A Practitioner's Guide to Cluster-Robust Inference

A. Colin Cameron and Douglas L. Miller

Abstract

We consider statistical inference for regression when data are grouped into clusters, with regression model errors independent across clusters but correlated within clusters. Examples include data on individuals with clustering on village or region or other category such as industry, and state-year differences-in-differences studies with clustering on state. In such settings default standard errors can greatly overstate estimator precision. Instead, if the number of clusters is large, statistical inference after OLS should be based on cluster-robust standard errors. We outline the basic method as well as many complications that can arise in practice. These include cluster-specific fixed effects, few clusters, multi-way clustering, and estimators other than OLS.

Colin Cameron is a Professor in the Department of Economics at UC- Davis. Doug Miller is an Associate Professor in the Department of Economics at UC- Davis. *They thank four referees and the journal editor for very helpful comments and for guidance, participants at the 2013 California Econometrics Conference, a workshop sponsored by the U.K. Programme Evaluation for Policy Analysis, seminars at University of Southern California and at University of Uppsala, and the many people who over time have sent them cluster-related puzzles (the solutions to some of which appear in this paper). Doug Miller acknowledges financial support from the Center for Health and Wellbeing at the Woodrow Wilson School of Public Policy at Princeton University.*

I. Introduction

In an empiricist's day-to-day practice, most effort is spent on getting unbiased or consistent point estimates. That is, a lot of attention focuses on the parameters ($\hat{\beta}$). In this paper we focus on getting accurate statistical inference, a fundamental component of which is obtaining accurate standard errors (se, the estimated standard deviation of $\hat{\beta}$). We begin with the basic reminder that empirical researchers should also really care about getting this part right. An asymptotic 95% confidence interval is $\hat{\beta} \pm 1.96 \times se$, and hypothesis testing is typically based on the Wald "t-statistic" $w = (\hat{\beta} - \beta_0)/se$. Both $\hat{\beta}$ and se are critical ingredients for statistical inference, and we should be paying as much attention to getting a good se as we do to obtain $\hat{\beta}$.

In this paper, we consider statistical inference in regression models where observations can be grouped into clusters, with model errors uncorrelated across clusters but correlated within cluster. One leading example of "clustered errors" is individual-level cross-section data with clustering on geographical region, such as village or state. Then model errors for individuals in the same region may be correlated, while model errors for individuals in different regions are assumed to be uncorrelated. A second leading example is panel data. Then model errors in different time periods for a given individual (e.g., person or firm or region) may be correlated, while model errors for different individuals are assumed to be uncorrelated.

Failure to control for within-cluster error correlation can lead to very misleadingly small standard errors, and consequent misleadingly narrow confidence intervals, large t-statistics and low p-values. It is not unusual to have applications where standard errors that control for within-cluster correlation are several times larger than default standard errors that ignore such correlation. As shown below, the need for such control increases not only with increase in the size of within-cluster error correlation, but the need also increases with the size of within-cluster correlation of regressors and with the number of observations within a cluster. A leading example, highlighted by Moulton (1986, 1990), is when interest lies in measuring the effect of a policy variable, or other aggregated regressor, that takes the same value for all observations within a cluster.

One way to control for clustered errors in a linear regression model is to additionally specify a model for the within-cluster error correlation, consistently estimate the parameters of this error correlation model, and then estimate the original model by feasible generalized least squares (FGLS) rather than ordinary least squares (OLS). Examples include random effects estimators and, more generally, random coefficient and hierarchical models. If all goes well this provides valid statistical inference, as well as estimates of the parameters of the original regression model that are more efficient than OLS. However, these desirable properties hold only under the very strong assumption that the model for within-cluster error correlation is correctly specified.

A more recent method to control for clustered errors is to estimate the regression model with limited or no control for within-cluster error correlation, and then post-estimation obtain "cluster-robust" standard errors proposed by White (1984, p.134-142) for OLS with a multivariate dependent variable (directly applicable to balanced clusters); by Liang and Zeger (1986) for linear and nonlinear models; and by Arellano (1987) for the fixed effects estimator in linear panel models. These cluster-robust standard errors do not require specification of a model for within-cluster error correlation, but do require the additional assumption that the number of clusters, rather than just the number of observations, goes to infinity.

Cluster-robust standard errors are now widely used, popularized in part by Rogers (1993) who incorporated the method in Stata, and by Bertrand, Duflo and Mullainathan (2004)

who pointed out that many differences-in-differences studies failed to control for clustered errors, and those that did often clustered at the wrong level. Cameron and Miller (2011) and Wooldridge (2003, 2006) provide surveys, and lengthy expositions are given in Angrist and Pischke (2009) and Wooldridge (2010).

One goal of this paper is to provide the practitioner with the methods to implement cluster-robust inference. To this end we include in the paper reference to relevant Stata commands (for version 13), since Stata is the computer package most often used in applied microeconometrics research. And we will post on our websites more expansive Stata code and the datasets used in this paper. A second goal is presenting how to deal with complications such as determining when there is a need to cluster, incorporating fixed effects, and inference when there are few clusters. A third goal is to provide an exposition of the underlying econometric theory as this can aid in understanding complications. In practice the most difficult complication to deal with can be "few" clusters, see Section VI. There is no clear-cut definition of "few"; depending on the situation "few" may range from less than 20 to less than 50 clusters in the balanced case.

We focus on OLS, for simplicity and because this is the most commonly-used estimation method in practice. Section II presents the basic results for OLS with clustered errors. In principle, implementation is straightforward as econometrics packages include cluster-robust as an option for the commonly-used estimators; in Stata it is the `vce(cluster)` option. The remainder of the survey concentrates on complications that often arise in practice. Section III addresses how the addition of fixed effects impacts cluster-robust inference. Section IV deals with the obvious complication that it is not always clear what to cluster over. Section V considers clustering when there is more than one way to do so and these ways are not nested in each other. Section VI considers how to adjust inference when there are just a few clusters as, without adjustment, test statistics based on the cluster-robust standard errors over-reject and confidence intervals are too narrow. Section VII presents extension to the full range of estimators – instrumental variables, nonlinear models such as logit and probit, and generalized method of moments. Section VIII presents both empirical examples and real-data based simulations. Concluding thoughts are given in Section IX.

II. Cluster-Robust Inference

In this section we present the fundamentals of cluster-robust inference. For these basic results we assume that the model does not include cluster-specific fixed effects, that it is clear how to form the clusters, and that there are many clusters. We relax these conditions in subsequent sections.

Clustered errors have two main consequences: they (usually) reduce the precision of $\hat{\beta}$, and the standard estimator for the variance of $\hat{\beta}$, $\widehat{\mathrm{V}}[\hat{\beta}]$, is (usually) biased downward from the true variance. Computing cluster-robust standard errors is a fix for the latter issue. We illustrate these issues, initially in the context of a very simple model and then in the following subsection in a more typical model.

A. A Simple Example

For simplicity, we begin with OLS with a single regressor that is nonstochastic, and assume no intercept in the model. The results extend to multiple regression with stochastic regressors.

Let $y_i = \beta x_i + u_i,\ i = 1,\ldots,N$, where x_i is nonstochastic and $\mathrm{E}[u_i] = 0$. The OLS estimator $\hat{\beta} = \sum_i x_i\, y_i / \sum_i x_i^2$ can be re-expressed as $\hat{\beta} - \beta = \sum_i x_i u_i \,/\, \sum_i x_i^2$, so in general

$$\mathrm{V}[\hat{\beta}] = \mathrm{E}[(\hat{\beta}-\beta)^2] = \mathrm{V}\left[\sum_i x_i u_i\right] / \left(\sum_i x_i^2\right)^2. \quad (1)$$

If errors are uncorrelated over i, then $\mathrm{V}[\sum_i x_i u_i] = \sum_i \mathrm{V}[x_i u_i] = \sum_i x_i^2 \mathrm{V}[u_i]$. In the simplest case of homoskedastic errors, $\mathrm{V}[u_i] = \sigma^2$ and (1) simplifies to $\mathrm{V}[\hat{\beta}] = \sigma^2 / \sum_i x_i^2$.

