New Developments in Econometrics
Lecture 7: Cluster Sampling

Jeff Wooldridge

Cemmap Lectures, UCL, June 2009

1. The Linear Model with Cluster Effects
2. Estimation with a Small Number of Groups and Large Group Sizes
3. What if G and M_g are Both “Large”?
4. Cluster Samples with Unit-Specific Panel Data
5. Nonlinear Models

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 g, y_{gm} is a scalar response, $\mathbf{x}_g$ is a $1 \times K$ vector containing explanatory variables that vary only at the group level, and $\mathbf{z}_{gm}$ is a $1 \times L$ vector of covariates that vary within (as well as across) groups.
- The linear model with an additive error is

$$y_{gm} = \alpha + \mathbf{x}_g \boldsymbol{\beta} + \mathbf{z}_{gm} \boldsymbol{\gamma} + v_{gm} \tag{1}$$

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

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

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

$$v_{gm} = c_g + u_{gm}, m = 1, \ldots, M_g, \tag{2}$$

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)?

- Easiest sampling scheme: From a large population of relatively small clusters, we draw a large number of clusters (G), where cluster g has M_g members. For example, sampling a large number of families, classrooms, or firms from a large population.

- In the panel data setting, G is the number of cross-sectional units and M_g is the number of time periods for unit g.

Large Group Asymptotics

- The theory with $G \rightarrow \infty$ and the group sizes, M_g, fixed is well developed [White (1984), Arellano (1987)]. How should one use these methods? If

$$E(v_{gm}|\mathbf{x}_g,\mathbf{z}_{gm}) = 0 \tag{3}$$

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 $G \rightarrow \infty$ 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

$$\left(\sum_{g=1}^{G} \mathbf{W}_g' \mathbf{W}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{W}_g' \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \mathbf{W}_g\right) \left(\sum_{g=1}^{G} \mathbf{W}_g' \mathbf{W}_g\right)^{-1} \quad (4)$$

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.

- Generalized Least Squares: Strengthen the exogeneity assumption to

$$E(v_{gm}|\mathbf{x}_g, \mathbf{Z}_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{5}$$

where $\mathbf{Z}_g$ is the $M_g \times L$ matrix of unit-specific covariates.

- 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,

$$Var(\mathbf{v}_g) = \sigma_c^2 \mathbf{j}'_{M_g} \mathbf{j}_{M_g} + \sigma_u^2 \mathbf{I}_{M_g}, \tag{6}$$

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,

$$Var(\mathbf{v}_g|\mathbf{x}_g, \mathbf{Z}_g) = Var(\mathbf{v}_g). \tag{7}$$

- The random effects estimator $\hat{\boldsymbol{\lambda}}_{RE}$ is asymptotically more efficient than pooled OLS under (5), (6), and (7) as $G \rightarrow \infty$ with the M_g fixed. The RE estimates and test statistics are computed routinely by popular software packages.
- Important point is often overlooked: one can, and in many cases should, make RE inference completely robust to an unknown form of $Var(\mathbf{v}_g|\mathbf{x}_g, \mathbf{Z}_g)$, whether we have a true cluster sample or panel data.

- Cluster sample example: random coefficient model,

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + v_{gm}. \tag{8}$$

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.

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

$$y_{gm} - \bar{y}_g = (\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)\boldsymbol{\gamma} + u_{gm} - \bar{u}_g. \tag{9}$$

- Often important to allow $Var(\mathbf{u}_g|\mathbf{Z}_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. Certainly should for panel data (serial correlation), but also for cluster sampling. From linear panel data notes, 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.

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

$$\left(\sum_{g=1}^{G}\ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}\left(\sum_{g=1}^{G}\ddot{\mathbf{Z}}_g'\hat{\ddot{\mathbf{u}}}_g\hat{\ddot{\mathbf{u}}}_g'\ddot{\mathbf{Z}}_g\right)\left(\sum_{g=1}^{G}\ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}, \tag{10}$$

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.

