Cluster-Sample Methods in Applied Econometrics

By JEFFREY M. WOOLDRIDGE*

Inference methods that recognize the clustering of individual observations have been available for more than 25 years. Brent Moulton (1990) caught the attention of economists when he demonstrated the serious biases that can result in estimating the effects of aggregate explanatory variables on individual-specific response variables. The source of the downward bias in the usual ordinary least-squares (OLS) standard errors is the presence of an unobserved, state-level effect in the error term. More recently, John Pepper (2002) showed how accounting for multi-level clustering can have dramatic effects on t statistics. While adjusting for clustering is much more common than it was 10 years ago, inference methods robust to cluster correlation are not used routinely across all relevant settings. In this paper, I provide an overview of applications of cluster-sample methods, both to cluster samples and to panel data sets.

Potential problems with inference in the presence of group effects when the number of groups is small have been highlighted in a recent paper by Stephen Donald and Kevin Lang (2001). I review different ways of handling the small number of groups case in Section III.

I. The Model

The goal is to estimate the parameters in the following linear model:

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

$$m = 1, \ldots, M_g \quad g = 1, \ldots, G$$

where g indexes the "group" or "cluster," m indexes observations within group, M_g is the group size, and G is the number of groups. The $1 \times K$ vector $\mathbf{x}_g$ contains explanatory variables that vary only at the group level, and the $1 \times L$ vector $\mathbf{z}_{gm}$ contains explanatory variables that vary within group. The approach to estimation and inference in equation (1) depends on several factors, including whether one is interested in the effects of aggregate variables ($\boldsymbol{\beta}$) or individual-specific variables ($\boldsymbol{\gamma}$). Plus, it is necessary to make assumptions about the error terms. An important issue is whether v_{gm} contains a common group effect, as in

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

where c_g is an unobserved cluster effect and u_{gm} is the idiosyncratic error. [In the statistics literature, (1) and (2) are referred to as a "hierarchical linear model."] One important issue is whether the explanatory variables in (1) can be taken to be appropriately exogenous. Under (2), exogeneity issues can be broken down by separately considering c_g and u_{gm}.

I assume that the sampling scheme generates observations that are independent across g. Appropriate sampling assumptions within cluster are more complicated. Theoretically, the simplest case also allows the most flexibility for robust inference: from a large population of relatively small clusters, draw a large number of clusters (G), of sizes M_g. This setup is appropriate in randomly sampling a large number of families or classrooms. The key feature is that the number of groups is large enough so that one can allow general 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) and the number of time periods is small. For panel data, G is the number of cross-sectional units, and M_g is the number of time periods for unit g.

Stratified sampling also results in data sets that can be arranged by group, where the population is first stratified into $G \geq 2$ nonoverlapping groups and then a random sample of size M_g is obtained from each group. Ideally,

* Department of Economics, Michigan State University, East Lansing, MI 48824-1038 (e-mail: wooldri1@msu.edu).
I thank Steven Haider and John Pepper for very helpful comments on an earlier draft. An expanded version of this paper is available from the author upon request.

the strata sizes are large in the population, resulting in large M_g. I consider this sampling scheme in Section III.

II. The Number of Groups is "Large"

The asymptotic theory for $G \to \infty$ is well developed; for a recent treatment, see Chapters 7, 10, and 11 in Wooldridge (2002). Suppose that the covariates are exogenous in the sense that

$$E(v_{gm}|\mathbf{x}_g, \mathbf{Z}_g) = 0 \quad m = 1, \ldots, M_g \quad g = 1, \ldots, G \tag{3}$$

where $\mathbf{Z}_g$ contains $\mathbf{z}_{gm}$, $m = 1, \ldots, M_g$. Then a pooled ordinary least-squares (OLS) estimator, where y_{gm} is regressed on 1, $\mathbf{x}_g$, $\mathbf{z}_{gm}$ ($m = 1, \ldots, M_g$; $g = 1, \ldots, G$) is consistent for $\boldsymbol{\lambda} \equiv (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}')'$ (as $G \to \infty$ 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 heteroscedasticity in $\text{Var}(v_{gm}|\mathbf{x}_g, \mathbf{Z}_g)$. When v_{gm} has the form in (2), the within-cluster correlation can be substantial, which means the usual OLS standard errors can be very misleading. Section 7.8 in Wooldridge (2002) gives the formula for a variance-matrix estimator that assumes no particular kind of within-cluster correlation nor a particular form of heteroscedasticity. These formulas apply without change to panel data with a large number of cross-sectional observations. Such variance matrices are easy to compute now with existing software packages.