If instead errors are heteroskedastic, then (1) becomes

$$\mathrm{V}_{\mathrm{het}}[\hat{\beta}] = \left(\sum_i x_i^2 \mathrm{E}[u_i^2]\right) / \left(\sum_i x_i^2\right)^2,$$

using $\mathrm{V}[u_i] = \mathrm{E}[u_i^2]$ since $\mathrm{E}[u_i] = 0$. Implementation seemingly requires consistent estimates of each of the N error variances $\mathrm{E}[u_i^2]$. In a very influential paper, one that extends naturally to the clustered setting, White (1980) noted that instead all that is needed is an estimate of the scalar $\sum_i x_i^2 \mathrm{E}[u_i^2]$, and that one can simply use $\sum_i x_i^2 \hat{u}_i^2$, where $\hat{u}_i = y_i - \hat{\beta} x_i$ is the OLS residual, provided $N \to \infty$. This leads to estimated variance

$$\widehat{\mathrm{V}}_{\mathrm{het}}[\hat{\beta}] = \left(\sum_i x_i^2 \hat{u}_i^2]\right) / \left(\sum_i x_i^2\right)^2.$$

The resulting standard error for $\hat{\beta}$ is often called a robust standard error, though a better, more precise term, is heteroskedastic-robust standard error.

What if errors are correlated over i? In the most general case where all errors are correlated with each other,

$$\mathrm{V}\left[\sum_i x_i u_i\right] = \sum_i \sum_j \mathrm{Cov}[x_i u_i, x_j u_j] = \sum_i \sum_j x_i x_j \mathrm{E}[u_i u_j],$$

so

$$\mathrm{V}_{\mathrm{cor}}[\hat{\beta}] = \left(\sum_i \sum_j x_i x_j \mathrm{E}[u_i u_j]\right) / \left(\sum_i x_i^2\right)^2.$$

The obvious extension of White (1980) is to use $\widehat{\mathrm{V}}[\hat{\beta}] = \left(\sum_i \sum_j x_i x_j \hat{u}_i \hat{u}_j]\right) / (\sum_i x_i^2)^2$, but this equals zero since $\sum_i x_i \hat{u}_i = 0$. Instead one needs to first set a large fraction of the error correlations $\mathrm{E}[u_i u_j]$ to zero. For time series data with errors assumed to be correlated only up to, say, m periods apart as well as heteroskedastic, White's result can be extended to yield a heteroskedastic- and autocorrelation-consistent (HAC) variance estimate; see Newey and West (1987).

In this paper we consider clustered errors, with $\mathrm{E}[u_i u_j] = 0$ unless observations i and j are in the same cluster (such as same region). Then

$$\mathrm{V}_{\mathrm{clu}}[\hat{\beta}] = \left(\sum_i \sum_j x_i x_j \mathrm{E}[u_i u_j] \mathbf{1}[i, j \text{ in same cluster}]\right) / \left(\sum_i x_i^2\right)^2, \quad (2)$$

where the indicator function $\mathbf{1}[A]$ equals 1 if event A happens and equals 0 if event A does not happen. Provided the number of clusters goes to infinity, we can use the variance estimate

$$\widehat{\mathrm{V}}_{\mathrm{clu}}[\hat{\beta}] = \left(\sum_i \sum_j x_i x_j \hat{u}_i \hat{u}_j \mathbf{1}[i,j \text{ in same cluster}]\right) / \left(\sum_i x_i^2\right)^2. \tag{3}$$

This estimate is called a cluster-robust estimate, though more precisely it is heteroskedastic- and cluster-robust. This estimate reduces to $\widehat{\mathrm{V}}_{\mathrm{het}}[\hat{\beta}]$ in the special case that there is only one observation in each cluster.

Typically $\widehat{\mathrm{V}}_{\mathrm{clu}}[\hat{\beta}]$ exceeds $\widehat{\mathrm{V}}_{\mathrm{het}}[\hat{\beta}]$ due to the addition of terms when $i \neq j$. The amount of increase is larger (1) the more positively associated are the regressors across observations in the same cluster (via $x_i x_j$ in (3)), (2) the more correlated are the errors (via $\mathrm{E}[u_i u_j]$ in (2)), and (3) the more observations are in the same cluster (via $\mathbf{1}[i,j$ in same cluster$]$ in (3)).

There are several take-away messages. First there can be great loss of efficiency in OLS estimation if errors are correlated within cluster rather than completely uncorrelated. Intuitively, if errors are positively correlated within cluster then an additional observation in the cluster no longer provides a completely independent piece of new information. Second, failure to control for this within-cluster error correlation can lead to using standard errors that are too small, with consequent overly-narrow confidence intervals, overly-large t-statistics, and over-rejection of true null hypotheses. Third, it is straightforward to obtain cluster-robust standard errors, though they do rely on the assumption that the number of clusters goes to infinity (see Section VI for the few clusters case).

B. Clustered Errors and Two Leading Examples

Let i denote the i^{th} of N individuals in the sample, and g denote the g^{th} of G clusters. Then for individual i in cluster g the linear model with (one-way) clustering is

$$y_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta} + u_{ig}, \tag{4}$$

where $\boldsymbol{x}_{ig}$ is a $K \times 1$ vector. As usual it is assumed that $\mathrm{E}[u_{ig}|\boldsymbol{x}_{ig}] = 0$. The key assumption is that errors are uncorrelated across clusters, while errors for individuals belonging to the same cluster may be correlated. Thus

$$\mathrm{E}[u_{ig}u_{jg'}|\boldsymbol{x}_{ig}, \boldsymbol{x}_{jg'}] = 0, \text{ unless } g = g'. \tag{5}$$

1. Example 1: Individuals in Cluster

Hersch (1998) uses cross-section individual-level data to estimate the impact of job injury risk on wages. Since there is no individual-level data on job injury rate, a more aggregated measure such as job injury risk in the individual's industry is used as a regressor. Then for individual i (with $N = 5960$) in industry g (with $G = 211$)

$$y_{ig} = \gamma \times x_g + \boldsymbol{z}_{ig}'\boldsymbol{\delta} + u_{ig}.$$

The regressor x_g is perfectly correlated within industry. The error term will be positively correlated within industry if the model systematically overpredicts (or underpredicts) wages in a given industry. In this case default OLS standard errors will be

downwards biased.

To measure the extent of this downwards bias, suppose errors are equicorrelated within cluster, so Cor$[u_{ig}, u_{jg}] = \rho$ for all $i \neq j$. This pattern is suitable when observations can be viewed as exchangeable, with ordering not mattering. Common examples include the current one, individuals or households within a village or other geographic unit (such as state), individuals within a household, and students within a school. Then a useful approximation is that for the k^{th} regressor the default OLS variance estimate based on $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, where s is the standard error of the regression, should be inflated by

$$\tau_k \simeq 1 + \rho_{x_k}\rho_u(\bar{N}_g - 1), \tag{6}$$

where ρ_{x_k} is a measure of the within-cluster correlation of x_{igk}, ρ_u is the within-cluster error correlation, and $\bar{N}_g$ is the average cluster size. The result (6) is exact if clusters are of equal size ("balanced" clusters) and $\rho_{x_k} = 1$ for all regressors (Kloek, 1981); see Scott and Holt (1982) and Greenwald (1983) for the general result with a single regressor.

This very important and insightful result is that the variance inflation factor is increasing in

1. the within-cluster correlation of the regressor
2. the within-cluster correlation of the error
3. the number of observations in each cluster.

For clusters of unequal size replace $(\bar{N}_g - 1)$ in (6) by $((\mathrm{V}[N_g]/\bar{N}_g) + \bar{N}_g - 1)$; see Moulton (1986, p.387). Note that there is no cluster problem if any one of the following occur: $\rho_{x_k} = 0$ or $\rho_u = 0$ or $\bar{N}_g = 1$.

