New Developments in Econometrics **Cemmap**, **UCL**, **June 2009**
Lectures 7&9, **Wednesday**, **June 17th**, **9:30-10:30** & **12:15-13:15**

Cluster and Stratified Sampling

These notes consider estimation and inference with cluster samples and samples obtained by stratifying the population. The main focus is on true cluster samples, although the case of applying cluster-sample methods to panel data is treated, including recent work where the sizes of the cross section and time series are similar. Wooldridge (2003, extended version 2006) contains a survey, but more recent work is discussed here.

1. The Linear Model with Cluster Effects

This section considers linear models estimated using cluster samples (of which a panel data set is a special case). 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.

1.1 Specification of the Model

The linear model with an additive error is

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma + v_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G. \quad (1.1)$$

Our approach to estimation and inference in equation (1.1) depends on several factors, including whether we are interested in the effects of aggregate variables (β) or individual-specific variables (γ). Plus, we need to make assumptions about the error terms. In the context of pure cluster sampling, an important issue is whether the v_{gm} contain a common

group effect that can be separated in an additive fashion, as in

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

where c_g is an unobserved cluster effect and u_{gm} is the idiosyncratic error. (In the statistics literature, (1.1) and (1.2) are referred to as a "hierarchical linear model.") One important issue is whether the explanatory variables in (1.1) can be taken to be appropriately exogenous. Under (1.2), exogeneity issues are usefully broken down by separately considering c_g and u_{gm}.

Throughout we assume that the sampling scheme generates observations that are independent across g. This assumption can be restrictive, particularly when the clusters are large geographical units. We do not consider problems of "spatial correlation" across clusters, although, as we will see, fixed effects estimators have advantages in such settings.

We treat two kinds of sampling schemes. The simplest case also allows the most flexibility for robust inference: from a large population of relatively small clusters, we draw a large number of clusters (G), where cluster g has M_g members. This setup is appropriate, for example, in randomly sampling a large number of families, classrooms, or firms from a large population. The key feature is that the number of groups is large enough relative to the group sizes so that we can allow essentially unrestricted within-cluster correlation. Randomly sampling a large number of clusters also applies to many panel data sets, where the cross-sectional population size is large (say, individuals, firms, even cities or counties) and the number of time periods is relatively small. 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 are 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. This is the perspective for the "small G" case in Section 1.3.

1.2. Large Group Asymptotics

In this section I review methods and estimators justified when the asymptotic approximations theory is with The theory with $G \to \infty$ and the group sizes, M_g, fixed is well developed; see, for example, White (1984), Arellano (1987), and Wooldridge (2002, Chapters 10, 11). Here, the emphasis is on how one might wish to use methods robust to cluster sampling even when it is not so obvious.

First suppose that the covariates satisfy

$$E(v_{gm}|x_g, z_{gm}) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{1.3}$$

For consistency, we can, of course, get by with zero correlation assumptions, but we use (1.3) for convenience because it meshes well with assumptions concerning conditional second moments. Importantly, the exogeneity in (1.3) only requires that z_{gm} and v_{gm} are uncorrelated. In particular, it does not specify assumptions concerning v_{gm} and z_{gp} for $m \neq p$. As we saw in the linear panel data notes, (1.3) is called the "contemporaneous exogeneity" assumption when m represents time. Allowing for correlation between v_{gm} and $z_{gp}, m \neq p$ is useful for some panel data applications and possibly even cluster samples (if the covariates of one unit can affect another unit's response). Under (1.3) and a standard rank condition on the covariates, the pooled OLS estimator, where we regress 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.

Without more assumptions, a robust variance matrix is needed to account for correlation within clusters or heteroskedasticity in $Var(v_{gm}|x_g, z_{gm})$, or both. When v_{gm} has the form in

(1.2), the amount of within-cluster correlation can be substantial, which means the usual OLS standard errors can be very misleading (and, in most cases, systematically too small). 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

$$\widehat{Avar}(\hat{\lambda}_{POLS}) = \left(\sum_{g=1}^{G} W_g' W_g\right)^{-1} \left(\sum_{g=1}^{G} W_g' \hat{v}_g \hat{v}_g' W_g\right) \left(\sum_{g=1}^{G} W_g' W_g\right)^{-1} \tag{1.4}$$

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.

Pooled OLS estimation of the parameters in (1.1) ignores the within-cluster correlation of the v_{gm}; even if the procedure is consistent (again, with $G \to \infty$ and the M_g fixed), the POLS estimators can be very inefficient. If we strengthen the exogeneity assumption to

$$E(v_{gm}|x_g, Z_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{1.5}$$

where Z_g is the $M_g \times L$ matrix of unit-specific covariates, then we can exploit the presence of c_g in (1.2) in a generalized least squares (GLS) analysis. With true cluster samples, (1.5) rules out the covariates from one member of the cluster affecting the outcomes on another, holding own covariates fixed. In the panel data case, (1.5) is the strict exogeneity assumption on $\{z_{gm} : m = 1, \ldots, M_g\}$ that we discussed in the linear panel data notes The standard 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,

$$Var(v_g) = \sigma_c^2 j_{M_g}' j_{M_g} + \sigma_u^2 I_{M_g}, \tag{1.6}$$

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. In the standard setup, we also make the "system homoskedasticity" assumption,

$$Var(v_g|x_g, Z_g) = Var(v_g). \tag{1.7}$$

It is important to understand the role of assumption (1,7): it implies that the conditional variance-covariance matrix is the same as the unconditional variance-covariance matrix, but it does not restrict $Var(v_g)$; it can be any $M_g \times M_g$ matrix under (1.7). The particular random effects structure on $Var(v_g)$ is given by (1.6). Under (1.6) and (1.7), the resulting GLS estimator is the well-known random effects (RE) estimator.

The random effects estimator $\hat{\lambda}_{RE}$ is asymptotically more efficient than pooled OLS under (1.5), (1.6), and (1.7) as $G \to \infty$ with the M_g fixed. The RE estimates and test statistics are computed routinely by popular software packages. Nevertheless, an important point is often overlooked in applications of RE: one can, and in many cases should, make inference completely robust to an unknown form of $Var(v_g|x_g, Z_g)$.

The idea in obtaining a fully robust variance matrix of RE is straightforward and we essentially discussed it in the notes on nonlinear panel data models. Even if $Var(v_g|x_g, Z_g)$ does not have the RE form, the RE estimator is still consistent and $\sqrt{G}$-asymptotically normal under (1.5), and it is likely to be more efficient than pooled OLS. Yet we should recognize that the RE second moment assumptions can be violated without causing inconsistency in the RE estimator. For panel data applications, making inference robust to serial correlation in the idiosyncratic errors, especially with more than a few time periods, can be very important. Further, within-group correlation in the idiosyncratic errors can arise for cluster samples, too, especially if underlying (1.1) is a random coefficient model,

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma_g + v_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{1.8}$$

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. Also, the idiosyncratic error will have heteroskedasticity that is a function of z_{gm}. Nevertheless, if we assume $E(\gamma_g|X_g, Z_g) = E(\gamma_g) \equiv \gamma$ along with (1.5), the random effects estimator still consistently estimates the average slopes, γ. Therefore, in applying random effects to panel data or cluster samples, it is sensible (with large G) to make the variance estimator of random effects robust to arbitrary heteroskedasticity and within-group correlation.

One way to see what the robust variance matrix looks like for $\hat{\lambda}_{RE}$ is to use the pooled OLS characterization of RE on a transformed set of data. For each g, define $\hat{\theta}_g = 1 - \{1/[1 + M_g(\hat{\sigma}_c^2/\hat{\sigma}_u^2)]\}^{1/2}$, where $\hat{\sigma}_c^2$ and $\hat{\sigma}_u^2$ are estimators of the variances of c_g and u_{gm}, respectively. Then the RE estimator is identical to the pooled OLS estimator of

$$y_{gm} - \hat{\theta}_g \bar{y}_g \text{ on } (1 - \hat{\theta}_g), (1 - \hat{\theta}_g)x_g, z_{gm} - \hat{\theta}_g \bar{z}_g, m = 1, \ldots, M_g; g = 1, \ldots, G; \tag{1.9}$$

see, for example, Hsiao (2003). For fully robust inference, we can just apply the fully robust variance matrix estimator in (1.4) but on the transformed data.

With panel data, it may make sense to estimate an unrestricted version of $Var(v_g)$, especially if G is large. Even in that case, it makes sense to obtain a variance matrix robust to $Var(v_{gm}|x_g, Z_g) \neq Var(v_g)$, as the GEE literature does. One can also specify a particular structure, such as an AR(1) model for the idiosyncratic errors. In any case, fully robust inference is still a good idea.

In summary, with large G and relatively small M_g, it makes sense to compute fully robust variance estimators even if we apply a GLS procedure that allows $Var(v_g)$ to be unrestricted. Nothing ever guarantees $Var(v_{gm}|x_g, Z_g) = Var(v_g)$. Because RE imposes a specific structure

on $Var(v_g)$, there is a strong case for making RE inference fully robust. When c_g is in the error term, it is even more critical to use robust inference when using pooled OLS because the usual standard errors ignore within-cluster correlation entirely.

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:

$$y_{gm} - \bar{y}_g = (z_{gm} - \bar{z}_g)\gamma + u_{gm} - \bar{u}_g, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{1.10}$$

and this equation is estimated by pooled OLS. (Of course, the x_g get swept away by the within-group demeaning.) Under a full set of "fixed effects" assumptions – which, unlike pooled OLS and random effects, allows arbitrary correlation between c_g and the z_{gm} – inference is straightforward using standard software. Nevertheless, analogous to the random effects case, it is often important to allow $Var(u_g|Z_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. For panel data, the idiosyncratic errors can always have serial correlation or heteroskedasticity, and it is easy to guard against these problems in inference. Reasons for wanting a fully robust variance matrix estimator for FE applied to cluster samples are similar to the RE case. For example, if we start with the model (1.8) then $(z_{gm} - \bar{z}_g)(\gamma_g - \gamma)$ appears in the error term. As we discussed in the linear panel data notes, the FE estimator is still consistent if $E(\gamma_g|z_{g1} - \bar{z}_g, \ldots, z_{g,M_g} - \bar{z}_g) = E(\gamma_g) = \gamma$, an assumption that allows γ_g to be correlated with $\bar{z}_g$. Nevertheless, u_{gm}, u_{gp} will be correlated for $m \neq p$. A fully robust variance matrix estimator is

$$\widehat{Avar}(\hat{\gamma}_{FE}) = \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g\right)^{-1} \left(\sum_{g=1}^{G} \ddot{Z}_g' \hat{u}_g \hat{u}_g' \ddot{Z}_g\right) \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g\right)^{-1}, \tag{1.11}$$

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.