- Even with unbalanced groups, estimating $y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \bar{\mathbf{z}}_g\boldsymbol{\xi} + e_{gm}$ by pooled OLS or RE gives the FE estimate of $\boldsymbol{\gamma}$. Can easily test $H_0 : \boldsymbol{\xi} = \mathbf{0}$.

- Above results are for “one-way clustering.” Cameron, Gelbach, and Miller (2006) have shown how to extend the formulas to multi-way clustering. For example, we have individual-level data with industry and occupation representing different clusters. So we have y_{ghm} for $g = 1,\ldots,G$, $h = 1,\ldots,H$, $m = 1,\ldots,M_{gh}$. An individual belongs to two clusters, implying some correlation across groups. Correlation across occupational groups occurs because some individuals in different occupations (indexed by g) are in the same industry (indexed by h).
- If explanatory variables vary by individual, two-way fixed effects is attractive and often eliminates the need for cluster-robust inference.

Should we Use the "Large" G Formulas with "Large" M_g?

- What if one applies robust inference in scenarios where the fixed M_g, $G \to \infty$ asymptotic analysis not realistic? Can apply recent results of Hansen (2007) to various scenarios.

- 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 still independence across groups). For example, if we have a sample of $G = 100$ schools and roughly $M_g = 100$ students per school, and we use pooled OLS leaving the school effects in the error term, we should expect the inference to have 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. (Section 2 below considers alternatives.)

- 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. The fixed G, large M_g results in Hansen (2007, Theorem 4) for cluster-robust inference apply, but are likely to be very costly: the usual variance matrix is multiplied by $G/(G-1)$ and the t statistics are approximately distributed as t_{G-1} (not standard normal).

- 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. In this case, any serial dependence in the idiosyncratic errors is assumed to be weakly dependent. The simulations in Bertrand, Duflo, and Mullainathan (2004) and Hansen (2007) verify that the fully robust cluster-robust variance matrix works well when *N* and *T* are about 50 and the idiosyncratic errors follow a stable AR(1) model.

2. Estimation with Few Groups and Large Group Sizes

- When G is small and each M_g is large, we probably 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 parameters associated with the different groups as outcomes of random draws.

- Simplest case: a single regressor that varies only by group:

$$\begin{aligned} y_{gm} &= \alpha + \beta x_g + c_g + u_{gm} \qquad (11) \\ &= \delta_g + \beta x_g + u_{gm}. \qquad (12) \end{aligned}$$

Notice how (12) is written as a model with common slope, β, but intercept, δ_g, that varies across g. Donald and Lang focus on (11), where c_g is assumed to be independent of x_g with zero mean. They use this formulation to highlight the problems of applying standard inference to (11), leaving c_g as part of the error term, $v_{gm} = c_g + u_{gm}$.

- We know that standard pooled OLS inference applied to (11) can be badly biased because it ignores the cluster correlation. Hansen's results do not apply. (We cannot use fixed effects here.)

- DL propose studying the regression in averages:

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1,\ldots,G. \tag{13}$$

If we add some strong assumptions, we can perform inference on (13) using standard methods. In particular, assume that $M_g = M$ for all g, $c_g|x_g$ ~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 $\bar{v}_g \sim Normal(0,\sigma_c^2 + \sigma_u^2/M)$. Because we assume independence across g, (13) satisfies the classical linear model assumptions.

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

$$\bar{y}_g \text{ on } 1, x_g, g = 1,\ldots,G; \tag{14}$$

identical to pooled OLS across g and m with same group sizes.

- 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 (14) 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

$$\bar{y}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \bar{\mathbf{z}}_g\boldsymbol{\gamma} + \bar{v}_g, g = 1,\ldots,G, \tag{15}$$

provided $G > K + L + 1$. If c_g is independent of $(\mathbf{x}_g, \bar{\mathbf{z}}_g)$ with a homoskedastic normal distribution, and the group sizes are large, inference can be carried out using the $t_{G-K-L-1}$ distribution. Regressions like (15) are reasonably common, at least as a check on results using disaggregated data, but usually with larger G then just a handful.

- If $G = 2$, should we give up? 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 in the DL setting, we cannot do inference (there are zero df). So, the DL setting rules out the standard comparison of means.
- Can we still obtain inference on estimated policy effects using randomized or quasi-randomized interventions when the policy effects are just identified? Not according the DL approach.