In an influential paper, Moulton (1990) pointed out that in many settings the inflation factor τ_k can be large even if ρ_u is small. He considered a log earnings regression using March CPS data $(N = 18{,}946)$, regressors aggregated at the state level $(G = 49)$, and errors correlated within state $(\hat{\rho}_u = 0.032)$. The average group size was $18{,}946/49 = 387$, $\rho_{x_k} = 1$ for a state-level regressor, so (6) yields $\widehat{\tau_k} \simeq 1 + 1 \times 0.032 \times 386 = 13.3$. The weak correlation of errors within state was still enough to lead to cluster-corrected standard errors being $\sqrt{13.3} = 3.7$ times larger than the (incorrect) default standard errors!

In such examples of cross-section data with an aggregated regressor, the cluster-robust standard errors can be much larger despite low within-cluster error correlation because the regressor of interest is perfectly correlated within cluster and there may be many observations per cluster.

2. Example 2: Differences-in-Differences (DiD) in a State-Year Panel

Interest may lie in how wages respond to a binary policy variable d_{ts} that varies by state and over time. Then at time t in state s

$$y_{ts} = \gamma \times d_{ts} + \boldsymbol{z}_{ts}'\boldsymbol{\gamma} + \alpha_s + \delta_t + u_{ts},$$

where we assume independence over states, so the ordering of subscripts (t, s) corresponds to (i, g) in (4), and α_s and δ_t are state and year effects.

The binary regressor d_{ts} equals one if the policy of interest is in effect and equals 0 otherwise. The regressor d_{ts} is often highly serially correlated since, for example, d_{ts} will equal a string of zeroes followed by a string of ones for a state that switches from never having the policy in place to forever after having the policy in place. The error u_{ts} is correlated over time for a given state if the model systematically overpredicts (or underpredicts) wages in a

given state. Again the default standard errors are likely to be downwards-biased.

In the panel data case, the within-cluster (i.e., within-individual) error correlation decreases as the time separation increases, so errors are not equicorrelated. A better model for the errors is a time series model, such as an autoregressive error of order one that implies that $\text{Cor}[u_{ts}, u_{t's}] = \rho^{|t-t'|}$. The true variance of the OLS estimator will again be larger than the OLS default, although the consequences of clustering are less extreme than in the case of equicorrelated errors (see Cameron and Miller (2011, Section 2.3) for more detail).

In such DiD examples with panel data, the cluster-robust standard errors can be much larger than the default because both the regressor of interest and the errors are highly correlated within cluster. Note also that this complication can exist even with the inclusion of fixed effects (see Section III).

The same problems arise if we additionally have data on individuals, so that

$$y_{its} = \gamma \times d_{ts} + \boldsymbol{z}_{its}'\boldsymbol{\delta} + \alpha_s + \delta_t + u_{its}.$$

In an influential paper, Bertrand, Duflo and Mullainathan (2004) demonstrated the importance of using cluster-robust standard errors in DiD settings. Furthermore, the clustering should be on state, assuming error independence across states. The clustering should not be on state-year pairs since, for example, the error for California in 2010 is likely to be correlated with the error for California in 2009.

The issues raised here are relevant for any panel data application, not just DiD studies. The DiD panel example with binary policy regressor is often emphasized in the cluster-robust literature because it is widely used and it has a regressor that is highly serially correlated, even after mean-differencing to control for fixed effects. This serial correlation leads to a potentially large difference between cluster-robust and default standard errors.

C. The Cluster-Robust Variance Matrix Estimate

Stacking all observations in the g^{th} cluster, the model (4) can be written as

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{u}_g, \quad g = 1, \ldots, G,$$

where $\boldsymbol{y}_g$ and $\boldsymbol{u}_g$ are $N_g \times 1$ vectors, $\boldsymbol{X}_g$ is an $N_g \times K$ matrix, and there are N_g observations in cluster g. Further stacking $\boldsymbol{y}_g$, $\boldsymbol{X}_g$ and $\boldsymbol{u}_g$ over the G clusters then yields the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

The OLS estimator is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{y}_g.$$

In general, the variance matrix conditional on $\boldsymbol{X}$ is

$$\text{V}[\widehat{\boldsymbol{\beta}}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{B}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{7}$$

With

$$\boldsymbol{B} = \boldsymbol{X}'\text{V}[\boldsymbol{u}|\boldsymbol{X}]\boldsymbol{X}. \tag{8}$$

Given error independence across clusters, $\text{V}[\boldsymbol{u}|\boldsymbol{X}]$ has a block-diagonal structure, and

(8) simplifies to

$$\boldsymbol{B}_{\text{clu}} = \sum_{g=1}^{G} \boldsymbol{X}_g' \text{E}\left[\boldsymbol{u}_g \boldsymbol{u}_g' | \boldsymbol{X}_g\right] \boldsymbol{X}_g. \tag{9}$$

The matrix $\boldsymbol{B}_{\text{clu}}$, the middle part of the "sandwich matrix" (7), corresponds to the numerator of (2). $\boldsymbol{B}_{\text{clu}}$ can be written as:

$$\boldsymbol{B}_{\text{clu}} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \boldsymbol{x}_{ig} \boldsymbol{x}_{jg}' \omega_{ig,jg},$$

where $\omega_{ig,jg} = \text{E}[u_{ig} u_{jg} | \boldsymbol{X}_g]$ is the error covariance for the ig^{th} and jg^{th} observations. We can gain a few insights from inspecting this equation. The term $\boldsymbol{B}$ (and hence $\text{V}[\widehat{\boldsymbol{\beta}}]$) will be bigger when: (1) regressors within cluster are correlated, (2) errors within cluster are correlated so $\omega_{ig,jg}$ is non-zero, (3) N_g is large, and (4) the within-cluster regressor and error correlations are of the same sign (the usual situation). These conditions mirror the more precise Moulton result for the special case of equicorrelated errors given in equation (6). Both examples in Section II had high within-cluster correlation of the regressor, the DiD example additionally had high within-cluster (serial) correlation of the error and the Moulton (1990) example additionally had N_g large.

Implementation requires an estimate of $\boldsymbol{B}_{\text{clu}}$ given in (9). The cluster-robust estimate of the variance matrix (CRVE) of the OLS estimator is the sandwich estimate

$$\widehat{\text{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}] = (\boldsymbol{X}'\boldsymbol{X})^{-1} \widehat{\boldsymbol{B}}_{\text{clu}} (\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{10}$$

where

$$\widehat{\boldsymbol{B}}_{\text{clu}} = \sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\boldsymbol{u}}_g \widehat{\boldsymbol{u}}_g' \boldsymbol{X}_g, \tag{11}$$

and $\widehat{\boldsymbol{u}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g \widehat{\boldsymbol{\beta}}$ is the vector of OLS residuals for the g^{th} cluster. Formally (10)-(11) provides a consistent estimate of the variance matrix if $G^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\boldsymbol{u}}_g \widehat{\boldsymbol{u}}_g' \boldsymbol{X}_g - G^{-1} \sum_{g=1}^{G} \text{E}\left[\boldsymbol{X}_g' \boldsymbol{u}_g \boldsymbol{u}_g' \boldsymbol{X}_g\right] \xrightarrow{p} \boldsymbol{0}$ as $G \to \infty$. Initial derivations of this estimator, by White (1984, p.134-142) for balanced clusters, and by Liang and Zeger (1986) for unbalanced, assumed a finite number of observations per cluster. Hansen (2007a) showed that the CRVE can also be used if $N_g \to \infty$, the case for long panels, in addition to $G \to \infty$. Carter, Schnepel and Steigerwald (2013) consider unbalanced panels with either N_g fixed or $N_g \to \infty$. The sandwich formula for the CRVE extends to many estimators other than OLS; see Section VII.

Algebraically, the estimator (10)-(11) equals (7) and (9) with $\text{E}[\boldsymbol{u}_g \boldsymbol{u}_g']$ replaced with $\widehat{\boldsymbol{u}}_g \widehat{\boldsymbol{u}}_g'$. What is striking about this is that for each cluster g, the $N_g \times N_g$ matrix $\widehat{\boldsymbol{u}}_g \widehat{\boldsymbol{u}}_g'$ is bound to be a very poor estimate of the $N_g \times N_g$ matrix $\text{E}[\boldsymbol{u}_g \boldsymbol{u}_g']$ – there is no averaging going on to enable use of a Law of Large Numbers. The "magic" of the CRVE is that despite this, by averaging across all G clusters in (11), we are able to get a consistent variance estimate. This fact helps us to understand one of the limitations of this method in practice – the averaging that makes $\widehat{\text{V}}[\widehat{\boldsymbol{\beta}}]$ accurate for $\text{V}[\widehat{\boldsymbol{\beta}}]$ is an average based on the number of clusters G. In applications with few clusters this can lead to problems that we discuss below in Section VI.