Under (2) one can use generalized least squares (GLS) to exploit the presence of c_g in v_{gm}. The standard assumptions imply that the $M_g \times M_g$ variance-covariance matrix of $\boldsymbol{v}_g = (v_{g1}, v_{g2}, \ldots, v_{g,M_g})'$ has the "random effects" form, $\text{Var}(\boldsymbol{v}_g) = \sigma_c^2 \mathbf{j}_{M_g} \mathbf{j}_{M_g}' + \sigma_u^2 \mathbf{I}_{M_g}$, where $\mathbf{j}_{M_g}$ is the $M_g \times 1$ vector of 1's and $\mathbf{I}_{M_g}$ is the $M_g \times M_g$ identity matrix. The standard assumptions also include the "system homoscedasticity" assumption, $\text{Var}(\boldsymbol{v}_g|\mathbf{x}_g, \mathbf{Z}_g) = \text{Var}(\boldsymbol{v}_g)$. The resulting GLS estimator is the well-known random-effects (RE) estimator (see Section 10.3 in Wooldridge [2002]).

The RE estimator is asymptotically more efficient than pooled OLS under the usual RE assumptions, and RE estimates and test statistics are computed by popular software packages. Something often overlooked in applications is that one can make inference completely robust to an unknown form of $\text{Var}(\boldsymbol{v}_g|\mathbf{x}_g, \mathbf{Z}_g)$. Equation 7.49 in Wooldridge (2002) gives the robust formula. Even if $\text{Var}(\boldsymbol{v}_g|\mathbf{x}_g, \mathbf{Z}_g)$ does not have the RE form, the RE estimator is still consistent and $\sqrt{G}$-asymptotically normal, and for interesting departures from the full RE assumptions, the RE estimator is likely to be more efficient than pooled OLS. Making inference robust to serial correlation in the idiosyncratic errors for panel-data applications can be very important. Within-group correlation in the u_{gm} can arise for cluster samples too. For example, suppose that underlying (1) is a random coefficient model where $\mathbf{z}_{gm}\boldsymbol{\gamma}_g$ replaces $\mathbf{z}_{gm}\boldsymbol{\gamma}$. By estimating an RE model, one effectively puts $\mathbf{z}_g(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ in the idiosyncratic error, and this induces correlation across u_{gm}. Under standard exogeneity assumptions, the RE estimator still consistently estimates the average effect, $\boldsymbol{\gamma} = E(\boldsymbol{\gamma}_g)$. For a large G one might estimate an unrestricted version of $\text{Var}(\boldsymbol{v}_g)$, but even in this case one should use a variance matrix robust to $\text{Var}(\boldsymbol{v}_{gm}|\mathbf{x}_g, \mathbf{Z}_g) \neq \text{Var}(\boldsymbol{v}_g)$.

In economics, the prevailing view is that robust inference is not necessary when using GLS, but the "generalized estimation equation" literature (see Kung-Yee Liang and Scott Zeger, 1986) explicitly recognizes that a specified variance matrix in panel-data applications need not be equal to the true conditional variance matrix.

If c_g is correlated with $(\mathbf{x}_g, \mathbf{Z}_g)$, neither $\boldsymbol{\beta}$ nor $\boldsymbol{\gamma}$ is consistently estimated by RE. Nevertheless, by using the "fixed-effects" (FE) or "within" estimator, one can still estimate $\boldsymbol{\gamma}$. The within transformation subtracts off group averages from the dependent variable and explanatory variables:

$$y_{gm} - \bar{y}_g = (\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)\boldsymbol{\gamma} + u_{gm} - \bar{u}_g \quad m = 1, \ldots, M_g \quad g = 1, \ldots, G \tag{4}$$

and this equation is estimated by pooled OLS. Under a full set of "fixed-effects" assumptions (which allows arbitrary correlation between c_g and the $\mathbf{z}_{gm}$), inference is straightforward using standard software. Nevertheless, analogous to the random-effects case, it is important to allow

$\text{Var}(\mathbf{u}_g|\mathbf{Z}_g)$ to have an arbitrary form, including within-group correlation and heteroscedasticity. Manuel Arellano (1987) proposed a fully robust variance-matrix estimator for the fixed-effects estimator, and it works with cluster samples or panel data (see also equation 10.59 in Wooldridge [2002]). Reasons for wanting a fully robust variance-matrix estimator for FE applied to cluster samples are similar to the RE case.