One benefit of a fixed effects approach, especially in the standard model with constant slopes but c_g in the composite error term, is that no adjustments are necessary if the c_g are correlated across groups. When the groups represent different geographical units, we might expect correlation across groups close to each other. If we think such correlation is largely captured through the unobserved effect c_g, then its elimination via the within transformation effectively solves the problem. If we use pooled OLS or a random effects approach, we would have to deal with spatial correlation across g, in addition to within-group correlation, and this is a difficult problem.

The previous discussion extends immediately to instrumental variables versions of all estimators. With large G, one can afford to make pooled two stage least squares (2SLS), random effects 2SLS, and fixed effects 2SLS robust to arbitrary within-cluster correlation and heteroskedasticity. Also, more efficient estimation is possible by applying generalized method of moments (GMM); again, GMM is justified with large G.

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

Until recently, the standard errors and test statistics obtained from pooled OLS, random effects, and fixed effects were known to be valid only as $G \to \infty$ with each M_g fixed. As a practical matter, that means one should have lots of small groups. Consider again formula (1.4), for pooled OLS, when the cluster effect, c_g, is left in the error term. With a large number of groups and small group sizes, we can get good estimates of the within-cluster correlations – technically, of the cluster correlations of the cross products of the regressors and errors – even if they are unrestricted, and that is why the robust variance matrix is consistent as $G \to \infty$ with

M_g fixed. In fact, in this scenario, one loses nothing in terms of asymptotic local power (with local alternatives shrinking to zero at the rate $G^{-1/2}$) if c_g is not present. In other words, based on first-order asymptotic analysis, there is no cost to being fully robust to any kind of within-group correlation or heteroskedasticity. These arguments apply equally to panel data sets with a large number of cross sections and relatively few time periods, whether or not the idiosyncratic errors are serially correlated.

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: the proper asymptotic analysis would be with G fixed, $M_g \to \infty$, and persistent within-cluster correlation (because of the presence of c_g in the

error). Hansen (2007, Theorem 4) is aimed at panel data where the time series dependence satisfies strong mixing assumptions, that is, where the correlation within each group g is weakly dependent. Even in this case, the variance matrix in (1.4) must be multiplied by $G/(G-1)$ and inference based on the t_{G-1} distribution. (Conveniently, this adjustment is standard in Stata's calculation of cluster-robust variance matrices.) Interestingly, Hansen finds, in simulations, that with $G = 10$ and $M_g = 50$ for all g, using the adjusted robust variance matrix estimator with critical values from the t_{G-1} distribution produces fairly small size distortions. But the simulation study is special (one covariate whose variance is as large as the variance of the composite error).

We probably 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. Interestingly, 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 (1.11) – that is, 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). With small G, inference based on cluster-robust statistics could be very conservative when it need not be. (Also, Hansen's Theorem 4 is not completely general, and may not apply with heterogeneous sampling across groups.)

In summary, for true cluster sample applications, cluster-robust inference using pooled OLS delivers statistics with proper size when G and M_g are both moderately large, but they should probably be avoided with large M_g and small G. When cluster fixed effects are included, the usual inference is often valid, perhaps made robust to heteroskedasticity, and is likely to be much more powerful than cluster-robust inference.

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. Unlike in the true cluster sampling case, it makes sense to treat the idiosyncratic errors as correlated with only weakly dependent. But Hansen's (2007, Theorem 4) does not allow time fixed effects (because the asymptotics is with fixed N and $T \to \infty$, and so the inclusion of time fixed effects means adding more and more parameters without more cross section data to estimate them). As a practical matter, it seems dangerous to rely on omitting time effects or unit effects with panel data. Hansen's result that applies in this case requires N and T both getting large.

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

We can summarize the findings of the previous section as follows: fully robust inference justified for large G (N) and small M_g (T) can also be relied on when M_g (T) is also large, provided G (N) is also reasonably large. However, whether or not we leave cluster (unobserved) effects in the error term, there are good reasons not to rely on cluster-robust inference when G (N) is small and M_g (T) is large.

In this section, we describe approaches to inference when G is small and the M_g are large. These results apply primarily to the true cluster sample case, although we will draw on them for difference-in-differences frameworks using pooled cross sections in a later set of notes.

In the large G, small M_g case, it often makes sense to think of sampling a large number of groups from a large population of clusters, where each cluster is relatively small. When G is small while each M_g is large, this thought experiment needs to be modified. For most data sets with small G, a stratified sampling scheme makes more sense: we have defined G groups in the population, and we obtain our data by randomly sampling from each group. As before, M_g is the sample size for group g. Except for the relative dimensions of G and M_g, the resulting data

set is essentially indistinguishable from that described in Section 1.2.

The problem of proper inference when M_g is large relative to G was brought to light by Moulton (1990), and has been recently studied by Donald and Lang (2007). DL focus on applications that seem well described by the stratified sampling scheme, but their approach seems to imply a different sampling experiment. In particular, they treat the parameters associated with the different groups as outcomes of random draws. One way to think about the sampling in this case is that a small number of groups is drawn from a (large) population of potential groups; therefore, the parameters common to all members of the group can be viewed as random. Given the groups, samples are then obtained via random sampling within each group.

To illustrate the issues considered by Donald and Lang, consider the simplest case, with a single regressor that varies only by group:

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm} \tag{2.1}$$

$$= \delta_g + \beta x_g + u_{gm}, \; m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{2.2}$$

Notice how (2.2) is written as a model with common slope, β, but intercept, δ_g, that varies across g. Donald and Lang focus on (2.1), 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 (2.1), leaving c_g as part of the composite error term, $v_{gm} = c_g + u_{gm}$. We know this is a bad idea even in the large G, small M_g case, as it ignores the persistent correlation in the errors within each group. Further, from the discussion of Hansen's (2007) results, using cluster-robust inference when G is small is likely to produce poor inference.

One way to see the problem with the usual inference in applying standard inference is to note that when $M_g = M$ for all $g = 1, \ldots, G$, the pooled OLS estimator, $\hat{\beta}$, is identical to the

"between" estimator obtained from the regression

$$\bar{y}_g \text{ on } 1, x_g, g = 1, \ldots, G. \tag{2.3}$$

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 $v_{gm} \equiv c_g + u_{gm}$. The presence of c_g means new observations within group do not provide additional information for estimating β beyond how they affect the group average, $\bar{y}_g$. In effect, we only have G useful pieces of information.

If we add some strong assumptions, there is a solution to the inference problem. In addition to assuming $M_g = M$ for all g, assume $c_g | x_g \sim \text{Normal}(0, \sigma_c^2)$ and assume $u_{gm} | x_g, c_g \sim \text{Normal}(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim \text{Normal}(0, \sigma_c^2 + \sigma_u^2/M)$ for all g. Because we assume independence across g, the equation

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G \tag{2.4}$$

satisfies the classical linear model assumptions. Therefore, we can use inference based on the t_{G-2} distribution to test hypotheses about β, provided $G > 2$. When G is very 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. (Interestingly, if we use cluster-robust inference and apply Hansen's results – even though they do not apply – we would use the t_{G-1} distribution.)

When x_g is a $1 \times K$ vector, we need $G > K + 1$ to use the t_{G-K-1} distribution for inference. [In Moulton (1990), $G = 50$ states and x_g contains 17 elements]

As pointed out by DL, performing the correct inference in the presence of c_g is *not* just a matter of correcting the pooled OLS standard errors for cluster correlation – something that does not appear to be valid for small G, anyway – or using the RE estimator. In the case of

common group sizes, there is only estimator: pooled OLS, random effects, and the between regression in (2.4) all lead to the *same* $\hat{\beta}$. The regression in (2.4), by using the t_{G-K-1} distribution, yields inference with appropriate size.

We can apply the DL method without normality of the u_{gm} if the common group size M is large: by the central limit theorem, $\bar{u}_g$ will be approximately normally distributed very generally. Then, because c_g is normally distributed, we can treat $\bar{v}_g$ as approximately normal with constant variance. Further, even if the group sizes differ across g, for very large group sizes $\bar{u}_g$ will be a negligible part of $\bar{v}_g$: $Var(\bar{v}_g) = \sigma_c^2 + \sigma_u^2/M_g$. Provided c_g is normally distributed and it dominates $\bar{v}_g$, a classical linear model analysis on (2.4) should be roughly valid.

The broadest applicability of DL's setup is when the average of the idiosyncratic errors, $\bar{u}_g$, can be ignored – either because σ_u^2 is small relative to σ_c^2, M_g is large, or both. In fact, applying DL with different group sizes or nonnormality of the u_{gm} is identical to ignoring the estimation error in the sample averages, $\bar{y}_g$. In other words, it is as if we are analyzing the simple regression $\mu_g = \alpha + \beta x_g + c_g$ using the classical linear model assumptions (where $\bar{y}_g$ is used in place of the unknown group mean, μ_g). With small G, we need to further assume c_g is normally distributed.

If z_{gm} appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + x_g\beta + \bar{z}_g\gamma + \bar{v}_g, g = 1, \ldots, G, \tag{2.5}$$

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.

The DL solution to the inference problem with small G is pretty common as a strategy to check robustness of results obtained from cluster samples, but often it is implemented with somewhat large G (say, $G = 50$). Often with cluster samples one estimates the parameters using the disaggregated data and also the averaged data. When some covariates that vary within cluster, using averaged data is generally inefficient. But it does mean that standard errors need not be made robust to within-cluster correlation. We now know that if G is reasonably large and the group sizes not too large, the cluster-robust inference can be acceptable. DL point out that with small G one should think about simply using the group averages in a classical linear model analysis.