Finite-sample modifications of (11) are typically used, to reduce downwards bias in

$\widehat{\mathbb{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}]$ due to finite G. Stata uses $\sqrt{c}\widehat{\boldsymbol{u}}_g$ in (11) rather than $\widehat{\boldsymbol{u}}_g$, with

$$c = \frac{G}{G-1}\frac{N-1}{N-K}. \tag{12}$$

In general $c \simeq G/(G-1)$, though see Subsection III.B for an important exception when fixed effects are directly estimated. Some other packages such as SAS use $c = G/(G-1)$, a simpler correction that is also used by Stata for extensions to nonlinear models. Either choice of c usually lessens, but does not fully eliminate, the usual downwards bias in the CRVE. Other finite-cluster corrections are discussed in Section VI, but there is no clear best correction.

D. Feasible GLS

If errors are correlated within cluster, then in general OLS is inefficient and feasible GLS may be more efficient.

Suppose we specify a model for $\Omega_g = \mathrm{E}[\boldsymbol{u}_g\boldsymbol{u}_g'|\boldsymbol{X}_g]$ in (9), such as within-cluster equicorrelation. Then the GLS estimator is $(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}\boldsymbol{y}$, where $\Omega = \mathrm{Diag}[\Omega_g]$. Given a consistent estimate $\widehat{\Omega}$ of Ω, the feasible GLS estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\text{FGLS}} = \left(\sum_{g=1}^{G}\boldsymbol{X}_g'\widehat{\Omega}_g^{-1}\boldsymbol{X}_g\right)^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g'\widehat{\Omega}_g^{-1}\boldsymbol{y}_g. \tag{13}$$

The FGLS estimator is second-moment efficient, with variance matrix

$$\widehat{\mathbb{V}}_{\text{def}}[\widehat{\boldsymbol{\beta}}_{\text{FGLS}}] = (\boldsymbol{X}'\widehat{\Omega}^{-1}\boldsymbol{X})^{-1}, \tag{14}$$

under the strong assumption that the error variance Ω is correctly specified.

Remarkably, the cluster-robust method of the previous section can be extended to FGLS. Essentially OLS is the special case where $\Omega_g = \sigma^2\boldsymbol{I}_{N_g}$. The cluster-robust estimate of the asymptotic variance matrix of the FGLS estimator is

$$\begin{aligned}&\widehat{\mathbb{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}_{\text{FGLS}}]\\ &= \left(\boldsymbol{X}'\widehat{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\widehat{\Omega}_g^{-1}\widehat{\boldsymbol{u}}_g\widehat{\boldsymbol{u}}_g'\widehat{\Omega}_g^{-1}\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\widehat{\Omega}^{-1}\boldsymbol{X}\right)^{-1},\end{aligned} \tag{15}$$

where $\widehat{\boldsymbol{u}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\beta}}_{\text{FGLS}}$. This estimator requires that $\boldsymbol{u}_g$ and $\boldsymbol{u}_h$ are uncorrelated when $g \neq h$, and that $G \to \infty$, but permits $\mathrm{E}[\boldsymbol{u}_g\boldsymbol{u}_g'|\boldsymbol{X}_g] \neq \Omega_g$. The approach of specifying a model for the error variances and then doing inference that guards against misspecification of this model is especially popular in the biostatistics literature that calls Ω_g a "working" variance matrix (see, for example, Liang and Zeger, 1986).

There are many possible candidate models for Ω_g, depending on the type of data being analyzed.

For individual-level data clustered by region, example 1 in Subsection II.B, a common starting point model is the random effects (RE) model. The error in model (4) is specified to have two components:

$$u_{ig} = \alpha_g + \varepsilon_{ig}, \tag{16}$$

where α_g is a cluster-specific error or common shock that is assumed to be independent and identically distributed (i.i.d.) $(0, \sigma_\alpha^2)$, and ε_{ig} is an idiosyncratic error that is assumed to be i.i.d. $(0, \sigma_\varepsilon^2)$. Then $\mathrm{V}[u_{ig}] = \sigma_\alpha^2 + \sigma_\varepsilon^2$ and $\mathrm{Cov}[u_{ig}, u_{jg}] = \sigma_\alpha^2$ for $i \neq j$. It follows that the intraclass correlation of the error $\rho_u = \mathrm{Cor}[u_{ig}, u_{jg}] = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$, so this model implies equicorrelated errors within cluster. Richer models that introduce heteroskedasticity include random coefficients models and hierarchical linear models.

For panel data, example 2 in Subsection II.B, a range of time series models for u_{it} may be used, including autoregressive and moving average error models. Analysis of within-cluster residual correlation patterns after OLS estimation can be helpful in selecting a model for Ω_g.

Note that in all cases if cluster-specific fixed effects are included as regressors and N_{g} is small then bias-corrected FGLS should be used; see Subsection III.C.

The efficiency gains of FGLS need not be great. As an extreme example, with equicorrelated errors, balanced clusters, and all regressors invariant within cluster $(\boldsymbol{x}_{ig} = \boldsymbol{x}_g)$ the FGLS estimator equals the OLS estimator - and so there is no efficiency gain to FGLS. With equicorrelated errors and general $\boldsymbol{X}$, Scott and Holt (1982) provide an upper bound to the maximum proportionate efficiency loss of OLS, compared to the variance of the FGLS estimator, of $1/\left[1 + \frac{4(1-\rho_u)[1+(N_{max}\ -1)\rho_u}{(N_{max}\ \times\rho_u)^2}\right]$, $N_{\max} = \max\{N_1, \ldots, N_G\}$. This upper bound is increasing in the error correlation ρ_u and the maximum cluster size $N_{\max}$. For low ρ_u the maximal efficiency gain can be low. For example, Scott and Holt (1982) note that for $\rho_u = .05$ and $N_{\max} = 20$ there is at most a 12% efficiency loss of OLS compared to FGLS. With $\rho_u = 0.2$ and $N_{\max} = 100$ the efficiency loss could be as much as 86%, though this depends on the nature of $\boldsymbol{X}$.

There is no clear guide to when FGLS may lead to considerable improvement in efficiency, and the efficiency gains can be modest. However, especially in models without cluster-specific fixed effects, implementation of FGLS and use of (15) to guard against misspecification of Ω_g is straightforward. And even modest efficiency gains can be beneficial. It is remarkable that current econometric practice with clustered errors ignores the potential efficiency gains of FGLS.

E. Implementation for OLS and FGLS

For regression software that provides a cluster-robust option, implementation of the CRVE for OLS simply requires defining for each observation a cluster identifier variable that takes one of G distinct values according to the observation's cluster, and then passing this cluster identifier to the estimation command's cluster-robust option. For example, if the cluster identifier is `id_clu`, then Stata OLS command `regress y x` becomes `regress y x, vce(cluster id_clu)`.

Wald hypothesis tests and confidence intervals are then implemented in the usual way. In some cases, however, joint tests of several hypotheses and of overall statistical significance may not be possible. The CRVE $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ is guaranteed to be positive semi-definite, so the estimated variance of the individual components of $\widehat{\boldsymbol{\beta}}$ are necessarily nonnegative. But $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ is not necessarily positive definite, so it is possible that the variance matrix of linear combinations of the components of $\widehat{\boldsymbol{\beta}}$ is singular. The rank of $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ equals the rank of $\widehat{\boldsymbol{B}}$ defined in (11). Since $\widehat{\boldsymbol{B}} = \boldsymbol{C}'\boldsymbol{C}$, where $\boldsymbol{C}' = [\boldsymbol{X}_1'\widehat{\boldsymbol{u}}_1 \cdots \boldsymbol{X}_G'\widehat{\boldsymbol{u}}_G]$ is a $K \times G$ matrix, it follows that the rank of $\widehat{\boldsymbol{B}}$, and hence that of $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$, is at most the rank of $\boldsymbol{C}$. Since $\boldsymbol{X}_1'\widehat{\boldsymbol{u}}_1 + \cdots + \boldsymbol{X}_G'\widehat{\boldsymbol{u}}_G = \boldsymbol{0}$, the rank of $\boldsymbol{C}$ is at most the minimum of K and $G - 1$. Effectively, the rank of $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ equals $\min(K, G - 1)$, though it can be less than this in some examples such as perfect collinearity of regressors and cluster-specific dummy regressors (see Subsection III.B

for the latter).

