases with *no* variation in x within cluster, the traditional (OLS) standard errors need to be inflated by the *deft*, or "Moulton factor":

$$d = \sqrt{1 + (n_c - 1)\rho}$$

where n_c is the number of observations per group/cluster. Holding total sample size constant, as n_c goes up, the number of clusters goes down.

Example 1: HS&B

Estimating the effect of Catholic school attendance using student-level data from High School & Beyond. See *hsb_subset.dta*

- Data include 10 students per school, 489 schools.
- Students originally sampled in a multi-stage design, with schools selected and then students.
- Regress *soph_scr* (sophomore test score) on a Catholic school indicator and controls (region, urban/rural, etc.)

There is likely to be correlation across students within schools due to a common factor. Use `loneway` or `xtreg` to estimate the intra-class correlation and calculate the Moulton factor.

Example 2: Angrist & Lavy

Estimating the effect of incentive pay for passing matriculation exams in Israel on passing rates.

- Data include 4,000 students in 40 schools, with $n_c = 100$.
- The treatment (offer of incentive pay) occurred at the school level, so x_i does not vary within school.
- Intra-class correlation of $\rho = 0.1$

Deft, or Moulton factor $= \sqrt{1 + (100 - 1)0.1} = 3.30$. Standard errors are approximately 3.3 times higher than those reported by the traditional standard error formula.

Clustered data and regressions

The Moulton factor can be modified for settings in which group sizes vary and x varies within group (but is also clustered):

$$d = \sqrt{1 + \left(\frac{V(n_c)}{\bar{n}} + \bar{n} - 1 \right) \rho_x \rho_v}$$

- $V(n_c)$ is the variance in cluster/group size
- $\bar{n}$ is the average group size
- ρ_x is the intra-class correlation of x
- ρ_v is the intra-class correlation of residuals

Clustered data and regressions

We have been talking about ICC in the context of v_{it} but explanatory variables can also exhibit clustering. That is, x_{it} may also be more similar within groups. As shown in the above formula, clustering has the biggest impact when there are variable group sizes and when ρ_x is large.

Note: the formulas for the Moulton factor above assume equi-correlated errors. That is, $\text{Corr}(v_{it}, v_{is}) = \rho$ for all $s \neq t$. This makes sense when observations are exchangeable, as when the order doesn't matter (e.g., individuals within a household). The consequences of clustering are less extreme when errors are not equi-correlated.

Clustered data: takeaways

Takeaways thus far:

1. With clustered data, observations in the same cluster are more similar to one another than two observations drawn at random from the population. The intra-class correlation ρ is a measure of this similarity.
2. For a given sample size, cluster sampling provides less variation than what one would obtain under a simple random sample. Standard error formulas that assume a SRS (i.i.d. observations) will be incorrect.
3. The same may be true for standard errors of statistics calculated using clustered data (e.g., panel data with a random effect).

Clustered data: takeaways

Takeaways thus far:

4. The design effect d^2 assuming equi-correlated errors and equal group sizes n_c is $1+(n_c-1)\rho$. This is the ratio of the sampling variance accounting for clustering to the sampling variance under a SRS.
5. The square root of this (the "Moulton factor") is the amount by which standard errors assuming SRS should be "inflated."
6. The need to account for clustering increases with ρ and the number of observations within a cluster n_c.
7. If group sizes vary and there is some within-cluster variation in x, the need to account for clustering also depends on ρ_x (the ICC for x) and the variance in group size.

Clustered data and power

Since clustered data affects precision, it also affects statistical *power*: our ability to correctly reject the null hypothesis in favor of an alternative.

To review, consider the sample mean $\bar{x}$, used to test a hypothesis about the population mean $H_0 : \mu = \mu_0$.

The significance level for the test is α (e.g., 0.05).

Suppose the test is one-sided, where we reject when the evidence favors the alternative $H_1 : \mu > \mu_0$.

Clustered data and power

The power of this hypothesis test depends on:

- The effect size of interest (call this δ): how far a specific alternative $\mu_1 = \mu_0 + \delta$ is away from μ_0. All else equal, the closer μ_1 is to μ_0, the *lower* the power.
- α, which determines when we reject. All else equal, a higher α the *greater* the power.
- The standard error of the sample mean ($\sigma/\sqrt{n}$). All else equal, the smaller the standard error, the *greater* the power of the test.
- Because *n* decreases the standard error, a larger *n increases* power, all else equal.