For small G and large M_g, inference obtained from analyzing (2.5) as a classical linear model will be very conservative in the absense of a cluster effect. Perhaps in some cases it is desirable to inject this kind of uncertainty, but it rules out some widely-used staples of policy analysis. For example, suppose we have two populations (maybe men and women, two different cities, or a treatment and a control group) with means $\mu_g, g = 1, 2$, and we would like to obtain a confidence interval for their difference. In almost all cases, it makes sense to view the data as being two random samples, one from each subgroup of the population. Under random sampling from each group, and assuming normality and equal population variances, the usual comparison-of-means statistic is distributed exactly as $t_{M_1+M_2-2}$ under the null hypothesis of equal population means. (Or, we can construct an exact 95% confidence interval of the difference in population means.) With even moderate sizes for M_1 and M_2, the $t_{M_1+M_2-2}$ distribution is close to the standard normal distribution. Plus, we can relax normality to obtain approximately valid inference, and it is easy to adjust the t statistic to allow for different population variances. With a controlled experiment the standard difference-in-means analysis

is often quite convincing. Yet we cannot even study this estimator in the DL setup because $G = 2$. The problem can be seen from (2.2): in effect, we have three parameters, δ_1, δ_2, and β, but only two observations.

DL criticize Card and Krueger (1994) for comparing mean wage changes of fast-food workers across two states because Card and Krueger fail to account for the state effect (New Jersery or Pennsylvania), c_g, in the composite error, v_{gm}. But the DL criticism in the $G = 2$ case is no different from a common question raised for any difference-in-differences analyses: How can we be sure that any observed difference in means is due entirely to the policy change? To characterize the problem as failing to account for an unobserved group effect is not necessarily helpful.

To further study the $G = 2$ case, recall that c_g is independent of x_g with mean zero. In other words, the expected deviation in using the simple comparison-of-means estimator is zero. In effect, it estimates

$$\mu_2 - \mu_1 = (\delta_2 + \beta) - \delta_1 = (\alpha + c_2 + \beta) - (\alpha + c_1) = \beta + (c_2 - c_1). \tag{2.6}$$

Under the DL assumptions, $c_2 - c_1$ has mean zero, and so estimating it should not bias the analysis. DL work under the assumption that β is the parameter of interest, but, if the experiment is properly randomized – as is maintained by DL – it is harmless to include the c_g in the estimated effect.

Consider now a case where the DL approach can be applied. Assume there are $G = 4$ 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

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2. \tag{2.7}$$

(The pooled OLS regression using the disaggregated data results in the weighted average $(p_3\bar{y}_3 + p_4\bar{y}_4) - (p_1\bar{y}_1 + p_2\bar{y}_2)$, where $p_1 = M_1/(M_1 + M_2)$, $p_2 = M_2/(M_1 + M_2)$, $p_3 = M_3/(M_3 + M_4)$, and $p_4 = M_4/(M_3 + M_4)$ are the relative proportions within the control and treatment groups, respectively.) With $\hat{\beta}$ written as in (2.7), we are left to wonder why we need to use the t_2 distribution for inference. Each $\bar{y}_g$ is usually obtained from a large sample – $M_g = 30$ or so is usually sufficient for approximate normality of the standardized mean – and so $\hat{\beta}$, when properly standardized, has an approximate standard normal distribution quite generally.

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 $\mu_3 = \mu_4$. In other words, the control group may be heterogeneous as might be the treatment group. But this in itself does not invalidate standard inference applied to (2.7). In fact, if we *define* the object of inference as

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2, \tag{2.8}$$

which is an average treatment effect of sorts, then $\hat{\beta}$ is consistent for τ and (when properly scaled) asymptotically normal as the M_g get large.

Equation (2.7) hints at a different way to view the small G, large M_g setup. In this particular application, we estimate two parameters, α and β, given four moments that we can estimate with the data. The OLS estimates from (2.4) in this case are 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 $\hat{\beta}$ as in (2.7) and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$. Using the MD approach, we see there are two overidentifying restrictions, which are easily tested. But even if

we reject them, it simply implies at least one pair of means within each of the control and treatment groups is different.

With large group sizes, and whether or not G is especially large, we can put the general problem 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

$$y_{gm} = \delta_g + z_{gm}\gamma_g + u_{gm}, m = 1, \ldots, M_g, \tag{2.9}$$

where we assume random sampling within group and independent sampling across groups. We make the standard assumptions for OLS to be consistent (as $M_g \to \infty$) and $\sqrt{M_g}$-asymptotically normal; see, for example, Wooldridge (2002, Chapter 4). The presence of group-level variables x_g in a "structural" model can be viewed as putting restrictions on the intercepts, δ_g, in the separate group models in (2.9). In particular,

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G, \tag{2.10}$$

where we 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. If M_g is large enough to estimate the δ_g precisely, a simple two-step estimation strategy suggests itself. First, obtain the $\hat{\delta}_g$, along with $\hat{\gamma}_g$, from an OLS regression within each group. If $G = K + 1$ then, typically, we can solve for $\hat{\theta} \equiv (\hat{\alpha}, \hat{\beta}')'$ uniquely in terms of the $G \times 1$ vector $\hat{\delta}$:. $\hat{\theta} = X^{-1}\hat{\delta}$, where X is the $(K + 1) \times (K + 1)$ matrix with g^{th} row $(1, x_g)$. If $G > K + 1$ then, in a second step, we can use a minimum distance approach, as described in Wooldridge (2002, Section 14.6). If we use as the weighting matrix I_G, the $G \times G$ identity matrix, then the minimum distance estimator can be computed from the OLS regression

$$\hat{\delta}_g \text{ on } 1, x_g, g = 1, \ldots, G. \tag{2.10}$$

Under asymptotics such that $M_g = \rho_g M$ where $0 < \rho_g \leq 1$ and $M \to \infty$, the minimum distance estimator $\hat{\theta}$ is consistent and $\sqrt{M}$-asymptotically normal. Still, this particular minimum distance estimator is asymptotically inefficient except under strong assumptions. Because the samples are assumed to be independent, it is not appreciably more difficult to obtain the efficient minimum distance (MD) estimator, also called the "minimum chi-square" estimator.

First consider the case where z_{gm} does not appear in the first stage estimation, so that the $\hat{\delta}_g$ is just $\bar{y}_g$, the sample mean for group g. Let $\hat{\sigma}_g^2$ denote the usual sample variance for group g. Because the $\bar{y}_g$ are independent across g, the efficient MD estimator uses a diagonal weighting matrix. As a computational device, the minimum chi-square estimator can be computed by using the weighted least squares (WLS) version of (2.11), where group g is weighted by $M_g/\hat{\sigma}_g^2$ (groups that have more data and smaller variance receive greater weight). Conveniently, the reported t statistics from the WLS regression are asymptotically standard normal as the group sizes M_g get large. (With fixed G, the WLS nature of the estimation is just a computational device; the standard asymptotic analysis of the WLS estimator has $G \to \infty$.). The minimum distance approach works with small G provided $G \geq K + 1$ and each M_g is large enough so that normality is a good approximation to the distribution of the (properly scaled) sample average within each group.

If z_{gm} is present in the first-stage estimation, we use as the minimum chi-square weights the inverses of the asymptotic variances for the g intercepts in the separate G regressions. With large M_g, we might make these fully robust to heteroskedasticity in $E(u_{gm}^2|z_{gm})$ using the White (1980) sandwich variance estimator. At a minimum we would want to allow different σ_g^2 even if we assume homoskedasticity within groups. Once we have the $\widehat{Avar}(\hat{\delta}_g)$ – which are just the

squared reported standard errors for the $\hat{\delta}_g$ – we use as weights $1/\widehat{Avar}(\hat{\delta}_g)$ in the computationally simple WLS procedure. We are still using independence across g in obtaining a diagonal weighting matrix in the MD estimation.

An important by-product of the WLS regression is a minimum chi-square statistic that can be used to test the $G - K - 1$ overidentifying restrictions. The statistic is easily obtained as the weighted sum of squared residuals, say SSR_w. Under the null hypothesis in (2.10), $SSR_w \overset{a}{\sim} \chi^2_{G-K-1}$ as the group sizes, M_g, get large. If we reject H_0 at a reasonably small significance level, the x_g are not sufficient for characterizing the changing intercepts across groups. If we fail to reject H_0, we can have some confidence in our specification, and perform inference using the standard normal distribution for t statistics for testing linear combinations of the population averages.

We might also be interested in how one of the slopes in γ_g depends on the group features, x_g. Then, we simple replace $\hat{\delta}_g$ with, say $\hat{\gamma}_{g1}$, the slope on the first element of z_{gm}. Naturally, we would use $1/\widehat{Avar}(\hat{\gamma}_{g1})$ as the weights in the MD estimation.

The minimum distance approach can also be applied if we impose $\gamma_g = \gamma$ for all g, as in the original model (1). Obtaining the $\hat{\delta}_g$ themselves is easy: run the pooled regression

$$y_{gm} \text{ on } d1_g, d2_g, \ldots, dG_g, z_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G \tag{2.11}$$

where $d1_g, d2_g, \ldots, dG_g$ are group dummy variables. Using the $\hat{\delta}_g$ from the pooled regression (2.12) in MD estimation is complicated by the fact that the $\hat{\delta}_g$ are no longer asymptotically independent; in fact, $\hat{\delta}_g = \bar{y}_g - \bar{z}_g\hat{\gamma}$, where $\hat{\gamma}$ is the vector of common slopes, and the presence of $\hat{\gamma}$ induces correlation among the intercept estimators. 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

$\hat{\theta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\hat{\delta}$ and its estimated asymptotic variance is $(X'\hat{V}^{-1}X)^{-1}$. If the OLS regression (2.11) is used, or the WLS version, the resulting standard errors will be incorrect because they ignore the across group correlation in the estimators. (With large group sizes the errors might be small; see the next section.)

Intermediate approaches are available, too. Loeb and Bound (1996) (LB for short) allow different group intercepts and group-specific slopes on education, but impose common slopes on demographic and family background variable. The main group-level covariate is the student-teacher ratio. Thus, LB are interested in seeing how the student-teach ratio affects the relationship between test scores and education levels. LB use both the unweighted estimator and the weighted estimator and find that the results differ in unimportant ways. Because they impose common slopes on a set of regressors, the estimated slopes on education (say $\hat{\gamma}_{g1}$) are not asymptotically independent, and perhaps using a nondiagonal estimated variance matrix $\hat{V}$ (which would be 36×36 in this case) is more appropriate; but see Section 3.