III. The Number of Groups is "Small"

The procedures described in Section II are easy to implement, so it is natural to ask: Even though those procedures are theoretically justified for large G, might they work well for moderate, or even small, G? Joshua D. Angrist and Victor Lavy (2002) provide references that show how cluster-robust estimators after pooled OLS do not work very well, even when G is as large as 40 or 50. Less is known about how well the fully robust variance-matrix estimator and the associated robust inference work after RE estimation.

Recently, in the context of fixed-effects estimation and panel data, Gábor Kézde (2001) and Marianne Bertrand et al. (2002) study the finite-sample properties of robust variance-matrix estimators that are theoretically justified only as $G \to \infty$. One common finding is that the fully robust estimator works reasonably well even when the cross-sectional sample size is not especially large relative to the time-series dimension. When $\text{Var}(\mathbf{u}_g|\mathbf{Z}_g)$ does not depend on $\mathbf{Z}_g$, a variance matrix that exploits system homoscedasticity can perform better than the fully robust variance-matrix estimator.

Importantly, the encouraging findings of the simulations for fixed effects with panel data are not in conflict with findings that the robust variance matrix for the pooled OLS estimator with a small number of groups can behave poorly. For FE estimation using panel data, the issue is serial correlation in $\{u_{gm}: m = 1, \ldots, M_g\}$, which dies out as the time periods get far apart. The pooled OLS estimator that keeps c_g in the error term suffers because of the constant correlation across all observations within cluster. Plus, FE estimates $\boldsymbol{\gamma}$, while for pooled OLS with clustering the focus is usually on $\boldsymbol{\beta}$.

When G is very small, relying on large G asymptotics can be very misleading. Donald and Lang (2001; hereafter, DL) have recently offered an alternative approach to inference, particularly for hypothesis testing about $\boldsymbol{\beta}$. To begin, consider a special case of (1) where $\mathbf{z}_{gm}$ is not in the equation and x_g is a scalar. The equation is

$$(5) \qquad y_{gm} = \alpha + \beta x_g + c_g + u_{gm}$$
$$m = 1, \ldots, M_g \quad g = 1, \ldots, G$$

where c_g and $\{u_{gm}: m = 1, \ldots, M_g\}$ are independent of x_g and $\{u_{gm}: m = 1, \ldots, M_g\}$ is a mean-zero, independent, identically distributed sequence for each g. Even with small G, the pooled OLS estimator is natural for estimating β. If the cluster effect c_g is absent from the model and $\text{Var}(u_{gm})$ is constant across g, then provided $N \equiv M_1 + \cdots + M_G$ is large enough (whether or not G is not large), we can use the usual t statistics from the pooled OLS regression as having an approximate standard normal distribution. Making inference robust to heteroscedasticity is straightforward for large N.

As pointed out by DL, the presence of c_g makes the usual pooled-OLS inference methods poorly behaved with small G. With a common cluster effect, there is no averaging out within cluster that allows application of the central-limit theorem. One way to see the problem is to note that the pooled-OLS estimator, $\hat{\beta}$, is identical to the "between" estimator obtained from the regression of $\bar{y}_g$ on 1, x_g $(g = 1, \ldots, G)$. Given the x_g, $\hat{\beta}$ inherits its distribution from $\{\bar{v}_g: g = 1, \ldots, G\}$, the within-group averages of the v_{gm}. The presence of c_g means new observations within group do not provide additional information for estimating β beyond affecting the group average, $\bar{y}_g$.

If some assumptions are added, there is a solution to the inference problem. In particular, assume $c_g \sim \mathcal{N}(0, \sigma_c^2)$ is independent of $u_{gm} \sim \mathcal{N}(0, \sigma_u^2)$ and $M_g = M$ for all g where $\mathcal{N}$ denotes a normal distribution. Then $\bar{v}_g \sim \mathcal{N}(0, \sigma_c^2 + \sigma_u^2/M)$. Since independence across g is assumed, the equation

$$(6) \quad \bar{y}_g = \alpha + \beta x_g + \bar{v}_g \qquad g = 1, \ldots, G$$

satisfies the classical linear-model assumptions. Therefore, one can use inference based on the t_{G-2} distribution to test hypotheses about β,

provided $G > 2$. When G is small, the requirements for a significant t statistic are much more stringent then if one uses the $t_{M_1+M_2+\cdots+M_G-2}$ distribution, which is what one would be doing by using the usual pooled OLS statistics. When $\mathbf{x}_g$ is a $1 \times K$ vector, one needs $G > K + 1$ to use the t_{G-K-1} distribution for inference after estimating the aggregated equation (6) by OLS. If $\mathbf{z}_{gm}$ is in the model, then one can add the group averages, $\bar{\mathbf{z}}_g$, to (6), provided $G > K + L + 1$, and use the $t_{G-K-L-1}$ distribution for inference. (An alternative approach that conserves on degrees of freedom, but is only approximately valid, is described below.)