Note: α is the probability of a Type I error. β is the probability of a Type II error. $1 - \beta$ is the power of the test.

Hypothesis test for μ

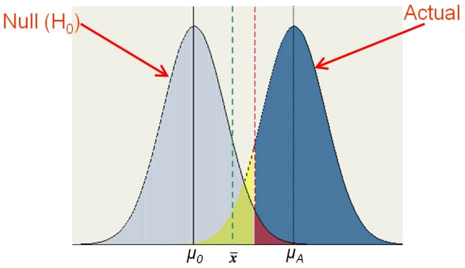
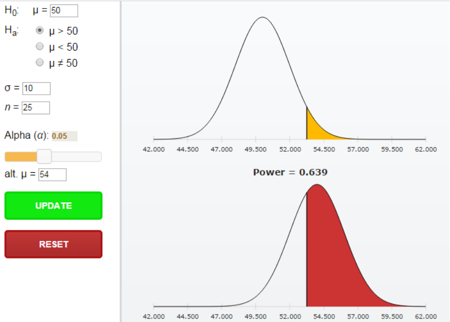


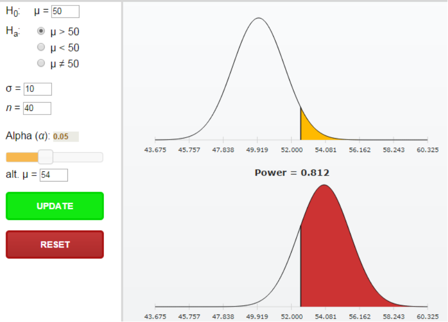
Figure: Distribution of $\bar{x}$ under H_0 and a specific alternative H_A

Power of a hypothesis test for μ



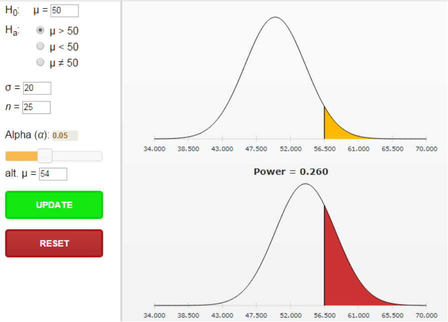
One-sided hypothesis test: $\mu_0 = 50$, $\sigma = 10$, $n = 25$, $\alpha = 0.05$. Find statistical power $(1 - \beta)$ when μ is actually 54.

Power of a hypothesis test for μ



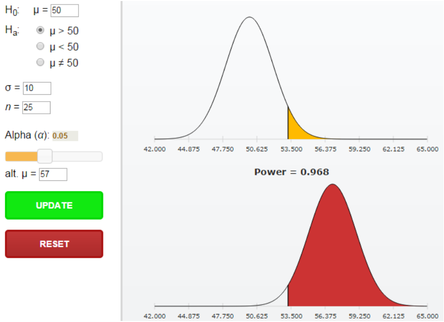
Consider what happens when n increases.

Power of a hypothesis test for μ



Consider what happens when σ increases.

Power of a hypothesis test for μ



Consider what happens when the alternative is further away (e.g. $\mu = 57$).

Clustered data and power

As we have seen, clustering reduces the *effective* sample size and thus reduces power.

Software packages like Optimal Design are useful for examining power to detect effects under different assumptions about clustering, and for determining the sample size needed to detect a given effect size.

- `sites.google.com/site/optimaldesignsoftware/home`

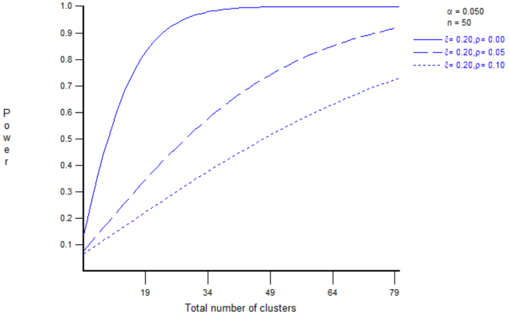
Clustered data and power: example

Example of a hypothetical SFA RCT in Murnane & Willett (ch. 7):

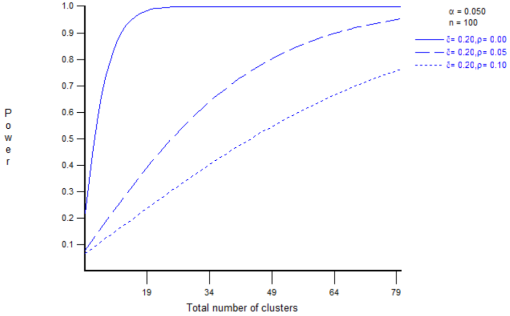
- Cluster-randomized trial (randomization at the school level): J =number of schools
- Observations per cluster (students per school): n_c =50 or 100
- Minimum detectable effect size of interest: $\delta = 0.2$
- Significance level: $\alpha = 0.05$, one-sided test
- ICC: ρ =0, 0.05, or 0.10.