If we reject the overidentifying restrictions, we are essentially concluding that $\delta_g = \alpha + x_g\beta + c_g,$ where c_g can be interpreted as the deviation from the restrictions in (2.10) for group *g*. As *G* increases relative to *K*, the likelihood of rejecting the restrictions increases. One possibility is to apply the Donald and Lang approach, which is to analyze the OLS regression in (2.11) in the context of the classical linear model (CLM), where inference is based on the t_{G-K-1} distribution. Why is a CLM analysis justified? Since $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, we can ingore the estimation error in $\hat{\delta}_g$ for large M_g (Recall that the same "large M_g" assumption underlies the minimum distance approach.) Then, it is as if we are estimating the equation $\delta_g = \alpha + x_g\beta + c_g, g = 1, \ldots, G$ by OLS. If the c_g are drawn from a normal distribution, classical analysis is applicable because c_g is assumed to be independent of

x_g. This approach is desirable when one cannot, or does not want to, find group-level observables that completely determine the δ_g. It is predicated on the assumption that the other factors in c_g are not systematically related to x_g, a reasonable assumption if, say, x_g is a randomly assigned treatment at the group level, a case considered by Angrist and Lavy (2002).

Beyond the treatment effect case, the issue of how to define parameters of interest appears complicated, and deserves further study. In the example with $G = 4$ and two control and two treatment groups, it can be shown that defining the treatment effect as (2.8) is the same as defining the parameters of interest as $\theta = (X'X)^{-1}X'\delta$, where X is the 4×2 matrix

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{2.12}$$

and $\beta = \tau$ is the second element of θ. Generally, if it makes sense to define the object of interest as $\theta = (X'X)^{-1}X'\delta$, and if we estimate θ as $\hat{\theta} = (X'X)^{-1}X'\hat{\delta}$, then $\sqrt{M}(\hat{\theta} - \theta)$ inherits its asymptotic distribution from that of $\sqrt{M}(\hat{\delta} - \delta)$, where we assume, as before, that $M_g = \rho_g M$ with $0 < \rho_g \leq 1$ and $M \to \infty$. Such a setting implies

$$\widehat{Avar(\hat{\theta})} = (X'X)^{-1}X'[\widehat{Avar(\hat{\delta})}]X(X'X)^{-1}. \tag{2.13}$$

3. What if G and M_g are Both "Large"?

Section 1 reviewed methods appropriate for a large number of groups and relatively small group sizes. Section 2 considered two approaches appropriate for large group sizes and a small number of groups. The DL and minimum distance approaches use the large group sizes

assumption differently: in its most applicable setting, DL use the large M_g assumption to ignore the first-stage estimation error entirely, with all uncertainty coming from unobserved group effects, while the asymptotics underlying the MD approach is based on applying the central limit theorem within each group. Not surprisingly, more flexibility is afforded if G and M_g are both "large."

For example, suppose we adopt the DL specification (with an unobserved cluster effect),

$$\delta_g = \alpha + x_g\beta + c_g, g = 1, \ldots, G, \tag{3.1}$$

and $G = 50$ (say, states in the U.S.). Further, assume first that the group sizes are large enough (or the cluster effects are so strong) that the first-stage estimation error can be ignored. Then, it matters not whether we impose some common slopes or run separate regressions for each group (state) in the first stage estimation. In the second step, we can treat the group-specific intercepts, $\hat{\delta}_g, g = 1, \ldots, G$, as independent "observations" to be used in the second stage. This means we apply regression (2.10) and apply the usual inference procedures. The difference now is that with $G = 50$, the usual t statistics have some robustness to nonnormality of the c_g, assuming the CLT approximation works well With small G, the exact inference was based on normality of the c_g.

Loeb and Bound (1996), with $G = 38$, essentially use regression (2.10), but with estimated slopes as the dependent variable in place of estimated intercepts. As mentioned in Section 2, LB impose some common slopes across groups, which means the estimated parameters are dependent across group. The minimum distance approach without cluster effects is one way to account for the dependence. Alternatively, one can simply adopt the DL perspective and just assume the estimation error is swamped by c_g; then standard OLS analysis is approximately

justfied.

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. For the small G case, we offer extensions of the Donald and Lang (2007) approach (with large group sizes) and the minimum distance approach.

Rather than using a general, abstract setting, the issues for nonlinear models are easily illustrated with the probit model. Wooldridge (2006) considers other models (which are also covered in the nonlinear panel data notes).

4.1. 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

$$y_{gm} = 1[\alpha + x_g\beta + z_{gm}\gamma + c_g + u_{gm} \geq 0] \tag{4.1}$$

$$u_{gm}|x_g, Z_g, c_g \sim Normal(0,1) \tag{4.2}$$

(where $1[\cdot]$ is the indicator function). Equations (4.1) and (4.2) imply

$$P(y_{gm} = 1|x_g, z_{gm}, c_g) = P(y_{gm} = 1|x_g, Z_g, c_g) = \Phi(\alpha + x_g\beta + z_{gm}\gamma + c_g), \tag{4.3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf). We assume throughout that only z_{gm} affects the response probability of y_{gm} conditional on x_g and c_g; the

outcomes of z_{gp} for $p \neq m$ are assumed not to matter. This is captured in (4.3). For pooled methods we could relax this restriction (as in the linear case), but, with the presence of c_g, this affords little generality in practice.

As in nonlinear panel data models, the presence of c_g in (4.3) raises several important issues, including how we estimate quantities of interest. As in the panel data case, we have some interest in estimating average partial or marginal effects. For example, if the first element of x_g is continuous,

$$\frac{\partial P(y_{gm} = 1|x_g, z_{gm}, c_g)}{\partial x_{g1}} = \beta_1 \phi(\alpha + x_g\beta + z_{gm}\gamma + c_g), \tag{4.4}$$

where $\phi(\cdot)$ is the standard normal density function. If

$$c_g|x_g, Z_g \sim Normal(0, \sigma_c^2), \tag{4.5}$$

where the zero mean is without loss of generality because (4.1) contains an intercept, α, then the APEs are obtained from the function

$$P(y_{gm} = 1|x_g, Z_g) = \Phi[(\alpha + x_g\beta + z_{gm}\gamma)/(1 + \sigma_c^2)^{1/2}] \equiv \Phi(\alpha_c + x_g\beta_c + z_{gm}\gamma_c), \tag{4.6}$$

where $\alpha_c = \alpha/(1 + \sigma_c^2)^{1/2}$, and so on. Conveniently, the scaled coefficients are exactly the coefficients estimated by using a simple pooled probit procedure. So, for estimating the average partial effects, pooled probit is perfectly acceptable. With large G and small group sizes, we can easily make the standard errors and test statistics robust to arbitarary within group correlation using standard sandwich variance estimators (robust to within-cluster correlation).

Some authors prefer to call procedures such as pooled probit applied to cluster samples *pseudo maximum likelihood*. Unfortunately, this term is used in contexts where only the

conditional mean is correctly specified in the context of the linear exponential family. Wooldridge (2002, Chapter 13) calls such methods *partial maximum likelihood* to emphasize that we have partially specified a distribution, namely the marginal distribution of y_{gm} given (x_g, Z_m), without specifying a joint distribution $(y_{g1}, \ldots, y_{g,M_g})$ conditional on (x_g, Z_g).

If we supplement (4.1),.(4.2), and (4.5) with

$$\{u_{g1}, \ldots, u_{g,M_g}\} \text{ are independent conditional on } (x_g, Z_g, c_g) \tag{4.7}$$

then we have the so-called *random effects probi*t model. Under the RE probit assumptions, α, β, γ and σ_c^2 are all identified, and estimable by MLE, which means we can estimate the APEs as well as the partial effects evaluated at the mean of c_g, which is zero. We can also compute partial effects at other values of c_g that we might select from the normal distribution with estimated standard deviation σ_c. The details for random effects probit in the balanced panel data case are given in Wooldridge (2002, Chapter 15). The unbalanced case is similar.

As we discussed in the nonlinear panel data notes, minimum distance estimator or generalized estimating equations can be used to obtain estimators (of the scaled coefficients) more efficient than pooled probit but just as robust. (Remember, the RE probit estimator has no known robustness properties to violation of assumption (4.7).)

A very 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 notational simplicity, we assume there are no group-level controls in the model, but these can always be added.) For linear models, we know that the within or 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 $(z_{g1}, z_{g2}, \ldots, z_{gM})$ is much more difficult in nonlinear models. In the balanced case, where the group sizes M_g are the same for all g, the Chamberlain (1980) device can be used:

$$c_g | Z_g \sim Normal(\eta + \bar{z}_g \xi, \sigma_a^2), \tag{4.8}$$

where σ_a^2 is the conditional variance $Var(c_g|Z_g)$. If we use all random effects probit assumptions but with (4.8) in place of (4.5), then we obtain a simple extension of the RE probit model: simply add the group averages, $\bar{z}_g$, as a set of additional explanatory variables. This is identical to the balanced panel case we covered earlier. The marginal distributions are

$$P(y_{gm} = 1|Z_g) = \Phi[(\eta + z_{gm}\gamma + \bar{z}_g\xi)/(1 + \sigma_a^2)^{1/2}] \equiv \Phi(\eta_a + z_{gm}\gamma_a + \bar{z}_g\xi_a) \tag{4.9}$$

where now the coefficients are scaled by a function of the conditional variance. This is just as in the case of a balanced panel, and all calculations, including those for APEs, follow immediately.