A common setting is to have a richly specified model with thousands of observations in far fewer clusters, leading to more regressors than clusters. Then $\widehat{\mathbf{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}]$ is rank-deficient, so it will not be possible to perform an overall F test of the joint statistical significance of all regressors. And in a log-wage regression with occupation dummies and clustering on state, we cannot test the joint statistical significance of the occupation dummies if there are more occupations than states. But it is still okay to perform statistical inference on individual regression coefficients, and to do joint tests on a limited number of restrictions (potentially as many as $\min(K, G-1)$).

Regression software usually also includes a panel data component. Panel commands may enable not only OLS with cluster-robust standard errors, but also FGLS for some models of within-cluster error correlation with default (and possibly cluster-robust) standard errors. It is important to note that those panel data commands that do not explicitly use time series methods, an example is FGLS with equicorrelation of errors within-cluster, can be applied more generally to other forms of clustered data, such as individual-level data with clustering on geographic region.

For example, in Stata first give the command `xtset id_clu` to let Stata know that the cluster-identifier is variable `id_clu`. Then the Stata command `xtreg y x, pa corr(ind) vce(robust)` yields OLS estimates with cluster-robust standard errors. Note that for Stata `xt` commands, option `vce(robust)` is generally interpreted as meaning cluster-robust; this is always the case from version 12.1 on. The `xt` commands use standard normal critical values whereas command `regress` uses Student's $T(G-1)$ critical values; see Sections VI and VIIA for further discussion.

For FGLS estimation the commands vary with the model for Ω_g. For equicorrelated errors, a starting point for example 1 in Subsection II.B, the FGLS estimator can be obtained using command `xtreg y x, pa corr(exch)` or command `xtreg y x, re`; slightly different estimates are obtained due to slightly different estimates of the equicorrelation. For FGLS estimation of hierarchical models that are richer than a random effects model, use Stata command `mixed` (version 13) or `xtmixed` (earlier versions). For FGLS with panel data and time variable `time`, first give the command `xtset id_clu time` to let Stata know both the cluster-identifier and time variable. A starting point for example 2 in Subsection II.B is an autoregressive error of order one, estimated using command `xtreg y x, pa corr(ar 1)`. Stata permits a wide range of possible models for serially correlated errors.

In all of these FGLS examples the reported standard errors are the default ones that assume correct specification of Ω_g. Better practice is to add option `vce(robust)` for `xtreg` commands, or option `vce(cluster id_clu)` for `mixed` commands, as this yields standard errors that are based on the cluster-robust variance defined in (15).

F. Cluster-Bootstrap Variance Matrix Estimate

Not all econometrics packages compute cluster-robust variance estimates, and even those that do may not do so for all estimators. In that case one can use a pairs cluster bootstrap that, like the CRVE, gives a consistent estimate of $\text{V}[\widehat{\boldsymbol{\beta}}]$ when errors are clustered.

To implement this bootstrap, do the following steps B times: (1) form G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), \ldots, (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample of clusters, and (2) compute the estimate of $\boldsymbol{\beta}$, denoted $\widehat{\boldsymbol{\beta}}_b$ in the b^{th} bootstrap sample. Then, given the B estimates $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_B$, compute the variance of these

$$\widehat{\mathbb{V}}_{\text{clu;boot}}[\widehat{\boldsymbol{\beta}}] = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\beta}}_b - \overline{\widehat{\boldsymbol{\beta}}})(\widehat{\boldsymbol{\beta}}_b - \overline{\widehat{\boldsymbol{\beta}}})',$$

where $\overline{\widehat{\boldsymbol{\beta}}} = B^{-1}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}_b$ and $B = 400$ should be more than adequate in most settings. It is important that the resampling be done over entire clusters, rather than over individual observations. Each bootstrap resample will have exactly G clusters, with some of the original clusters not appearing at all while others of the original clusters may be repeated in the resample two or more times. The terms "pairs" is used as $(\boldsymbol{y}_g, \boldsymbol{X}_g)$ are resampled as a pair. The term "nonparametric" is also used for this bootstrap. Some alternative bootstraps hold $\boldsymbol{X}_g$ fixed while resampling. For finite clusters, if $\widehat{\mathbb{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}]$ uses $\sqrt{c}\hat{u}_g$ in (11) then for comparability $\widehat{\mathbb{V}}_{\text{clu;boot}}[\widehat{\boldsymbol{\beta}}]$ should be multiplied by the constant c defined in (12). The pairs cluster bootstrap leads to a cluster-robust variance matrix for OLS with rank K even if $K > G$.

An alternative resampling method that can be used is the leave-one-cluster-out jackknife. Then, letting $\widehat{\boldsymbol{\beta}}_g$ denote the estimator of $\boldsymbol{\beta}$ when the g^{th} cluster is deleted,

$$\widehat{\mathbb{V}}_{\text{clu;jack}}[\widehat{\boldsymbol{\beta}}] = \frac{G-1}{G}\sum_{g=1}^{G}(\widehat{\boldsymbol{\beta}}_g - \overline{\widehat{\boldsymbol{\beta}}})(\widehat{\boldsymbol{\beta}}_g - \overline{\widehat{\boldsymbol{\beta}}})',$$

where $\overline{\widehat{\boldsymbol{\beta}}} = G^{-1}\sum_{g=1}^{G}\widehat{\boldsymbol{\beta}}_g$. This older method can be viewed as an approximation to the bootstrap that does not work as well for nonlinear estimators. It is used less often than the bootstrap, and has the same rank as the CRVE.

Unlike a percentile-t cluster bootstrap, presented later, the pairs cluster bootstrap and cluster jackknife variance matrix estimates are no better asymptotically than the CRVE, so it is best and quickest to use the CRVE if it is available. But the CRVE is not always available, especially for estimators more complicated than OLS. In that case one can instead use the pairs cluster bootstrap, though see the end of Subsection VI.C for potential pitfalls if there are few clusters, or even the cluster jackknife.

In Stata the pairs cluster bootstrap for OLS without fixed effects can be implemented in several equivalent ways including: `regress y x, vce(boot, cluster(id_clu) reps(400) seed(10101))`; `xtreg y x, pa corr(ind) vce(boot, reps(400) seed(10101))`; and `bootstrap, cluster(id_clu) reps(400) seed(10101) : regress y x`. The last variant can be used for estimation commands and user-written programs that do not have a `vce(boot)` option. We recommend 400 bootstrap iterations for published results and for replicability one should always set the seed.

For the jackknife the commands are instead, respectively, `regress y x, vce(jack, cluster(id_clu))`; `xtreg y x, pa corr(ind) vce(jack)`; and `jackknife, cluster(id_clu): regress y x`. For Stata `xt` commands, options `vce(boot)` and `vce(jack)` are generally interpreted as meaning cluster bootstrap and cluster jackknife; always so from Stata 12.1 on.

III. Cluster-Specific Fixed Effects

The cluster-specific fixed effects (FE) model includes a separate intercept for each cluster, so

$$y_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta} + \alpha_g + u_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta} + \sum_{h=1}^{G}\alpha_g dh_{ig} + u_{ig}, \quad (17)$$

where dh_{ig}, the h^{th} of G dummy variables, equals one if the ig^{th} observation is in cluster h and equals zero otherwise.

