What's New in Econometrics?
Lecture 8
Cluster and Stratified Sampling

Jeff Wooldridge
NBER Summer Institute, 2007

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 $\{(y_{gm}, x_g, z_{gm}) : m = 1, \ldots, M_g\}$ be the observable data, where M_g is the number of units in cluster g, y_{gm} is a scalar response, x_g is a $1 \times K$ vector containing explanatory variables that vary only at the group level, and 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

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma + v_{gm} \tag{1}$$

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

- Key questions: Are we primarily interested in β or γ? Does v_{gm} contain a common group effect, as in

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

where c_g is an unobserved cluster effect and u_{gm} is the idiosyncratic error? Are the regressors (x_g, z_{gm}) appropriately exogenous? How big are the group sizes (M_g) and number of groups (G)?

- Two kinds of sampling schemes. First, 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. This is like the panel data setup we have covered. 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.
- A different sampling scheme results in data sets that also can be arranged by group, but is better

interpreted in the context of sampling from different populations or different strata within a population. We stratify the population into into $G \geq 2$ nonoverlapping groups. Then, we obtain a random sample of size M_g from each group. Ideally, the group sizes are large in the population, hopefully resulting in large M_g.

Large Group Asymptotics

- The theory with $G \to \infty$ and the group sizes, M_g, fixed is well developed. How should one use these methods? If

$$E(v_{gm}|x_g, z_{gm}) = 0 \tag{3}$$

then pooled OLS estimator of y_{gm} on $1, x_g, z_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G,$ is consistent for $\lambda \equiv (\alpha, \beta', \gamma')'$ (as $G \to \infty$ with M_g fixed) and $\sqrt{G}$-asymptotically normal. In panel data case, (3)

does allow for non-strictly exogenous covariates, but only if there is no unobserved effect.

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

$$\widehat{Avar}(\hat{\lambda}_{POLS}) = \left(\sum_{g=1}^{G} W_g' W_g \right)^{-1} \left(\sum_{g=1}^{G} W_g' \hat{v}_g \hat{v}_g' W_g \right) \cdot \left(\sum_{g=1}^{G} W_g' W_g \right)^{-1} \quad (4)$$

where $\hat{v}_g$ is the $M_g \times 1$ vector of pooled OLS residuals for group g. This asymptotic variance is now computed routinely using "cluster" options.

- If we strengthen the exogeneity assumption to

$$E(v_{gm}|x_g, Z_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \quad (5)$$

where Z_g is the $M_g \times L$ matrix of unit-specific covariates, then we can use GLS. This is about the strongest assumption we can make. As discussed in the linear panel data notes, the random effects approach makes enough assumptions so that the $M_g \times M_g$ variance-covariance matrix of $v_g = (v_{g1}, v_{g2}, \ldots, v_{g,M_g})'$ has the so-called "random effects" form,

$$Var(v_g) = \sigma_c^2 j'_{M_g} j_{M_g} + \sigma_u^2 I_{M_g}, \quad (6)$$

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

$$Var(v_g|x_g, Z_g) = Var(v_g). \quad (7)$$

- The random effects estimator $\hat{\lambda}_{RE}$ is asymptotically more efficient than pooled OLS under (5), (6), and (7) as $G \to \infty$ 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 inference completely robust to an unknown form of $Var(v_g|x_g, Z_g)$, whether we have a true cluster sample or panel data.
- Cluster sample example: random coefficient model,

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma_g + v_{gm}. \quad (8)$$

By estimating a standard random effects model that

assumes common slopes γ, we effectively include $z_{gm}(\gamma_g - \gamma)$ in the idiosyncratic error; this generally creates within-group correlation because $z_{gm}(\gamma_g - \gamma)$ and $z_{gp}(\gamma_g - \gamma)$ will be correlated for $m \neq p$, conditional on Z_g.