The Chamberlain-Mundlak needs to be modified for the unbalanced case. [One possibility is to discard observations and balance the cluster sample under the assumption that the cluster sizes are exogenous, and that might be desirable if there is not much variation in the cluster sizes.] An alternative is to use the cluster setup and assuming a kind of exchangeability assumption concerning the correlation between the cluster effect and the covariates. At a minimum, (4.8) should be modified to allow the variances to depend on the cluster size, M_g. Under restrictive assumptions, such as joint normality of $(c_g, z_{g1}, \ldots, z_{g,M_g})$, with the z_{gm} independent and identically distributed within a cluster, one can derive $Var(c_g|Z_g)$. But these are strong assumptions. We might just assume

$$c_g | (z_{g1}, \ldots, z_{g,M_g}) \sim \text{Normal}(\eta + \bar{z}_g \xi, \sigma_{a,M_g}^2), \tag{4.10}$$

where σ_{a,M_g}^2 denotes a different variance for each group size, M_g. Then the marginal

distributions are

$$P(y_{gm} = 1|Z_g) = \Phi[(\eta + z_{gm}\gamma + \bar{z}_g\xi)/(1 + \sigma^2_{a,M_g})^{1/2}]. \tag{4.11}$$

Equation (4.11) can be estimated by pooled probit, allowing for different group variances. (A normalization is also required.) A simpler approach is to estimate a different set of parameters, $(\eta_{M_g}, \xi_{M_g}, \gamma_{M_g})$, for each group size, and then to impose the restrictions in (4.11) by minimum distance estimation. 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) – see also Wooldridge (2005) – can be applied. A completely nonparametric approach is based on

$$P(y_{gm} = 1|Z_g, c_g) = P(y_{gm} = 1|z_{gm}, c_g) \equiv F(z_{gm}, c_g) \tag{4.12}$$

and

$$D(c_g|z_{g1}, z_{g2}, \ldots, z_{g,M_g}) = D(c_g|\bar{z}_g). \tag{4.13}$$

Define $H_g(z_{gm}, \bar{z}_g) = P(y_{gm} = 1|z_{gm}, \bar{z}_g)$. As discussed in the nonlinear panel data notes, under (4.12) and (4.13), it can be show that the APEs are obtained from

$$E_{\bar{z}_g}[H_g(z, \bar{z}_g)]. \tag{4.14}$$

If the group sizes differ, $H_g(\cdot, \cdot)$ generally depends on g. If there are relatively few group sizes, it makes sense to estimate the $H_g(\cdot, \cdot)$ separately for each group size M_g. Then, the APEs can be estimated from

$$G^{-1}\sum_{g=1}^{G}\hat{H}_g(z,\bar{z}_g). \tag{4.15}$$

As discussed before, as a practical matter we might just use flexible parametric models, such as probit with flexible functional forms.

Other strategies are available for estimating APEs. We can apply "fixed effects probit" to cluster samples just as with panel data and treat the c_g as parameters to estimate in

$$P(y_{gm} = 1|Z_g, c_g) = P(y_{gm} = 1|z_{gm}, c_g) = \Phi(z_{gm}\gamma + c_g). \tag{4.16}$$

The same issues arise as in the panel data case, except with true cluster samples the conditional indepdence assumption likely is more reasonable than in the panel data case. With small group sizes M_g (say, siblings or short panel data sets), treating the c_g as parameters to estimate creates an incidental parameters problem. As before, we might use

$$G^{-1}\sum_{g=1}^{G}\Phi(z\hat{\gamma} + \hat{c}_g), \tag{4.17}$$

to estimate the APEs.

The logit conditional MLE can be applied to cluster samples essentially without change, which means we can estimate the parameters, γ, without restricting $D(c_g|Z_g)$. This is especially convenient in the unbalanced case.

4.2. 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. Nevertheless, if the group sizes M_g are reasonably large, we can extend the DL approach to nonlinear models and obtain approximate inference. In addition, the the minimum distance approach carries over essentially without change.

We can apply the methods to any nonlinear model that has an index structure – which includes all of the common ones, and many other models besides, but we again consider the probit case. With small G and random sampling of $\{(y_{gm}, z_{gm}) : m = 1, \ldots, M_g\}$ within each g, write

$$P(y_{gm} = 1|z_{gm}) = \Phi(\delta_g + z_{gm}\gamma_g), m = 1, \ldots, M_g \tag{4.18}$$

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G. \tag{4.19}$$

As with the linear model, we assume the intercept, δ_g in (4.18), is a function of the group features x_g. With the M_g moderately large, we can get good estimates of the δ_g. The $\hat{\delta}_g, g = 1, \ldots, G$, are easily obtained by estimating a separate probit for each group. Or, we can impose common γ_g and just estimate different group intercepts (sometimes called "group fixed effects").

Under (4.19), we can apply the minimum distance approach just as before. Let $\widehat{Avar}(\hat{\delta}_g)$ denote the estimated asymptotic variances of the $\hat{\delta}_g$ (so these shrink to zero at the rate $1/M_g$). If the $\hat{\delta}_g$ are obtained from G separate probits, they are independent, and the $\widehat{Avar}(\hat{\delta}_g)$ are all we need. As in the linear case, if a pooled method is used, the $G \times G$ matrix $\widehat{Avar}(\hat{\delta})$ should be obtained as the weighting matrix. For binary response, we use the usual MLE estimated variance. If we are using fractional probit for a fractional response, these would be from a sandwich estimate of the asymptotic variance. In the case where the $\hat{\delta}_g$ are obtained from separate probits, we can obtain the minimum distance estimates as the WLS estimates from

$$\hat{\delta}_g \text{ on } 1, x_g, g = 1, \ldots, G \tag{4.20}$$

using weights $1/\widehat{Avar}(\hat{\delta}_g)$ are used as the weights. This is the efficient minimum distance

estimator and, conveniently, the proper asymptotic standard errors are reported from the WLS estimation (even though we are doing large M_g, not large G, asymptotics.) Generally, we can write the MD estimator as in the linear case, $\hat{\theta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\hat{\delta},$ where $\hat{\delta}$ is the $G \times 1$ vector of $\hat{\delta}_g$ and $\hat{V} = \widehat{Avar}(\hat{\delta})$. The overidentification test is obtained exactly as in the linear case: there are $G - K - 1$ degrees-of-freedom in the chi-square distribution.

The same cautions about using the overidentification test to reject the minimum distance approach apply here as well. In particular, in the treatment effect setup, where x_g is zero or one, we might reject a comparision of means across multiple groups simply because the means within the control or within the treatment group differ, or both. It might make sense to define the treatment effect as the difference in averages between treatment and control, or use weighted averages, without worrying about whether the means are the same. (We consider an alternative, namely, using data to choose a synthetic control from a set of potentil control groups, the the notes on difference-in-differences.)

If we reject the overidentification restrictions, we can adapt Donald and Lang (2007) and treat

$$\hat{\delta}_g = \alpha + x_g\beta + error_g, g = 1, \ldots, G \tag{4.21}$$

as approximately satisfying the classical linear model assumptions, provided $G > K + 1$, just as before. As in the linear case, this approach is justified if $\delta_g = \alpha + x_g\beta + c_g$ with c_g independent of x_g and c_g drawn from a homoskedastic normal distribution. It assumes that we can ignore the estimation error in $\hat{\delta}_g$, based on $\hat{\delta}_g = \delta_g + O(1/\sqrt{M_g})$. Because the DL approach ignores the estimation error in $\hat{\delta}_g$, it is unchanged if one imposes some constant slopes across the groups, as with the linear model.

Once we have estimated α and β, the estimated effect on the response probability can be obtained by averaging the response probability for a given x:

$$G^{-1}\sum_{g=1}^{G}\left(M_g^{-1}\sum_{m=1}^{M_g}\Phi(\hat{\alpha}+x\hat{\beta}+z_{gm}\hat{\gamma}_g)\right), \tag{4.22}$$

where derivatives or differences with respect to the elements of x can be computed. Here, the minimum distance approach has an important advantage over the DL approach: the finite sample properties of (4.22) are viritually impossible to obtain, whereas the large-M_g asymptotics underlying minimum distance would be straightforward using the delta method. How the bootstrap might work in this situation is an interesting question.

Particularly with binary response problems, the two-step methods described here are problematical when the response does not vary within group. For example, suppose that x_g is a binary treatment – equal to one for receiving a voucher to attend college – and y_{gm} is an indicator of attending college. Each group is a high school class, say. If some high schools have all students attend college, one cannot use probit (or logit) of y_{gm} on $z_{gm}, m = 1,\ldots,M_g$. A linear regression returns zero slope coefficients and intercept equal to unity. Of course, if randomization occurs at the group level – that is, x_g is independent of group attributes – then it is not necessary to control for the z_{gm}. Instead, the within-group averages can be used in a simple minimum distance approach. In this case, as y_{gm} is binary, the DL approximation will not be valid, as the CLM assumptions will not even approximately hold in the model $\bar{y}_g = \alpha + x_g\beta + e_g$ (because $\bar{y}_g$ is always a fraction regardless of the size of M_g).

4.3. 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. If we can ignore the estimation error in the $\hat{\delta}_g$, then, in implementing the DL approach –

with or without common slopes imposed in the first stage – one gains robustness to nonnormality of c_g if G is large enough so that $G^{-1/2}\sum_{g=1}^{G} c_g$ and $G^{-1/2}\sum_{g=1}^{G} x_g c_g$ are approximately normally distributed. The second step is the same as in the linear model case: $\hat{\delta}_g$ is regressed on 1, x_g, $g = 1,\ldots,G$; one can use heteroskedasticity-robust inference with large G to partly account for the estimation error in the $\hat{\delta}_g$.

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. BLP's original application was where g indexes different automobile models. Petrin and Train (2003) cover the case of about 170 television markets and four TV services. To handle this case, assume that H products are available in each market. Therefore, we now think of δ_g as an H-vector for each g, and so is c_g. The main difference here with the previous setup is that, for reasons discussed in BLP and Petrin and Train, we must allow the c_{gh} to be correlated with the x_{gh} (which contains the price of good j in market g, in addition to product/market attributes). BLP propose a two-step estimation strategy. In the first step, a choice model, such as multinomial logit, is estimated using the individual-level data pooled across markets. The key estimates are what we call the $\hat{\delta}_g$ – the market "fixed effects." Typically, most or all of the "slope" parameters in the multinomial logit estimation are assumed to be constant across g, although, with many individuals per market, that is not necessary.

In the second step, the $\hat{\delta}_{gh}$ are used in place of δ_{gh} in the market/good-level equation

$$\delta_{gh} = \alpha + x_{gh}\beta + c_{gh}, \mathrm{h} = 1,\ldots,H;\ g = 1,\ldots,G, \tag{4.23}$$

where, say, w_g is a set of instruments for x_{gh}. (Typically, w_g varies only by market, g, and not by good, h.) This allows for market/good-specific unobservables in the individual choice equations to be correlated with prices. If we could observe the δ_{gh}, then (4.23) would be a standard problem in IV estimation for a cross section system of equations, provided G is large enough to invoke the law of large numbers and central limit theorem. Replacing δ_g with $\hat{\delta}_g$ is justified if the M_g are large because the variance of c_g will dominate that of the $\hat{\delta}_g$. Further, any correlation induced in the $\hat{\delta}_g$ by pooling in the first-stage estimation shrinks to zero at the rate $1/M$, where we can think of M as the average group size. In other words, we just apply, say, 2SLS in the second step.