- Even when DL approach applies, should we? 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

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2, \tag{16}$$

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 $\mu_3 = \mu_4$.

- Could just define the treatment effect as

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

- 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 can be interpreted as minimum distance estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. If we use the 4×4 identity matrix as the weight matrix, we get $\hat{\beta}$ and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$.
- With large group sizes, 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.

For each group g, write

$$y_{gm} = \delta_g + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + u_{gm}. \tag{17}$$

Again, random sampling within group and independence across groups. OLS estimates withing group are $\sqrt{M_g}$ -asymptotically normal. The presence of $\mathbf{x}_g$ can be viewed as putting restrictions on the intercepts:

$$\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta}, g = 1,\ldots,G, \tag{18}$$

where we now think of x_g as fixed, observed attributes of heterogeneous groups. With K attributes 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 $\boldsymbol{\gamma}_g$.

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

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\delta}} \tag{19}$$

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

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

- Can test the overidentification restrictions. If reject, can go back to the DL approach or find more elements to put in $\mathbf{x}_g$. With large group sizes, can analyze

$$\hat{\delta}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g, g = 1,\ldots,G \tag{20}$$

as a classical linear model because $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, provided c_g is homoskedastic, normally distributed, and independent of $\mathbf{x}_g$.

3. **What if G and M_g are Both “Large”**?

- If we have a reasonably large G in addition to large M_g, we have more flexibility. In addition to ignoring the estimation error in $\hat{\delta}_g$ (because of large M_g), we can also drop the normality assumption in c_g (because, as G gets large, we can apply the central limit theorem). The regression approach still requires that the deviations, c_g, in $\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g$, are uncorrelated with $\mathbf{x}_g$. Alternatively, if we have suitable instruments, we can apply IV methods.
- Can view applications to $G = 50$ states and many individuals this way. Still unclear how big G should be.

4. 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*.
- 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

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

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 – maintaining independence across clusters – is to allow any kind of serial correlation across units or time, or both, within a cluster.

- In Stata:

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

- Compare this 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.”

- Can apply a generalized least squares analysis that makes assumptions about the components of the composite error. Typically, it is assumed that the 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 this 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 $\mathbf{\Omega}_g = \mathrm{Var}(\mathbf{v}_g)$ 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 $\mathrm{Var}(\mathbf{v}_g)$, and that is why robust inference is generally needed after RE estimation.

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

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

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

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

- Unfortunately, some routines intended for estimating HLMs (or mixed models) often assume that the structure imposed on $\mathbf{\Omega}_g$ is correct, and that $\mathrm{Var}(\mathbf{v}_g|\mathbf{R}_g) = \mathrm{Var}(\mathbf{v}_g)$. The resulting inference could be misleading, especially if serial correlation in $\{u_{gmt}\}$ is not allowed.

- 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 were allowed to change groups over time, 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

```
. 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 =
       between = 0.0292                                        avg =       4
       overall = 0.1514                                        max =

------------------------------------------------------------------------------
       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)
----------------------------------------------------------------------------
```

- Other data structures can be difficult to work with. For example, need to wonder about clustering county-level panel data at the state level with many years. Allowing unrestricted correlation across county within a state and across time (regardless of county within a state) violates the “large-*G*, small group-size” framework. In particular, cannot expect Hansen (2007) results to hold.
- If *T* is not too large (around 40 or 50), can aggregate data to the state level and apply Hansen.

5. Nonlinear Models

- Many of the issues for nonlinear models are the same as for linear models. The biggest difference is that, in many cases, standard approaches require distributional assumptions about the unobserved group effects.
- In addition, it is more difficult in nonlinear models to allow for group effects correlated with covariates, especially when group sizes differ. Can use extensions of correlated random effects approaches, but must allow conditional distributions to depend at least on the group sizes as well as group averages.

- In DL setting, no exact inference available, so need “large” M_g so that first-stage estimation error can be ignored.
- Minimum distance estimation can be employed without substantive change (but, for example, probit models are not always estimable).