A Course in Applied Econometrics
Lecture 7: Cluster Sampling

Jeff Wooldridge
IRP Lectures, UW Madison, August 2008

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. 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

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

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

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

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,

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

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,

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

- 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,

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

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:

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

- 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.

- 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. **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.
- 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).