Ignoring the estimation in $\hat{\delta}_g$, efficient estimation is obtained by writing the system of equations as

$$\hat{\delta}_g \approx X_g\theta + c_g \tag{4.24}$$

where X_g is the $J \times (K+1)$ matrix of attributes (including an intercept and prices). Because (4.24) is a system of equations with instruments $I_J \otimes w_g$, we can use the 3SLS estimator or GMM to efficiently account for the correlation across $\{c_{gh} : h = 1,\ldots,H\}$.

5. Estimation of Population Parameters with Stratified Samples

We now provide a brief, modern treatment of estimation with stratified samples. The emphasis here is in estimation parameters from a population that has been stratified. Typically, with stratified sampling, some segments of the population are over- or underrepresented by the sampling scheme. Fortunately, if we know enough information about the stratification scheme, we can often modify standard econometric methods and consistently estimate population parameters.

There are two common types of stratified sampling, standard stratified (SS) sampling and variable probability (VP) sampling. A third type of sampling, typically called multinomial sampling, is practically indistinguishable from SS sampling, but it generates a random sample from a modified population (thereby simplifying certain theoretical analyses). See Cosslett (1993), Imbens (1992), Imbens and Lancaster (1996), and Wooldridge (1999) for further discussion. We focus on SS and VP sampling here.

SS sampling begins by partitioning the sample space (set of possible outcomes), say W, into G non-overlapping, exhaustive groups, $\{W_g : g = 1,\dots G\}$. Then, a random sample is taken from each group g, say $\{w_{gi} : i = 1,\dots,N_g\}$, where N_g is the number of observations drawn from stratum g and $N = N_1 + N_2 +\dots+N_G$ is the total number of observations. If w is a random vector representing the population, and taking values in W, then each random draw from stratum g has the same distribution as w conditional on w belonging to W_g:

$$D(w_{gi}) = D(w|w \in W_g),\ i = 1,\dots,N_g.$$

Therefore, the resulting sample across all strata consists of independent but not identically distributed observations. Unless we are told, we have no way of knowing that our data came from SS sampling.

What if we want to estimate the mean of w from an SS sample? It turns out we cannot get an unbiased or consistent estimator of unless we have some additional information. Typically, the information comes in the form of population frequencies for each of the strata. Specifically, let $\pi_g = P(w \in W_g)$ be the probability that w falls into stratum g; the π_g are often called the "aggregate shares."

If we know the π_g (or can consistently estimate them), then $\mu_w = E(w)$ is identified by a weighted average of the expected values for the strata:

$$\mu_w = \pi_1 E(w|w \in W_1) + \ldots + \pi_G E(w|w \in W_G). \tag{5.1}$$

Because we can estimate each of the conditional means using the random sample from the appropriate stratum, an unbiased estimator of is simply

$$\hat{\mu}_w = \pi_1 \bar{w}_1 + \pi_2 \bar{w}_2 \ldots + \pi_G \bar{w}_G, \tag{5.2}$$

where $\bar{w}_g$ is the sample average from stratum g. As the strata sample sizes grow, $\hat{\mu}_w$ is also a consistent estimator of μ_w. The variance of $\hat{\mu}_w$ is easily obtained because of independence withing and between strata:

$$Var(\hat{\mu}_w) = \pi_1^2 Var(\bar{w}_1) + \ldots + \pi_G^2 Var(\bar{w}_G). \tag{5.3}$$

Because $Var(\bar{w}_g) = \sigma_g^2/N_g$, each of the variances can be estimated in an unbiased fashion by using the usual unbiased variance estimator,

$$\hat{\sigma}_g^2 = (N_g - 1)^{-1} \sum_{i=1}^{N_g} (w_{gi} - \bar{w}_g)^2. \tag{5.4}$$

Sometimes it is useful to have a formula for $\hat{\mu}_w$ that shows it is a weighted average across all observations:

$$\begin{aligned} \hat{\mu}_w &= (\pi_1/h_1) N^{-1} \sum_{i=1}^{N_1} w_{1i} + \ldots + (\pi_G/h_G) N^{-1} \sum_{i=1}^{N_G} w_{Gi} \\ &= N^{-1} \sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) w_i \end{aligned} \tag{5.5}$$

where $h_g = N_g/N$ is the fraction of observations in stratum g and in (5.5) we drop the strata index on the observations and use the stratum for observation i, g_i, to pick out the appropriate weight, π_{g_i}/h_{g_i}. Formula (5.5) is useful because the sampling weights associated with SS samples are reported as (π_{g_i}/h_{g_i}), and so applying these weights in averaging across all N

produces an unbiased, consistent estimator. Nevertheless, the large sample properties of estimators from SS samples are properly derived from (5.2) and its extensions.

A different sampling scheme is usually called *variable probability (VP) sampling*, which is more convenient for telephone or email surveys, where little, if anything, is known ahead of time about those being contacted. With VP sampling, each stratum g is assigned a nonzero sampling probability, p_g. Now, a random draw w_i is taking from the population, and it is kept with probability p_g if $w_i \in W_g$. With VP sampling, the population is sampled N times. Often N is not reported with VP samples (although, as we discuss latter, knowing how many times each stratum was sampled can improve efficiency). Instead, we know how many data points were kept, and we call this M. Because of the randomness in whether an observation is kept, M is properly viewed as a random variable. With VP sampling, it is handy for each draw from the population to have a selection indicator, s_i, which is one of observation i is kept (and then its stratum is also known). Then $M = \sum_{i=1}^{N} s_i$. Let z_i be a G-vector of stratum indicators, and let $p(z_i) = p_1 z_{i1} + \ldots + p_G z_{iG}$ be the function that delivers the sampling probability for any random draw i.

A key assumption for VP sampling is that conditional on being in stratum g, the chance of keeping an observation is p_g. Statistically, this means that, conditional on z_i, s_i and w_i are independent. Using this assumption, we can show, just as in the treatment effect case,

$$E[(s_i/p(z_i))w_i] = E(w_i); \tag{5.6}$$

that is, weighting a selected observation by the inverse of its sampling probability allows us to recover the population mean. (We will use a more general version of this result when we discuss missing data general. Like estimating counterfactual means in program evaluation, VP

sampling is, in effect, a missing data problem. But it is usually treated along with other stratified sampling schemes.) It follows that

$$N^{-1}\sum_{i=1}^{N}(s_i/p(z_i))w_i \tag{5.7}$$

is a consistent estimator of $E(w_i)$. We can also write this as

$$(M/N)M^{-1}\sum_{i=1}^{N}(s_i/p(z_i))w_i; \tag{5.8}$$

if we define weights as $\hat{v}_i = \hat{\rho}/p(z_i)$ where $\hat{\rho} = M/N$ is the fraction of observations retained from the sampling scheme, then (5.8) is $M^{-1}\sum_{i=1}^{M}\hat{v}_i w_i,$ where only the observed points are included in the sum. Thus, like in the SS case, we can write the esimator for the mean under VP sampling as a weighted average of the observed data points. In the VP case, the weight is (an estimate of) the probability of keeping an observation, $\rho = P(s_i = 1)$, over the probability that an observation in its stratum is kept. If $p_g < \rho$, the observations for stratum g are underpresented in the eventual sample (asymptotically), and they receive weight greater than one.

In both the SS and VP cases, one may replace the number of observed data points in the average with the sum of the weights described in each case.

Virtually any estimation method can be used with SS or VP sampled data. Wooldridge (1999, 2001) covers M-estimation for the VP and SS cases, respectively. This includes a wide variety of estimators, including least squares, MLE, and quasi-MLE. There are several interesting findings concerning asymptotic efficiency and estimating the asymptotic variances. Consider the problem of linear regression for simplicity; analogous claims hold for MLE, NLS, and many other estimators. The model in the population is

$$y = \mathbf{x}\boldsymbol{\beta} + u, \tag{5.9}$$

where $\boldsymbol{\beta}$ may index the conditional mean, but consistency follows from $E(\mathbf{x}'u) = \mathbf{0}$. Consider SS sampling. Then a consistent estimator $\hat{\boldsymbol{\beta}}$ is obtained from the "weighted" least squares problem

$$\min_{\mathbf{b}} \sum_{i=1}^{N} v_i \cdot (y_i - \mathbf{x}_i\mathbf{b})^2, \tag{5.10}$$

where $v_i = \pi_{g_i}/h_{g_i}$ is the weight for observation i. Remember, the weighting used here is not to solve any heteroskedasticity problem; it is to reweight the sample in order to consistently estimate the population parameter $\boldsymbol{\beta}$.

One possibility for performing inference on $\hat{\boldsymbol{\beta}}$ is to use the White (1980) robust sandwich estimator and associated statistics. This estimator is routinely reported by regression packages when sampling weights are included. In fact, sometimes this estimator is consistent for $Avar\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. There are two assumptions that imply consistency of this widely used variance matrix estimator: (i) $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, so that we are actually estimating a conditional mean; and (ii) the strata are determined by the explanatory variables, $\mathbf{x}$. It turns out that when the White estimator is not consistent, it is actually conservative. In other words, the White estimator converges to a matrix that is larger, in the matrix sense, than the correct asymptotic variance.