- If we are only interested in estimating γ, the "fixed effects" (FE) or "within" estimator is attractive. The within transformation subtracts off group averages from the dependent variable and explanatory variables:

$$y_{gm} - \bar{y}_g = (z_{gm} - \bar{z}_g)\gamma + u_{gm} - \bar{u}_g, \qquad (9)$$

and this equation is estimated by pooled OLS. (Of course, the x_g get swept away by the within-group demeaning.) Often important to allow $Var(u_g|Z_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. Certainly should

for panel data (serial correlation), but also for cluster sampling. In linear panel data notes, we saw that FE can consistently estimate the average effect in the random coefficient case. But $(z_{gm} - \bar{z}_g)(\gamma_g - \gamma)$ appears in the error term. A fully robust variance matrix estimator is

$$\widehat{Avar}(\hat{\gamma}_{FE}) = \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g \right)^{-1} \left(\sum_{g=1}^{G} \ddot{Z}_g' \hat{u}_g \hat{u}_g' \ddot{Z}_g \right) \cdot \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g \right)^{-1}, \quad (10)$$

where $\ddot{Z}_g$ is the matrix of within-group deviations from means and $\hat{u}_g$ is the $M_g \times 1$ vector of fixed effects residuals. This estimator is justified with large-G asymptotics.

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? Hansen (2007) has recently derived properties of the cluster-robust variance matrix and related test statistics under various scenarios that help us more fully understand the properties of cluster robust inference across different data configurations.
- First consider how his results apply to true cluster samples. Hansen (2007, Theorem 2) shows that, with G and M_g both getting large, the usual inference based on (1.4) is valid with arbitrary correlation among the errors, v_{gm}, within each group. Because we usually think of v_{gm} as including the group effect c_g, this means that, with

large group sizes, we can obtain valid inference using the cluster-robust variance matrix, provided that G is also large. So, 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. Probably we leave the school effects in the error term because we are interested in a school-specific explanatory variable, perhaps indicating a policy change.

- Unfortunately, pooled OLS with cluster effects when G is small and group sizes are large fall outside Hansen's theoretical findings. Generally, we should not expect good properties of the cluster-robust inference with small groups and very

large group sizes when cluster effects are left in the error term.

As an example, suppose that $G = 10$ hospitals have been sampled with several hundred patients per hospital. If the explanatory variable of interest is exogenous and varies only at the hospital level, it is 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. In the next section we discuss other approaches available with small G and large M_g.

- If the explanatory variables of interest vary within group, FE is attractive for a couple of reasons. The first advantage is the usal one about allowing c_g to be arbitrarily correlated with the z_{gm}. The second advantage is that, with large M_g, we

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 asymptotic results in Theorem 4 of Hansen (2007) for cluster-robust inference apply in this case. But using cluster-robust inference is likely to be very costly in this situation: the cluster-robust variance matrix actually converges to a random variable, and t statistics based on the adjusted version of (10) – multiplied by $G/(G-1)$ – have an asymptotic t_{G-1} distribution. Therefore, while the usual or heteroskedasticity-robust inference can be based on the standard normal distribution, the cluster-robust

inference is based on the t_{G-1} distribution (and the cluster-robust standard errors may be larger than the usual standard errors).

- 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 similulations in Bertrand, Duflo, and Mullainathan (2004) and Hansen (2007) verify that the fully robust

cluster-robust variance matrix works well.

- There is some scope for applying the fully robust variance matrix estimator when N is small relative to T when unit-specific fixed effects are included. But allowing time effects causes problems in this case. Really want "large" N and T to allow for a full set of time and unit-specific effects.

2. Estimation with a Small Number of Groups and Large Group Sizes

- When G is small and each M_g is large, thinking of sampling from different strata in a population, or even different populations, makes more sense. Alternatively, we might think that the clusters were randomly drawn from a large population, but only a small number were drawn. Either way, except for the relative dimensions of G and M_g, the resulting

data set is essentially indistinguishable from a data set obtained by sampling 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 (so it seems more like the second sampling experiment). Simplest case: a single regressor that varies only by group:

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm} \quad (11)$$

$$= \delta_g + \beta x_g + u_{gm}. \quad (12)$$

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 can be badly biased because it ignores the cluster correlation. And Hansen's results do not apply. (We cannot use fixed effects here.)
- The DL solution is to study the OLS estimate in the regression "between" regresssion

$$\bar{y}_g \text{ on } 1, x_g, g = 1, \ldots, G, \tag{13}$$

which is identical to pooled OLS when the group sizes are the same. 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.