There are several different ways to obtain the same cluster-specific fixed effects estimator. The two most commonly-used are the following. The least squares dummy variable (LSDV) estimator directly estimates the second line of (17), with OLS regression of y_{ig} on $\boldsymbol{x}_{ig}$ and the G dummy variables $d1_{ig}, \ldots, dG_{ig}$, in which case the dummy variable coefficients $\hat{\alpha}_g = \bar{y}_g - \bar{\boldsymbol{x}}_g'\hat{\boldsymbol{\beta}}$ where $\bar{y}_g = N_g^{-1}\sum_{i=1}^{G} y_{ig}$ and $\bar{\boldsymbol{x}}_g = N_g^{-1}\sum_{i=1}^{G} \boldsymbol{x}_{ig}$. The within estimator, also called the fixed effects estimator, estimates $\boldsymbol{\beta}$ just by OLS regression in the within or mean-differenced model

$$(y_{ig} - \bar{y}_g) = (\boldsymbol{x}_{ig} - \bar{\boldsymbol{x}}_g)'\boldsymbol{\beta} + (u_{ig} - \bar{u}_g). \tag{18}$$

The main reason that empirical economists use the cluster-specific FE estimator is that it controls for a limited form of endogeneity of regressors. Suppose in (17) that we view $\alpha_g + u_{ig}$ as the error, and the regressors $\boldsymbol{x}_{ig}$ are correlated with this error, but only with the cluster-invariant component, i.e., $\text{Cov}[\boldsymbol{x}_{ig}, \alpha_g] \neq \mathbf{0}$ while $\text{Cov}[\boldsymbol{x}_{ig}, u_{ig}] = \mathbf{0}$. Then OLS and FGLS regression of y_{ig} on $\boldsymbol{x}_{ig}$, as in Section II, leads to inconsistent estimation of $\boldsymbol{\beta}$. Mean-differencing (17) leads to the within model (18) that has eliminated the problematic cluster-invariant error component α_g. The resulting FE estimator is consistent for $\boldsymbol{\beta}$ if either $G \longrightarrow \infty$ or $N_g \longrightarrow \infty$.

The cluster-robust variance matrix formula given in Section II carries over immediately to OLS estimation in the FE model, again assuming $G \longrightarrow \infty$.

In the remainder of this section we consider some practicalities. First, including fixed effects generally does not control for all the within-cluster correlation of the error and one should still use the CRVE. Second, when cluster sizes are small and degrees-of-freedom corrections are used the CRVE should be computed by within rather than LSDV estimation. Third, FGLS estimators need to be bias-corrected when cluster sizes are small. Fourth, tests of fixed versus random effects models should use a modified version of the Hausman test.

A. Do Fixed Effects Fully Control for Within-Cluster Correlation?

While cluster-specific effects will control for part of the within-cluster correlation of the error, in general they will not completely control for within-cluster error correlation (not to mention heteroskedasticity). So the CRVE should still be used. There are several ways to make this important point.

Suppose we have data on students in classrooms in schools. A natural model, a special case of a hierarchical model, is to suppose that there is both an unobserved school effect and, on top of that, an unobserved classroom effect. Letting i denote individual, s school, and c classroom, we have $y_{isc} = \boldsymbol{x}_{isc}'\boldsymbol{\beta} + \alpha_s + \delta_c + \varepsilon_{isc}$. A regression with school-level fixed effects (or random effects) will control for within-school correlation, but not the additional within-classroom correlation.

Suppose we have a short panel (T fixed, $N \to \infty$) of uncorrelated individuals and estimate $y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}$. Then the error u_{it} may be correlated over time (i.e., within-cluster) due to omitted factors that evolve progressively over time. Inoue and Solon (2006) provide a test for this serial correlation. Cameron and Trivedi (2005, p.710) present an FE individual-level panel data log-earnings regressed on log-hours example with cluster-robust standard errors four times the default. Serial correlation in the error may be due to omitting lagged y as a regressor. When $y_{i,t-1}$ is included as an additional regressor in the FE model, the Arellano-Bond estimator is used and even with $y_{i,t-1}$ included the Arellano-Bond

methodology requires that we first test whether the remaining error u_{it} is serially correlated.

Finally, suppose we have a single cross-section (or a single time series). This can be viewed as regression on a single cluster. Then in the model $y_i = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$ (or $y_t = \alpha + \boldsymbol{x}_t'\boldsymbol{\beta} + u_t$), the intercept is the cluster-specific fixed effect. There are many reasons for why the error u_i (or u_t) may be correlated in this regression.

B. Cluster-Robust Variance Matrix with Fixed Effects

Since inclusion of cluster-specific fixed effects may not fully control for cluster correlation (and/or heteroskedasticity), default standard errors that assume u_{ig} to be i.i.d. may be invalid. So one should use cluster-robust standard errors.

Arellano (1987) showed that $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ defined in (10)-(11) remains valid for the within estimator that controls for inclusion of G cluster-specific fixed effects, provided $G \to \infty$ and N_g is small. If instead one obtains the LSDV estimator, the CRVE formula gives the same CRVE for $\widehat{\boldsymbol{\beta}}$ as that for the within estimator, with the important proviso that the same degrees-of-freedom adjustment must be used – see below. The fixed effects estimates $\hat{\alpha}_g$ obtained for the LSDV estimator are essentially based only on N_g observations, so $\widehat{\mathrm{V}}[\hat{\alpha}_g]$ is inconsistent for $\mathrm{V}[\hat{\alpha}_g]$, just as $\hat{\alpha}_g$ is inconsistent for α_g.

Hansen (2007a, p.600) shows that this CRVE can also be used if additionally $N_g \to \infty$, for both the case where within-cluster correlation is always present (e.g. for many individuals in each village) and for the case where within-cluster correlation eventually disappears (e.g. for panel data where time series correlation disappears for observations far apart in time). The rates of convergence are $\sqrt{G}$ in the first case and $\sqrt{GN_g}$ in the second case, but the same asymptotic variance matrix is obtained in either case. Kézdi (2004) analyzed the CRVE in FE models for a range of values of G and N_g.

It is important to note that while LSDV and within estimation lead to identical estimates of $\boldsymbol{\beta}$, they can yield different standard errors due to different finite sample degrees-of-freedom correction.

It is well known that if default standard errors are used, i.e. it is assumed that u_{ig} in (17) is i.i.d., then one can safely use standard errors after LSDV estimation as this correctly views the number of parameters as $G + K$ rather than K. If instead the within estimator is used, however, manual OLS estimation of (18) will mistakenly view the number of parameters to equal K rather than $G + K$. (Built-in panel estimation commands for the within estimator, i.e. a fixed effects command, should remain okay to use, since they should be programmed to use $G + K$ in calculating the standard errors.)

It is not well known that if cluster-robust standard errors are used, and cluster sizes are small, then inference should be based on the within estimator standard errors. We thank Arindrajit Dube and Jason Lindo for bringing this issue to our attention. Within and LSDV estimation lead to the same cluster-robust standard errors if we apply formula (11) to the respective regressions, or if we multiply this estimate by $c = G/(G-1)$. Differences arise, however, if we multiply by the small-sample correction c given in (12). Within estimation sets $c = \frac{G}{G-1} \times \frac{N-1}{N-(K-1)}$ since there are only $(K-1)$ regressors – the within model is estimated without an intercept. LSDV estimation uses $c = \frac{G}{G-1}\frac{N-1}{N-G-(K-1)}$ since the G cluster dummies are also included as regressors. For balanced clusters with $N_g = N_*$ and G large relative to K, $c \simeq 1$ for within estimation and $c \simeq N_*/(N_*-1)$ for LSDV estimation. Suppose there are only two observations per cluster, due to only two individuals per household or two time periods in a panel setting, so $N_g = N_* = 2$. Then $c \simeq 2/(2-1) = 2$ for LSDV estimation,

leading to CRVE that is twice that from within estimation. Within estimation leads to the correct finite-sample correction.

In Stata the within command `xtreg y x, fe vce(robust)` gives the desired CRVE. The alternative LSDV commands `regress y x i.id_clu, vce(cluster id_clu)` and, equivalently, `regress y x, absorb(id_clu) vce(cluster id_clu)` use the wrong degrees-of-freedom correction. If a CRVE is needed, then use `xtreg`. If there is reason to instead use `regress i.id` then the cluster-robust standard errors should be multiplied by the square root of $[N-(K-1)]/[N-G-(K-1)]$, especially if N_g is small.