To obtain the correct asymptotic variance, we need to use a more detailed formulation of the estimation problem, which is

$$\min_{\mathbf{b}} \left\{ \sum_{g=1}^{G} \pi_g \left[N_g^{-1} \sum_{i=1}^{N_g} (y_{gi} - \mathbf{x}_{gi}\mathbf{b})^2 \right] \right\} \tag{5.11}$$

so that we are minimizing the a weighted average sum of squared residuals. (Equation (5.11) is a consistent estimator of $E[(y - \mathbf{xb})^2]$, and we know the population value, $\boldsymbol{\beta}$, minimizes $E[(y - \mathbf{xb})^2]$.) Using this formulation – actually, the M-estimator version of it – Wooldridge (2001) showed that a consistent estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}\widehat{Avar(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} = &\left[\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})\mathbf{x}_i'\mathbf{x}_i\right]^{-1} \\ &\cdot\left\{\sum_{g=1}^{G}(\pi_g/h_g)^2\left[\sum_{i=1}^{N_g}(\mathbf{x}_{gi}'\hat{u}_{gi} - \overline{\mathbf{x}_g'\hat{u}_g})(\mathbf{x}_{gi}'\hat{u}_{gi} - \overline{\mathbf{x}_g'\hat{u}_g})'\right]\right\} \\ &\cdot\left[\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})\mathbf{x}_i'\mathbf{x}_i\right]^{-1}.\end{aligned} \tag{5.12}$$

This formula looks a bit daunting, but, it can be seen that the outer parts of the sandwich are identical to the usual White sandwich estimator. The difference is in the middle. The usual estimator ignores the information on the strata of the observations, which is the same as dropping the within-strata averages, $\overline{\mathbf{x}_g'\hat{u}_g}$. Because a smaller sum of squared residuals (in a matrix sense) is obtained by subtracting off the same average – rather than centering around zero – the matrix in (5.12) is smaller than the usual White matrix. That happens asymptotically, too, provided the means $E(\mathbf{x}'u|\mathbf{w} \in W_g)$, where $\mathbf{w} = (\mathbf{x},y)$, are nonzero. So, it is the difference between subtracting off within-strata averages and not that produces the more precise inference with stratified sampled data. Econometrics packages, such as Stata, will compute (5.12) provided strata membership is included along with the weights. If only the weights are provided, the larger asymptotic variance is computed.

One case where there is no gain from subtracting within-strata means is when $E(u|\mathbf{x}) = 0$ and $\mathbf{w} \in W_g$ is the same as $\mathbf{x} \in X_g$ for some partition of the regressor space. In fact, if we add

the homoskedasticity assumption $Var(u|\mathbf{x}) = \sigma^2$, then the weighted estmator is less efficient than the unweighted estimator, which, of course, is also consistent because $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ and stratification is based on $\mathbf{x}$. So, the cost to weighting when the classical linear model assumptions hold and stratification is exogenous is in terms of efficiency loss.

Some argue that even if stratification is based on $\mathbf{x}$, one should use the weighted estimator. The argument is based on consistently estimating the linear projection, $L(y|\mathbf{x})$, even if the conditional mean is not linear. If we can only assume $L(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, then the weighted estimator consistently estiimates $\boldsymbol{\beta}$ whether or not the stratification is based on $\mathbf{x}$. The unweighted estimator does not consistently estimate $\boldsymbol{\beta}$ in either case.

The previous discussion applies to nonlinear least squares and maximum likelihood problems, and others. Now, to exploit the stratification, strata means should be subtracted from the gradient of the objective function when computing the asymptotic variance. This requires knowing the stratum and its weight for each observation. A conservative estimate is obtained by the Huber-White sandwich form for misspecified MLE – but with sampling weights. This is the proper formula for, say, MLE if the conditional density $f(y|\mathbf{x},\boldsymbol{\theta})$ is correctly specified and stratification is based on $\mathbf{x}$. But in that case the unweighted MLE is fully efficient, and the usual variance matrix estimators can be used. The weighted estimator does consistently estimate the solution to the population problem $\min_{\mathbf{t}} E[\log f(y|\mathbf{x},\mathbf{t})]$ if the density is misspecified, and that is valuable in some situations.

The above findings have analogs for VP sampling. One interesting finding is that while the Huber-White sandwich matrix applied to the weighted objective function (weighted by the $1/p_g$) *is* consistent when the known p_g are used, an asymptotically more efficient estimator is available when the retention frequencies, $\hat{p}_g = M_g/N_g$, are observed, where M_g is the number

of observed data points in stratum g and N_g is the number of times stratum g was sampled. We always know M_g if we are given a stratum indicator with each observation. Generally, N_g might not be known. If it is, we should use the $\hat{p}_g$ in place of p_g. Results such as this are discussed in Imbens (1992), Imbens and Lancaster (1996), and Wooldridge (1999, 2007). The VP sampling example in Wooldridge (2007) can be used to show that the following matrix is valid:

$$\begin{aligned}\widehat{Avar}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = &\left[\sum_{i=1}^{M} \mathbf{x}_i'\mathbf{x}_i/\hat{p}_{g_i}\right]^{-1} \\ &\cdot \left\{\sum_{g=1}^{G} \hat{p}_g^{-2}\left[\sum_{i=1}^{M_g}(\mathbf{x}_{gi}'\hat{u}_{gi} - \overline{\mathbf{x}_g'\hat{u}_g})(\mathbf{x}_{gi}'\hat{u}_{gi} - \overline{\mathbf{x}_g'\hat{u}_g})'\right]\right\} \\ &\cdot \left[\sum_{i=1}^{M} \mathbf{x}_i'\mathbf{x}_i/\hat{p}_{g_i}\right]^{-1},\end{aligned} \tag{5.13}$$

where, remember, M_g is the number of observed data points in stratum g, and the above sums are over the observed data points. This formula is essentially the same as (5.12) in that the quantities are weighted so that the sample represents the population and $\mathbf{x}_{gi}'\hat{u}_{gi}$ are centered about the within-strata means. If we use the known sampling weights, we drop $\mathbf{x}_g'\hat{u}_g$ from (5.13). If $E(u|\mathbf{x}) = 0$ and the sampling is exogenous, we also drop this term because $E(\mathbf{x}'u|\mathbf{w} \in W_g) = \mathbf{0}$ for all g, and this is whether or not we estimate the p_g. See Wooldridge (2007) for how these same claims carry over to general nonlinear models and estimation methods.

6. Clustering and Stratified Sampling

Often, survey data are often characterized by clustering and variable probability sampling.

For example, suppose that g represents the primary sampling unit (say, city) and individuals or families (indexed by m) are sampled within each PSU with probability p_{gm}. Consider the problem of regression using such a data set. If $\hat{\boldsymbol{\beta}}$ is the pooled OLS estimator across PSUs and individuals, then its variance is estimated as

$$\left(\sum_{g=1}^{G}\sum_{m=1}^{M_g}\mathbf{x}_{gm}'\mathbf{x}_{gm}/p_{gm}\right)^{-1}\cdot\left[\sum_{g=1}^{G}\sum_{m=1}^{M_g}\sum_{r=1}^{M_g}\hat{u}_{gm}\hat{u}_{gr}\mathbf{x}_{gm}'\mathbf{x}_{gr}/(p_{gm}p_{gr})\right]\left(\sum_{g=1}^{G}\sum_{m=1}^{M_g}\mathbf{x}_{gm}'\mathbf{x}_{gm}/p_{gm}\right)^{-1}. \quad (6.1)$$

The middle of the sandwich accounts for cluster correlation along with unequal sampling probabilities. If the probabilities are estimated using retention frequencies, (6.1) is conservative, as before.

Multi-stage sampling schemes introduce even more complications. Consider the following setup, closely related to Bhattacharya (2005). Let there be S strata (e.g., states in the U.S.), exhaustive and mutually exclusive. Within stratum s, there are C_s clusters (e.g., neighborhoods). We require some sort of large-sample approximation. Therefore, we assume that in each stratum a large number of clusters is sampled, with replacement. (The assumption of with replacement can be relaxed, but is not important of the number of clusters samples, N_s, is "large.") The setup allows for arbitrary correlation (say, across households) within each cluster.

Within stratum s and cluster c, let there be M_{sc} total units (household or individuals). Therefore, the total number of units in the population is

$$M = \sum_{s=1}^{S}\sum_{c=1}^{C_s}M_{sc}. \quad (6.2)$$

Let z be a variable whose mean we want to estimate. List all population values as

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

$$\mu = M^{-1}\sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{m=1}^{M_{sc}} z_{scm}^o. \tag{6.3}$$

Define the total in the population as

$$\tau = \sum_{s=1}^{S}\sum_{c=1}^{C_s}\sum_{m=1}^{M_{sc}} z_{scm}^o = M\mu. \tag{6.4}$$

It is also useful to define the totals within each cluster and stratum:

$$\tau_{sc} = \sum_{m=1}^{M_{sc}} z_{scm}^o \tag{6.5}$$

and

$$\tau_s = \sum_{c=1}^{C_s} \tau_{sc}. \tag{6.6}$$

Now we can define the sampling scheme: (i) For each stratum s, randomly draw N_s clusters, with replacement. (ii) For each cluster c drawn in step (i), randomly sample K_{sc} households with replacement. For each pair (s,c), define

$$\hat{\mu}_{sc} = K_{sc}^{-1}\sum_{m=1}^{K_{sc}} z_{scm}. \tag{6.7}$$

Because this is an average based on a random sample within (s,c),

$$E(\hat{\mu}_{sc}) = \mu_{sc} = M_{sc}^{-1}\sum_{m=1}^{M_{sc}} z_{scm}^o. \tag{6.8}$$

To continue up to the cluster level we need the total, $\tau_{sc} = M_{sc}\mu_{sc}$. So, $\hat{\tau}_{sc} = M_{sc}\hat{\mu}_{sc}$ is an unbiased estimator of τ_{sc} for all $\{(s,c) : c = 1,\ldots,C_s, s = 1,\ldots,S\}$ (even if we eventually do

not use some clusters because they are not sampled). Now, for each stratum s, the estimator

$$N_s^{-1} \sum_{c=1}^{N_s} \hat{\tau}_{sc}, \tag{6.9}$$

which is the average of the cluster totals withing stratum s, has expected value which is the population average (for stratum s), that is,

$$C_s^{-1} \sum_{c=1}^{C_s} \tau_{sc} = C_s^{-1} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z_{scm}^o = C_s^{-1} \tau_s. \tag{6.10}$$

[In general, $C_s^{-1}\tau_s \neq \mu_s = \left(\sum_{c=1}^{C_s} M_{sc}\right)^{-1} \tau_s$ unless each cluster has only one observation.] It follows that an unbiased estimator of the total τ_s for stratum s is

$$C_s \cdot N_s^{-1} \sum_{c=1}^{N_s} \hat{\tau}_{sc}. \tag{6.11}$$

Finally, the total in the entire population is estimated as

$$\begin{aligned} \sum_{s=1}^{S} \left(C_s \cdot N_s^{-1} \sum_{c=1}^{N_s} \hat{\tau}_{sc} \right) &= \sum_{s=1}^{S} (C_s/N_s) \sum_{c=1}^{N_s} (M_{sc}/K_{sc}) \sum_{m=1}^{K_{sc}} z_{scm} \\ &= \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \left(\frac{C_s}{N_s} \cdot \frac{M_{sc}}{K_{sc}} \right) z_{scm} \\ &\equiv \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} z_{scm} \end{aligned} \tag{6.12}$$

where

$$\omega_{sc} \equiv \frac{C_s}{N_s} \cdot \frac{M_{sc}}{K_{sc}} \tag{6.13}$$

is the weight for every unit sampled in stratum-cluster pair (s,c). Note how (6.13) $\omega_{sc} = (C_s/N_s)(M_{sc}/K_{sc})$ accounts for under- or over-sampled clusters within strata and under- or over-sampled units within clusters. The expression in (6.12) appears in the literature on

complex survey sampling, sometimes without M_{sc}/K_{sc} when each cluster is sampled as a complete unit, and so $M_{sc}/K_{sc} = 1$. To estimate the population mean, μ, we just divide by M, the total number of units in the population,

$$\hat{\mu} = M^{-1}\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \omega_{sc} z_{scm}\right). \tag{6.14}$$

(The alternative is to use the regression formulation of estimating a mean that we now turn to, which does not require knowing M.)