Importantly, 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, or using the RE estimator. All three estimation methods lead to the same $\hat{\beta}$. But using the between regression in (6) gives the appropriate standard error *and* reports the small degrees of freedom in the t distribution.

If the common group size M is large, then $\bar{u}_g$ will be approximately normal very generally, so $\bar{v}_g$ is approximately normal with constant variance. Even if the group sizes differ, for very large group sizes $\bar{u}_g$ will be a negligible part of $\bar{v}_g$. Provided c_g is normally distributed, classical linear model analysis on (6) should be roughly valid.

For small G and large M_g, inference obtained from analyzing (6) as a classical linear model can be very conservative if there is no cluster effect. Perhaps this is desirable, but it also excludes some staples of policy analysis. In the simplest case, suppose there are two populations with means μ_g ($g = 1, 2$), and the question is whether their difference is zero. Under random sampling from each population, 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 means. With even moderate-sized M_1 and M_2, one can relax normality and adjust the statistic for different population variances. In the DL setup, the standard comparison-of-means case cannot even be analyzed, because $G = 2$. DL criticize David Card and Alan B. Krueger (1994) for comparing mean wage changes of fast-food workers because Card and Krueger fail to account for c_g in v_{gm}, but the criticism in the $G = 2$ case is indistinguishable from a common criticism of difference-in-differences (DID) analyses: How can one be sure that any observed difference in means is due entirely to the policy change?

More generally, in studies with $G > 2$ it often makes sense to view the observations as coming from standard stratified sampling. With large group sample sizes one can get precise estimates of the group population means, μ_g. For example, suppose that $G = 4$ and groups 1 and 2 are control groups, while groups 3 and 4 are treated groups. One might estimate the policy effect by $\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2$, or different fixed weights could be used to allow for different group population sizes. In any case, one can get a good estimator of τ by plugging in the group means, and when properly standardized, $\hat{\tau}$ will be approximately standard normal even if the M_g are as small as, say, 30. To obtain a valid standard error, it is not necessary to assume that the group means or variances within, say, the treated group, are the same. In the DL approach, the estimated treatment effect, $\hat{\beta}$, is obtained by pooling within the treated and control groups, and then differencing the treatment and control means. Their inference using the t_2 distribution is a different way of accounting for $\mu_1 \neq \mu_2$ or $\mu_3 \neq \mu_4$. It seems that more work is needed to reconcile the two approaches when G is small.

With large group sizes, a minimum distance (MD) approach to estimating $\boldsymbol{\beta}$ sheds additional insight. For each group g, write a linear model with individual-specific covariates as

$$(7)\quad y_{gm} = \delta_g + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + u_{gm} \qquad m = 1, \ldots, M_g$$

assuming random sampling within groups. Also, make the assumptions for OLS to be consistent (as $M_g \to \infty$) and $\sqrt{M_g}$-asymptotically normal (see Wooldridge, 2002 Ch. 4). The presence of group-level variables $\mathbf{x}_g$ in (1) can be viewed as putting restrictions on the intercepts, δ_g. In particular,

$$(8)\quad \delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} \qquad g = 1, \ldots, G$$

where we now think of $\mathbf{x}_g$ as fixed observed attributes of the different groups. Given that one can 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. Alternatively, to impose $\gamma_g = \gamma$ for all g, then pool across groups and include group dummy variables to get the $\hat{\delta}_g$. If $G = K + 1$ then one can solve for $\hat{\theta} \equiv (\hat{\alpha}, \hat{\beta}')'$ uniquely in terms of the $G \times 1$ vector $\hat{\delta}$: $\hat{\theta} = \mathbf{X}^{-1}\hat{\delta}$, where $\mathbf{X}$ is the $(K + 1) \times (K + 1)$ matrix with gth row $(1, \mathbf{x}_g)$. If $G > K + 1$, then in a second step, one can use an MD approach, as described in Section 14.6 of Wooldridge (2002). If the $G \times G$ identity matrix is the weighting matrix, the MD estimator can be computed from the OLS regression of

$$(9) \qquad \hat{\delta}_g \text{ on } 1, \mathbf{x}_g \qquad g = 1, \ldots, G.$$