The inclusion of cluster-specific dummy variables increases the dimension of the CRVE, but does not lead to a corresponding increase in its rank. To see this, stack the dummy variable dh_{ig} for cluster g into the $N_g \times 1$ vector $\boldsymbol{dh}_g$. Then $\boldsymbol{dh}_g'\hat{\boldsymbol{u}}_g = \mathbf{0}$, by the OLS normal equations, leading to the rank of $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ falling by one for each cluster-specific effect. If there are k regressors varying within cluster and $G-1$ dummies then, even though there are $K+G-1$ parameters $\boldsymbol{\beta}$, the rank of $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ is only the minimum of K and $G-1$. And a test that $\alpha_1,\ldots,\alpha_G$ are jointly statistically significant is a test of $G-1$ restrictions (since the intercept or one of the fixed effects needs to be dropped). So even if the cluster-specific fixed effects are consistently estimated (i.e., if $N_g \to \infty$), it is not possible to perform this test if $K < G-1$, which is often the case.

If cluster-specific effects are present then the pairs cluster bootstrap must be adapted to account for the following complication. Suppose cluster 3 appears twice in a bootstrap resample. Then if clusters in the bootstrap resample are identified from the original cluster-identifier, the two occurrences of cluster 3 will be incorrectly treated as one large cluster rather than two distinct clusters.

In Stata, the bootstrap option `idcluster` ensures that distinct identifiers are used in each bootstrap resample. Examples are `regress y x i.id_clu, vce(boot, cluster(id_clu) idcluster(newid) reps(400) seed(10101))` and, more simply, `xtreg y x, fe vce(boot, reps(400) seed(10101))`, as in this latter case Stata automatically accounts for this complication.

C. Feasible GLS with Fixed Effects

When cluster-specific fixed effects are present, more efficient FGLS estimation can become more complicated. In particular, if asymptotic theory relies on $G \to \infty$ with N_g fixed, the α_g cannot be consistently estimated. The within estimator of $\boldsymbol{\beta}$ is nonetheless consistent, as α_g disappears in the mean-differenced model. But the resulting residuals $\hat{u}_{ig}$ are contaminated, since they depend on both $\widehat{\boldsymbol{\beta}}$ and $\hat{\alpha}_g$, and these residuals will be used to form a FGLS estimator. This leads to bias in the FGLS estimator, so one needs to use bias-corrected FGLS unless $N_g \to \infty$. The correction method varies with the model for $\Omega_g = \mathrm{V}[\boldsymbol{u}_g]$, and currently there are no Stata user-written commands to implement these methods.

For panel data a commonly-used model specifies an AR(p) model for the errors u_{ig} in (17). If fixed effects are present, then there is a bias (of order N_g^{-1}) in estimation of the AR(p) coefficients. Hansen (2007b) obtains bias-corrected estimates of the AR(p) coefficients and uses these in FGLS estimation. Hansen (2007b) in simulations shows considerable efficiency gains in bias-corrected FGLS compared to OLS.

Brewer, Crossley, and Joyce (2013) consider a DiD model with individual-level U.S. panel data with $N = 750{,}127$, $T = 30$, and a placebo state-level law so clustering is on state with $G = 50$. They find that bias-corrected FGLS for AR(2) errors, using the Hansen (2007b)

correction, leads to higher power than FE estimation. In their example ignoring the bias correction does not change results much, perhaps because $T = 30$ is reasonably large.

For balanced clusters with Ω_g the same for all g, say $\Omega_g = \Omega_*$, and for N_g small, then the FGLS estimator in (13) can be used without need to specify a model for Ω_*. Instead we can let $\widehat{\Omega}_*$ have ij^{th} entry $G^{-1}\sum_{g=1}^{G}\hat{u}_{ig}\,\hat{u}_{jg}$, where $\hat{u}_{ig}$ are the residuals from initial OLS estimation. These assumptions may be reasonable for a balanced panel. Two complications can arise. First, even without fixed effects there may be many off-diagonal elements to estimate and this number can be large relative to the number of observations. Second, the fixed effects lead to bias in estimating the off-diagonal covariances. Hausman and Kuersteiner (2008) present fixes for both complications.

D. Testing the Need for Fixed Effects

FE estimation is often accompanied by a loss of precision in estimation, as only within-cluster variation is used (recall we regress $(y_{ig} - \bar{y}_g)$ on $(\boldsymbol{x}_{ig} - \bar{\boldsymbol{x}}_g)$). Furthermore, the coefficient of a cluster-invariant regressor is not identified, since then $x_{ig} - \bar{x}_g = 0$. Thus it is standard to test whether it is sufficient to estimate by OLS or FGLS, without cluster-specific fixed effects.

The usual test is a Hausman test based on the difference between the FE estimator, $\widehat{\boldsymbol{\beta}}_{\text{FE}}$, and the RE estimator, $\widehat{\boldsymbol{\beta}}_{\text{RE}}$. Let $\boldsymbol{\beta}_1$ denote a subcomponent of $\boldsymbol{\beta}$, possibly just the coefficient of a single regressor of key interest; at most $\boldsymbol{\beta}_1$ contains the coefficients of all regressors that are not invariant within cluster or, in the case of panel data, that are not aggregate time effects that take the same value for each individual. The chi-squared distributed test statistic is

$$\text{T}_{\text{Haus}} = (\widehat{\boldsymbol{\beta}}_{1;\text{FE}} - \widehat{\boldsymbol{\beta}}_{1;\text{RE}})'\widehat{\text{V}}^{-1}(\widehat{\boldsymbol{\beta}}_{1;\text{FE}} - \widehat{\boldsymbol{\beta}}_{1;\text{RE}}),$$

where $\widehat{\text{V}}$ is a consistent estimate of $\text{V}[\widehat{\boldsymbol{\beta}}_{1;\text{FE}} - \widehat{\boldsymbol{\beta}}_{1;\text{RE}}]$.

Many studies use the standard form of the Hausman test. This obtains $\widehat{\text{V}}$ under the strong assumption that $\widehat{\boldsymbol{\beta}}_{\text{RE}}$ is fully efficient under the null hypothesis. This requires the unreasonably strong assumptions that α_g and ε_{ig} in (16) are i.i.d., requiring that neither α_g nor ε_{ig} is heteroskedastic and that ε_{ig} has no within-cluster correlation. As already noted, these assumptions are likely to fail and one should not use default standard errors. Instead a CRVE should be used. For similar reasons the standard form of the Hausman test should not be used.

Wooldridge (2010, p.332) instead proposes implementing a cluster-robust version of the Hausman test by the following OLS regression

$$y_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta} + \bar{\boldsymbol{w}}_g'\boldsymbol{\gamma} + u_{ig},$$

where $\boldsymbol{w}_g$ denotes the subcomponent of $\boldsymbol{x}_{ig}$ that varies within cluster and $\bar{\boldsymbol{w}}_g = N_g^{-1}\sum_{i=1}^{G}\boldsymbol{w}_{ig}$. If $H_0\colon \boldsymbol{\gamma} = \boldsymbol{0}$ is rejected using a Wald test based on a cluster-robust estimate of the variance matrix, then the fixed effects model is necessary. The Stata user-written command `xtoverid`, due to Schaffer and Stillman (2010), implements this test.

An alternative is to use the pairs cluster bootstrap to obtain $\widehat{\text{V}}$, in each resample obtaining $\widehat{\boldsymbol{\beta}}_{1;\text{FE}}$ and $\widehat{\boldsymbol{\beta}}_{1;\text{RE}}$, leading to B resample estimates of $(\widehat{\boldsymbol{\beta}}_{1;\text{FE}} - \widehat{\boldsymbol{\beta}}_{1;\text{RE}})$. We are unaware of studies comparing these two cluster-robust versions of the Hausman test.

IV. What to Cluster Over?

It is not always clear what to cluster over - that is, how to define the clusters - and there may even be more than one way to cluster.