To study the asymptotic properties of regression (and many other estimation methods), it is more convenient to modify the weights so that they are constant, or converge to a constant. The weights ω_{sc} converge to zero at rate N_s^{-1} because C_s and M_{sc} are fixed and K_{sc} is treated as fixed. (We assume a relative small number of households sampled per cluster.) Let $N = N_1 + N_2 + \ldots + N_S$ be the total number of clusters sampled.

$$\upsilon_{sc} = \frac{C_s}{(N_s/N)} \cdot \frac{M_{sc}}{K_{sc}} = N\omega_{sc}. \tag{6.15}$$

As in Bhattacharya (2005), it is easiest just to assume $N_s = a_s N$ for a_s fixed, $0 < a_s < 1$, $a_1 + \ldots + a_S = 1$. But we can also just assume N_s/N converges to a_s with the same property. Therefore, by writing $\upsilon_{sc} = (C_s/a_s)(M_s/K_s)$, we see that υ_{sc} is constant. Further, any optimization problem that uses ω_{sc} as weights gives the same answer when υ_{sc} is used because the scale factor in (6.15) does not depend on s or c. The key in the formulas for the asymptotic variance below is that υ_{sc} is (roughly) constant, and so N_s/N is critical in the formula.

While (6.15) is the most natural definition of the weights for obtaining the limiting distribution results, we can use different formulations without changing the end formulas. For example, let $C = C_1 + \ldots + C_S$ be the total number of clusters in the population, let M be the

total number of units in the population, and let K be the total units samples. Then, for the final formulas, we could use the weights defined as

$$\upsilon_{sc} = \frac{(C_s/C)}{(N_s/N)} \cdot \frac{(M_{sc}/M)}{(K_{sc}/K)} = \frac{(NK)}{(CM)}\omega_{sc}. \tag{6.16}$$

Because C, M, and K are fixed, the factor $K/(CM)$ has no effect on estimation or inference. Equation (6.16) has a nice interpretation because it is expressed in terms of frequencies of the population relative to the sample frequencies. For example, if $(C_s/C) > (N_s/N)$, which means that stratum s is underrepresented in terms of number of clusters, (6.16) gives more weight to such strata. The same is true of the fractions involving the number of units (say, households).

While we can consider general M-estimation problems, or generalized method of moments as in Bhattachary (2005), we consider least squares for concreteness. The weighted minimization problem is

$$\min_{\boldsymbol{\beta}} N^{-1} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \upsilon_{sc}(y_{scm} - \mathbf{x}_{scm}\boldsymbol{\beta})^2, \tag{6.17}$$

where it is helpful to divide by N to facilitate the asymptotic analysis as $N \to \infty$. The first order condition is

$$N^{-1} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \upsilon_{sc}\mathbf{x}_{scm}'(y_{scm} - \mathbf{x}_{scm}\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{6.18}$$

Using arguments similar to the SS sampling case, but accounting for the clustering (by, in effect, treating each cluster as its own observation), we can show that an appropriate estimator of Avar$(\hat{\boldsymbol{\beta}})$ – in the sense that it is consistent for Avar$[\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$ when multiplied by N – is

$$\left(\sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \upsilon_{sc}\mathbf{x}_{scm}'\mathbf{x}_{scm} \right)^{-1} \hat{\mathbf{B}} \left(\sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \upsilon_{sc}\mathbf{x}_{scm}'\mathbf{x}_{scm} \right)^{-1} \tag{6.19}$$

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

$$\hat{\mathbf{B}} = \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \upsilon_{sc}^2 \hat{u}_{scm}^2 \mathbf{x}_{scm}'\mathbf{x}_{scm} + \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\sum_{r\neq m}^{K_{sc}} \upsilon_{sc}^2 \hat{u}_{scm}\hat{u}_{scr}\mathbf{x}_{scm}'\mathbf{x}_{scr} \tag{6.20}$$
$$- \sum_{s=1}^{S} N_s^{-1}\left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \upsilon_{sc}\mathbf{x}_{scm}'\hat{u}_{scm}\right)\left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \upsilon_{sc}\mathbf{x}_{scm}'\hat{u}_{scm}\right)'.$$

The first part of $\hat{\mathbf{B}}$ is obtained using the White "heteroskedasticity"-robust form. The second piece accounts for the correlation within clusters. The third piece reduces the variance by accounting for the nonzero means of the "score" within strata, just as in the standard stratified sampling case.

If each cluster has just one unit, so $M_{sc} = K_{sc} = 1$, then (6.19) reduced to

$$\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s} \upsilon_{sc}\mathbf{x}_{sc}'\mathbf{x}_{sc}\right)^{-1}\left[\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s} \upsilon_{sc}^2\hat{u}_{sc}^2\mathbf{x}_{sc}'\mathbf{x}_{sc}\right) - \sum_{s=1}^{S} N_s^{-1}\left(\sum_{c=1}^{N_s} \upsilon_{sc}\mathbf{x}_{sc}'\hat{u}_{sc}\right)\left(\sum_{c=1}^{N_s} \upsilon_{sc}\mathbf{x}_{sc}'\hat{u}_{sc}\right)'\right]\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s}\right.$$

which is the formuala for standard stratified sampling with a finite number of units in each stratum.

References

Altonji, J.G, and R.L. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica* 73, 1053-1102.

Angrist, J.D. and V. Lavy (2002), "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," NBER Working Paper 9389.

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics* 49, 431-434.

Baker, M. and N.M. Fortin (2001), "Occupational Gender Composition and Wages in Canada, 1987-1988," *Canadian Journal of Economics* 34, 345-376.

Bhattacharya, D. (2005), "Asymptotic Inference from Multi-stage Samples," *Journal of Econometrics* 126, 145-171.

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-890.

Bertrand, M., E. Duflo, and S. Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119, 249-275.

Cameron, A.C., J.B. Gelbach, and D.L. Miller (2006), "Robust Inference with Multi-way Clustering," NBER Technical Working Paper Number 327.

Campolieti, M. (2004), "Disability Insurance Benefits and Labor Supply: Some Additional Evidence," *Journal of Labor Economics* 22, 863-889.

Card, D., and A.B. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84, 772-793.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of*

Economic Studies 47, 225-238.

Cosslett, S.R. (1993), "Estimation from Endogenously Stratified Samples," in G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., *Handbook of Statistics*, Volume 11, 1-43. North-Holland: Amsterdam.

Donald, S.G. and K. Lang (2001), "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics* 89, 221-233.

Hansen, C.B. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics* 141, 597-620.

Hausman, J.A., B.H. Hall, and Z. Griliches (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica* 52, 909-938.

Hsiao, C. (2003), *Analysis of Panel Data*. Cambridge: Cambridge University Press, second edition.

Imbens, G.W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling," *Econometrica* 60, 1187-1214.

Imbens, G.W. and T. Lancaster (1996), "Efficient Estimation and Stratified Sampling," *Journal of Econometrics* 74, 289-318.

Kézdi, G. (2001), "Robust Standard Error Estimation in Fixed-Effects Panel Models," mimeo, University of Michigan Department of Economics.

Liang, K.-Y., and S.L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika* 73, 13-22.

Loeb, S. and J. Bound (1996), "The Effect of Measured School Inputs on Academic Achievement: Evidence form the 1920s, 1930s and 1940s Birth Cohorts," *Review of Economics and Statistics* 78, 653-664

Moulton, B.R. (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics* 72, 334-338.

Pepper, J.V. (2002), "Robust Inferences from Random Clustered Samples: An Application Using Data from the Panel Study of Income Dynamics," Economics Letters 75, 341-345.

Petrin, A. and K. Train (2003), "Omitted Product Attributes in Discrete Choice Models," NBER Working Paper Number 9452.

White, H. (1980), "A Heterosekdasticity-Consistent Covariance Matrix Estimator and A Direct Test for Heteroskeasticity," *Econometrica* 48, 817-838.

White, H. (1982), "Maximum Likelihood Estimation with Misspecified Models," *Econometrica* 50, 1-26.

White, H. (1984), *Asymptotic Theory for Econometricians*. Academic Press: Orlando, FL.

Wooldridge, J.M. (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples," *Econometrica* 67, 1385-1406.

Wooldridge, J.M. (2001), "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Econometric Theory* 17, 451-470.

Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA.

Wooldridge, J.M. (2003), "Cluster-Sample Methods in Applied Econometrics," *American Economic Review* 93, 133-138.

Wooldridge, J.M. (2005), "Unobserved Heterogeneity and Estimation of Average Partial Effects," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. D.W.K. Andrews and J.H. Stock (eds.). Cambridge: Cambridge University Press, 27-55.

Wooldridge, J.M. (2006), "Cluster-Sample Methods in Applied Econometrics: An Extended Analysis," manuscript, Michigan State University Department of Economics.

Wooldridge, J.M. (2007), "Inverse Probability Weighted M-Estimation for General Missing Data Problems," *Journal of Econometrics* 141, 1281-1301.