- If we add some strong assumptions, there is an exact solution to the inference problem. In addition

to assuming $M_g = M$ for all g, assume $c_g|x_g$ ~Normal$(0, \sigma_c^2)$ and assume $u_{gm}|x_g, c_g \sim$ Normal$(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim$ Normal$(0, \sigma_c^2 + \sigma_u^2/M)$ for all g. Because we assume independence across g, the equation

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G \qquad (14)$$

satisfies the classical linear model assumptions. 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 $t_{M_1+M_2+...+M_G-2}$ 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 even justified and, besides, 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 z_{gm} appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + x_g\beta + \bar{z}_g\gamma + \bar{v}_g, g = 1, \ldots, G, \qquad (15)$$

provided $G > K + L + 1$. If c_g is independent of $(x_g, \bar{z}_g)$ 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 few.

- If $G = 2$, should we give up? Suppose x_g is binary, indicating treatment and control. The DL estimate of β is the usual one: $\hat{\beta} = \bar{y}_1 - \bar{y}_0$. But in the DL setting, we cannot do inference (there are zero df). So, the DL setting rules out the standard comparison of means. It also rules out the typical setup for difference-in-differences, where there would be four groups, for the same reason.
- 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 we can apply the approach, should we? Suppose there $G = 4$ groups with groups one

and two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$). The DL approach would involve computing the averages for each group, $\bar{y}_g$, and running the regression $\bar{y}_g$ on $1, x_g$, $g = 1, \ldots, 4$. Inference is based on the t_2 distribution. The estimator $\hat{\beta}$ in this case can be written as

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

With $\hat{\beta}$ written as in (16), it is clearly it is approximately normal (for almost any underlying population distribution) provided the group sizes M_g are moderate. The DL approach would base inference on a t_2 distribution. In effect, the DL approach rejects the usual inference based on group means from large sample sizes because it may not be the case that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$.

- Equation (16) 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 (16) 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 probably generally into an MD framework, as done, for example, by Loeb and Bound (1996), who had $G = 36$ cohort-division groups and many observations per group. For each group g, write

$$y_{gm} = \delta_g + z_{gm}\gamma_g + u_{gm}. \tag{17}$$

where we assume random sampling within group and independent sampling across groups. Generally, the OLS estimates withing group are $\sqrt{M_g}$ -asymptotically normal. The presence of x_g can be viewed as putting restrictions on the intercepts, δ_g, in the separate group models in (2.8). In particular,

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G, \tag{18}$$

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 β. In the first stage, we obtain the $\hat{\delta}_g$, either by group-specific regressions or pooling to impose some common slope elements in γ_g. Let $\hat{V}$ be the $G \times G$ estimated (asymptotic) variance matrix of

the $G \times 1$ vector $\hat{\delta}$. Then the MD estimator is

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

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

- Can test the overidentification restrictions. If reject, can go back to the DL approach (or find more elements of x_g). With large group sizes, can justify analyzing

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

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

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). But, of course, we are still assuming that the deviations, c_g, in $\delta_g = \alpha + x_g\beta + c_g$, are at least uncorrelated with x_g. We can apply IV methods in this setting, though, if we have suitable instruments.

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.

Large Group Asymptotics

We can illustrate many issues using an unobserved effects probit model. Let y_{gm} be a binary response, with x_g and z_{gm}, $m = 1, \ldots, M_g, g = 1, \ldots, G$ defined as in Section 1. Assume that

$$y_{gm} = 1[\alpha + x_g\beta + z_{gm}\gamma + c_g + u_{gm} \geq 0] \quad (21)$$

$$u_{gm}|x_g, Z_g, c_g \sim Normal(0, 1) \quad (22)$$

(where $1[\cdot]$ is the indicator function). Then

$$P(y_{gm} = 1|x_g, z_{gm}, c_g) = \Phi(\alpha + x_g\beta + z_{gm}\gamma + c_g), \quad (23)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf).