If $M_g = \rho_g M$ where $0 < \rho_g \leq 1$ and $M \to \infty$, the MD estimator $\hat{\theta}$ is consistent and $\sqrt{M}$-asymptotically normal. However, this MD estimator is asymptotically inefficient except under strong assumptions. It is not difficult to obtain the efficient MD estimator—also called the "minimum chi-square" estimator. The simplest case is when $\mathbf{z}_{gm}$ does not appear in the first-stage estimation, so that the $\hat{\delta}_g$ are sample means. Let $\hat{\sigma}_g^2$ denote the usual sample variance for group g. The minimum chi-square estimator can be computed by using the weighted-least-squares (WLS) version of (9), where group g is weighted by $M_g/\hat{\sigma}_g^2$. Conveniently, the reported t statistics from the WLS regression are asymptotically standard normal as the group sizes M_g get large. An example of this kind of procedure is given by Susanna Loeb and John Bound (1996).

A 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 (SSR): under the null hypothesis in (8), $\text{SSR}_w \overset{a}{\sim} \chi^2_{G-K-1}$. If the null hypothesis H_0 is rejected at a small significance level, the x_g are not sufficient for characterizing the changing intercepts across groups. If one fails to reject H_0, one can have some confidence in the specification and perform inference using the standard normal distribution for t statistics.

If $\mathbf{z}_{gm}$ appears in the first stage, one can use as weights the asymptotic variances of the G intercepts. These might be made fully robust to heteroscedasticity in $E(u_{gm}^2|\mathbf{z}_{gm})$, or at least allow different σ_g^2. In any case, the weights are given by $1/[\text{SE}(\hat{\delta}_g)]^2$ $(g = 1, \ldots, G)$, where $\text{SE}(\hat{\delta}_g)$ are the asymptotic standard errors.

For example, suppose x_g is a binary treatment indicator. Then $\hat{\beta}$ is an estimate of an average treatment effect. If $G = 2$ there are no restrictions to test. With $G > 2$ one can test the overidentifying restrictions. Rejection implies that there are missing group-level characteristics, and one might re-specify the model by adding elements to x_g, even if the new elements are not thought to be systematically related to the original elements (as when treatment is randomly assigned at the group level).

Alternatively, if the restrictions in (8) are rejected, one concludes that $\delta_g = \alpha + x_g\beta + c_g$, where c_g is the error made in imposing the restrictions. This leads to the DL approach, which is to analyze the OLS regression in (9) where inference is based on the t_{G-K-1} distribution. Why is this approach justified? Since $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, for large M_g one might ignore the estimation error in δ_g. Then, it is as if the equation $\delta_g = \alpha + x_g\beta + c_g$ $(g = 1, \ldots, G)$ is being estimated by OLS. Classical analysis is applicable when $c_g \sim \mathcal{N}(0, \sigma_c^2)$ and c_g is independent of x_g. The latter assumption means that differences in the intercepts δ_g not due to x_g must be unrelated to x_g, which seems reasonable if G is not too small and x_g is a randomly assigned treatment variable assigned at the group level, as in Angrist and Lavy (2002).

REFERENCES

Angrist, Joshua D. and Lavy, Victor. "The Effect of High School Matriculation Awards: Evidence from Randomized Trials." Working paper, Massachusetts Institute of Technology, 2002.

Arellano, Manuel. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics*, November 1987, *49*(4), pp. 431–34.

Bertrand, Marianne; Duflo, Esther and Mullainathan, Sendhil. "How Much Should We Trust Differences-in-Differences Estimates?" Working paper, Massachusetts Institute of Technology, 2002.

Card, David and Krueger, Alan B. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and

Pennsylvania." *American Economic Review*, September 1994, *84*(4), pp. 772–93.

Donald, Stephen G. and Lang, Kevin. "Inference with Difference in Differences and Other Panel Data." Working paper, University of Texas, 2001.

Kézde, Gábor. "Robust Standard Error Estimation in Fixed-Effects Panel Models." Working paper, University of Michigan, 2001.

Liang, Kung-Yee and Zeger, Scott L. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, April 1986, *73*(1), pp. 13–22.

Loeb, Susanna and Bound, John. "The Effect of Measured School Inputs on Academic Achievement: Evidence from the 1920s, 1930s, and 1940s Birth Cohorts." *Review of Economics and Statistics*, November 1996, *78*(4), pp. 653–64.

Moulton, Brent R. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics*, May 1990, *72*(2), pp. 334–38.

Pepper, John V. "Robust Inferences from Random Clustered Samples: An Application Using Data from the Panel Study of Income Dynamics." *Economics Letters*, May 2002, *75*(3), pp. 341–45.

Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press, 2002.