The next two slides use Optimal Design to plot power against the number of schools J. The $\rho = 0$ case is shown as the benchmark of no clustering. Note $\beta = 0.80$ is a commonly-used threshold for acceptable power.

Clustered data and power: example



Clustered data and power: example



Clustered data and power: example

Commands in Optimal Design to produce these figures:

- Design
- Cluster randomized trial with person-level outcomes
- Cluster randomized trial
- Treatment at level 2
- Power vs. ICC

Murnane & Willett advise that you choose an ICC that applies to the point in the time that will be analyzed. For example, if the analysis will be of grade 4 outcomes post-treatment, use the ICC for grade 4 outcomes, not the grade 3 baseline outcomes.

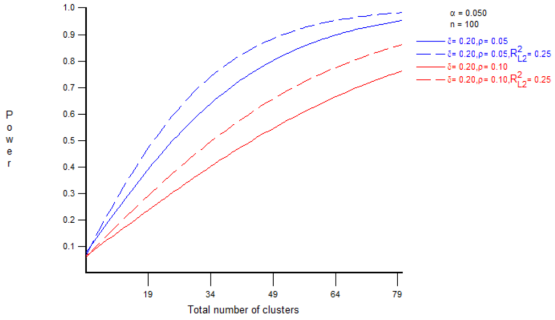
Clustered data and power: example

Some takeaways:

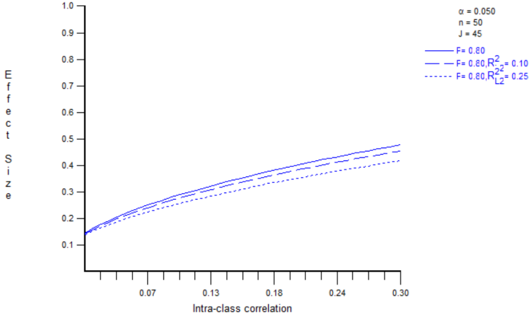
- With $\rho = 0$ can achieve power of 0.8 with a small number of schools ($J = 13$, with $n_c = 50$ students per school).
- With $\rho = 0.05$ or $\rho = 0.10$, need substantially more schools to achieve the same power ($J = 45$ or $J = 75$).
- Increasing the number of students per school has minimal effects on power when there is clustering

Including covariates can reduce residual variance and (possibly) the ICC. *Group*-level covariates will generally yield the biggest reduction of group-level residual variance. In Optimal Design, can set R^2_{L2}: the proportion of variation in the outcome explained by the group-level covariates.

Clustered data and power: example



Clustered data and power: example



Optimal design can produce other plots: e.g., MDES vs. ICC

Cluster-robust inference

When and how to account for the clustering of errors within a regression model? Practical advice (and differing perspectives) from Cameron & Miller (2015) and Abadie et al. (2017).

Approaches:

- Specify a model for the within-cluster correlation, as in the random effects model (`xtreg, re`). Makes strong assumptions about the correlation of errors within clusters.
- After estimation, calculate "cluster-robust" standard errors which does not require such assumptions. Assumes number of clusters goes to infinity. (The usual panel data econometric theory applies to a given N and $T \to \infty$.)

Cluster-robust inference

From Cameron & Miller (2015), A Practitioner's Guide to Cluster-Robust Inference (JHR). Example using simple OLS:

$$y_i = \beta x_i + u_i$$

- If $\text{Var}(u_i) = \sigma^2$ (homoskedasticity), $\text{Var}(\hat{\beta}) = \sigma^2 / \sum_i x_i^2$. Estimate using traditional formula:

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{\sum_i \hat{u}_i^2}{n-1}$$

- If $\text{Var}(u_i) = E(u_i^2)$ (heteroskedasticity), $\text{Var}_{\text{het}}(\hat{\beta}) = \frac{\sum_i x_i^2 E(u_i^2)}{\left(\sum_i x_i^2\right)^2}$. Estimate with White's **heteroskedasticity-robust** variance:

$$\widehat{\text{Var}}_{\text{het}}(\hat{\beta}) = \frac{\sum_i x_i^2 \hat{u}_i^2}{\left(\sum_i x_i^2\right)^2}$$

Cluster-robust inference

- If errors are correlated over i, as with some time series data, there is a **heteroskedasticity and autocorrelation-consistent** variance estimator (Newey & West):

$$\widehat{\text{Var}}_{\text{cor}}(\hat{\beta}) = \frac{\sum_i \sum_j x_i x_j \hat{u}_i \hat{u}_j}{\left(\sum_i x_i^2\right)^2}$$

- This is a generalization of White's robust variance calculation. However, requires some assumptions about the correlation structure between observations since $\sum_i x_i \hat{u}_i = 0$. A large fraction of the error correlations $E(u_i u_j)$ must be zero for this to work.
- Within-cluster (but not between) correlation is a special case of this.

Cluster-robust inference

- Suppose errors are clustered, with $E(u_i u_j) \neq 0$ if i and j are in the same cluster and $E(u_i u_j) = 0$ otherwise. Then:

$$\text{Var}_{\text{clu}}(\hat{\beta}) = \frac{\sum_i \sum_j x_i x_j E(u_i u_j) \mathbf{1}[i,j \text{ in same cluster}]}{\left(\sum_i x_i^2\right)^2}$$

For large number of clusters can estimate using:

$$\widehat{\text{Var}}_{\text{clu}}(\hat{\beta}) = \frac{\sum_i \sum_j x_i x_j \hat{u}_i \hat{u}_j \mathbf{1}[i,j \text{ in same cluster}]}{\left(\sum_i x_i^2\right)^2}$$

This is the Liang-Zeger **cluster-robust** (or, heteroskedasticity *and* cluster-robust) standard error. Simplifies to $\widehat{\text{Var}}_{\text{het}}(\hat{\beta})$ if there is only one observation per cluster.

Cluster-robust inference

From looking at the above formula we can see that $\widehat{\text{Var}}_{\text{clu}}(\hat{\beta})$ will be larger than $\widehat{\text{Var}}_{\text{het}}(\hat{\beta})$. The difference is larger:

- The more positively associated across observations are the *regressors* (via $x_i x_j$)
- The more correlated are the errors (via $u_i u_j$)
- The more observations in the same cluster (via the $\mathbf{1}[]$ indicator for observations sharing a cluster)

This is consistent with what we saw earlier regarding clustered data.

Implementation of cluster-robust inference

Once the cluster level is known, include the `vce(cluster id)` option in Stata, with *id* representing the cluster identifying variable. With `xt` commands, the option `vce(robust)` is also interpreted as cluster-robust.

Clustering decisions: Cameron & Miller

Advice: be conservative. Abadie et al. disagree—more in a moment.

- If we believe the regressors and errors may be correlated within cluster, we should think about accounting for that clustering.
- "the consensus is to be conservative and avoid bias and use bigger and more aggregate clusters when possible, up to and including the point at which there is concern about having too few clusters" (p. 16)
- If clusters are large (and there are few clusters) the cluster-robust variance formula will be a poor approximation of the true variance. Clustering is a bad idea if the number of clusters is small.
- If the regressor of interest is randomly assigned within cluster, then there is no need to account for clustering of errors.

Clustering decisions: Abadie et al.

Abadie et al. argue the "model-based approach," in which error correlation within groups is the primary motivation for clustering, is problematic.

- This motivation could be used to justify clustering for all sorts of groups (e.g., age cohorts, states)
- There is good reason *not* to take the conservative approach and simply cluster at the most aggregate level.
- One should *not* simply adjust standard errors for clustering when doing so makes a difference for inference.

Clustering decisions: Abadie et al.

Abadie et al. view clustering as a *design issue*, either a *sampling design* issue or an *experimental design* issue.

- Clustering as a sampling design issue: if the sample is a two-stage cluster sample, a clear case to be made for clustering errors.
- Clustering as an experimental design issue: if clusters of units, rather than individual units, are assigned to treatments, another case to be made for clustering errors.

Their advice: (1) is the sampling process clustered? (2) is the assignment to treatment mechanism clustered? If the answer to both is no, one should not adjust standard errors for clustering, irrespective of whether doing so would change the standard errors.