- We already discussed the issue of quantities of interest, including parameters and average partial effects.
- For estimation, if we assume c_g is independent of (x_g, Z_g) with a Normal$(0, \sigma_g^2)$ distribution, then pooled probit consistently estimates the scaled coefficients (multiplied by $(1 + \sigma_c^2)^{-1/2}$). The pooled or partial maximum likelihood estimator is sometimes called a *pseudo maximum likelihood* estimator
- If we add the conditional independence assumption that $\{u_{g1}, \ldots, u_{g,M_g}\}$ are independent conditional on (x_g, Z_g, c_g) then we can use random effects probit, albeit in an unbalanced case. As we saw before, all parameters are identified.
- A challenging task, and one that appears not to

have gotten much attention for true cluster samples, is allowing correlation between the unobserved heterogeneity, c_g, and the covariates that vary within group, z_{gm}. For linear models, we know that the fixed effects estimator allows arbitrary correlation, and does not restrict the within-cluster dependence of $\{u_{g1}, \ldots, u_{g,M_g}\}$. Unfortunately, allowing correlation between c_g and $(z_{g1}, z_{g2}, \ldots, z_{gM_g})$. Even if we assume normality and exchangeability in the mean, we must at least allow for difference variances:

$$c_g|(z_{g1}, \ldots, z_{g,M_g}) \sim \text{Normal}(\eta + \bar{z}_g\xi, \sigma^2_{a,M_g}), \qquad (24)$$

where σ^2_{a,M_g} denotes a different variance for each group size, M_g. Then the marginal distributions are

$$P(y_{gm} = 1|Z_g) = \Phi[(\eta + z_{gm}\gamma + \bar{z}_g\xi)/(1 + \sigma^2_{a,M_g})^{1/2}]. \qquad (25)$$

Any estimation must account for the different variances for different group sizes. With very large G and little variation in M_g, we might just use the unrestricted estimates $(\hat{\eta}_{M_g}, \hat{\xi}_{M_g}, \hat{\gamma}_{M_g})$, estimate the APEs for each group size, and then average these across group size. But more work needs to be done to see if such an approach loses too much in terms of efficiency.

- The methods of Altonji and Matzkin (2005) can be applied to allow more flexible relationships between c_g and $\bar{z}_g$, say, or other functions of $\{z_{g1}, \ldots, z_{g,M_g}\}$
- The logit conditional MLE applies to cluster samples without change, so we can estimate parameters without restricting $D(c_g | z_{g1}, \ldots, z_{gM_g})$.

A Small Number of Groups and Large Group

Sizes

- Unlike in the linear case, for nonlinear models exact inference is unavailable even under the strongest set of assumptions. But approximate inference is if the group sizes M_g are reasonably large
- With small G and random sampling of $\{(y_{gm}, z_{gm}) : m = 1, \ldots, M_g\}$ write

$$P(y_{gm} = 1|z_{gm}) = \Phi(\delta_g + z_{gm}\gamma_g) \tag{26}$$

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G. \tag{27}$$

Using a minimum distance approach, in a first step we estimate a series of G probits (or pool across g to impose common slopes), obtain the group "fixed effects" $\hat{\delta}_g, g = 1, \ldots, G$. Then, we impose the restrictions in (26) using linear MD estimation –

just as before. Now, the asymptotic variances $\widehat{Avar}(\hat{\delta}_g)$ come from the probits.

- The DL approach also applies with large M_g but we again must assume $\delta_g = \alpha + x_g\beta + c_g$ where c_g is independent of x_g and homoskedastic normal. As in the linear case, we just use classical linear model inference in the equation $\hat{\delta}_g = \alpha + x_g\beta + c_g$, provide $G > K + 1$.
- The same holds for virtually any nonlinear model with an index structure: the second step is linear regression.

Large G and Large M_g

- As in the linear case, more flexibility is afforded if G is somewhat large along with large M_g because we can relax the normality assumption in c_g in analyzing the regression $\hat{\delta}_g$ on $1, x_g, g = 1, \ldots, G$.

- A version of the method proposed by Berry, Levinsohn, and Pakes (1995) for estimating structural models using both individual-level and product-level data, or market-level data, or both can be treated in the large G, large M_g framework, where g indexes good or market and m indexes individuals within a market. Suppose there is a single good across many markets (so G is large) and we have many individuals within each market (the M_g are large). The main difference with what we have done up until now is that BLP must allow correlation between x_g (particularly, price) and the unobserved product attributes in market g, c_g. So, the second step involves instrumental variables estimation of the equation $\hat{\delta}_g = \alpha + x_g\beta + c_g$. If the M_g are large enough to ignore estimation error in

the $\hat{\delta}_g$, then the second step can be analyzed just like a standard cross section IV estimation. (BLP actually derive an asymptotic variance that accounts for estimation error in $\hat{\delta}_g$, along with the uncertainty in c_g, and simulation error – which comes from difficult computational problems in the first stage.)