Before providing some guidance, we note that it is possible for cluster-robust errors to actually be smaller than default standard errors. First, in some rare cases errors may be negatively correlated, most likely when $N_g = 2$, in which case (6) predicts a reduction in the standard error. Second, cluster-robust is also heteroskedastic-robust and White heteroskedastic-robust standard errors in practice are sometimes larger and sometimes smaller than the default. Third, if clustering has a modest effect, so cluster-robust and default standard errors are similar in expectation, then cluster-robust may be smaller due to noise. In cases where the cluster-robust standard errors are smaller they are usually not much smaller than the default, whereas in other applications they can be much, much larger.

A. Factors Determining What to Cluster Over

There are two guiding principles that determine what to cluster over.

First, given $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ defined in (7) and (9) whenever there is reason to believe that both the regressors and the errors might be correlated within cluster, we should think about clustering defined in a broad enough way to account for that clustering. Going the other way, if we think that either the regressors or the errors are likely to be uncorrelated within a potential group, then there is no need to cluster within that group.

Second, $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ is an average of G terms that gets closer to $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ only as G gets large. If we define very large clusters, so that there are very few clusters to average over in equation (11), then the resulting $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ can be a very poor estimate of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$. This complication, and discussion of how few is "few", is the subject of Section VI.

These two principles mirror the bias-variance trade-off that is common in many estimation problems – larger and fewer clusters have less bias but more variability. There is no general solution to this trade-off, and there is no formal test of the level at which to cluster. The consensus is to be conservative and avoid bias and use bigger and more aggregate clusters when possible, up to and including the point at which there is concern about having too few clusters.

For example, suppose your dataset included individuals within counties within states, and you were considering whether to cluster at the county level or the state level. We have been inclined to recommend clustering at the state level. If there was within-state cross-county correlation of the regressors and errors, then ignoring this correlation (for example, by clustering at the county level) would lead to incorrect inference. In practice researchers often cluster at progressively higher (i.e., broader) levels and stop clustering when there is relatively little change in the standard errors. This seems to be a reasonable approach.

There are settings where one may not need to use cluster-robust standard errors. We outline several, though note that in all these cases it is always possible to still obtain cluster-robust standard errors and contrast them to default standard errors. If there is an appreciable difference, then use cluster-robust standard errors.

If a key regressor is randomly assigned within clusters, or is as good as randomly assigned, then the within-cluster correlation of the regressor is likely to be zero. Thus there is no need to cluster standard errors, even if the model's errors are clustered. In this setting, if there are additionally control variables of interest, and if these are not randomly assigned within cluster, then we may wish to cluster our standard errors for the sake of correct inference on the control variables.

If the model includes cluster-specific fixed effects, and we believe that within-cluster

correlation of errors is solely driven by a common shock process, then we may not be worried about clustering. The fixed effects will absorb away the common shock, and the remaining errors will have zero within-cluster correlation. More generally, control variables may absorb systematic within-cluster correlation. For example, in a state-year panel setting, control variables may capture the state-specific business cycle.

However, as already noted in Subsection III.A, the within-cluster correlation is usually not fully eliminated. And even if it is eliminated, the errors may still be heteroskedastic. Stock and Watson (2008) show that applying the usual White (1980) heteroskedastic-consistent variance matrix estimate to the FE estimator leads, surprisingly, to inconsistent estimation of $V[\widehat{\boldsymbol{\beta}}]$ if N_g is small (though it is correct if $N_g = 2$). They derive a bias-corrected formula for heteroskedastic-robust standard errors. Alternatively, and more simply, the CRVE is consistent for $V[\widehat{\boldsymbol{\beta}}]$, even if the errors are only heteroskedastic, though this estimator of $V[\widehat{\boldsymbol{\beta}}]$ is more variable.

Finally, as already noted in Subsection II.D we can always build a parametric model of the correlation structure of the errors and estimate by FGLS. If we believe that this parametric model captures the salient features of the error correlations, then default FGLS standard errors can be used.

B. Clustering Due to Survey Design

Clustering routinely arises due to the sampling methods used in complex surveys. Rather than randomly draw individuals from the entire population, costs are reduced by sampling only a randomly-selected subset of primary sampling units (such as a geographic area), followed by random selection, or stratified selection, of people within the chosen primary sampling units.

The survey methods literature uses methods to control for clustering that predate the cluster-robust approach of this paper. The loss of estimator precision due to clustered sampling is called the design effect: "The design effect or Deff is the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements" (Kish (1965), p.258)). Kish and Frankel (1974) give the variance inflation formula (6) in the non-regression case of estimation of the mean. Pfeffermann and Nathan (1981) consider the more general regression case. The CRVE is called the linearization formula, and the common use of $G - 1$ as the degrees of freedom used in hypothesis testing comes from the survey methods literature; see Shah, Holt and Folsom (1977) which predates the econometrics literature.

Applied economists routinely use data from complex surveys, controlling for clustering by using a cluster-robust variance matrix estimate. At the minimum one should cluster at the level of the primary sampling unit, though often there is reason to cluster at a broader level, such as clustering on state if regressors and errors are correlated within state.

The survey methods literature additionally controls for two other features of survey data – weighting and stratification. These methods are well-established and are incorporated in specialized software, as well as in some broad-based packages such as Stata. Bhattacharya (2005) provides a comprehensive treatment in a GMM framework.

If sample weights are provided then it is common to perform weighted least squares. Then the CRVE for $\widehat{\boldsymbol{\beta}}_{\text{WLS}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$ is that given in (15) with $\widehat{\Omega}_g^{-1}$ replaced by $\boldsymbol{W}_g$. The need to weight can be ignored if stratification is on only the exogenous regressors and we assume correct specification of the model, so that in our sample $E[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$. In that special case both weighted and unweighted estimators are consistent, and weighted OLS may actually be less efficient if, for example, model errors are i.i.d.; see, for example, Solon, Haider, and Wooldridge (2013). Another situation in which to use weighted least squares,

unrelated to complex surveys, is when data for the ig^{th} observation is obtained by in turn averaging N_{ig} observations and N_{ig} varies.

Stratification of the sample can enable more precise statistical inference. These gains can be beneficial in the nonregression case, such as estimating the monthly national unemployment rate. The gains can become much smaller once regressors are included, since these can partially control for stratification; see, for example, the application in Bhattacharya (2005). Econometrics applications therefore usually do not adjust standard errors for stratification, leading to conservative inference due to some relatively small over-estimation of the standard errors.

V. Multi-way Clustering

The discussion to date has presumed that if there is more than one potential way to cluster, these ways are nested in each other, such as households within states. But when clusters are non-nested, traditional cluster-robust inference can only deal with one of the dimensions.

In some applications it is possible to include sufficient regressors to eliminate concern about error correlation in all but one dimension, and then do cluster-robust inference for that remaining dimension. A leading example is that in a state-year panel there may be clustering both within years and within states. If the within-year clustering is due to shocks that are the same across all observations in a given year, then including year fixed effects as regressors will absorb within-year clustering, and inference then need only control for clustering on state.

When this approach is not applicable, the one-way cluster robust variance can be extended to multi-way clustering. Before discussing this topic, we highlight one error that we find some practitioners make, which is to cluster at the intersection of the two groupings. In the preceding example, some might be tempted to cluster at the state-year level. A Stata example is to use the command `regress y x, vce(cluster id_styr)` where `id_styr` is a state-year identifier. This will be very inadequate, since it imposes the restriction that observations are independent if they are in the same state but in different years. Indeed if the data is aggregated at the state-year level, there is only one observation at the state-year level, so this is identical to using heteroskedastic-robust standard errors, i.e. not clustering at all. This point was highlighted by Bertrand, Duflo, and Mullainathan (2004) who advocated clustering on the state.

A. Multi-way Cluster-Robust Variance Matrix Estimate

The cluster-robust estimate of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ defined in (10)-(11) can be generalized to clustering in multiple dimensions. In a change of notation, suppress the subscript for cluster and more simply denote the model for an individual observation as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i. \tag{19}$$

Regular one-way clustering is based on the assumption that $\mathrm{E}[u_i u_j|\boldsymbol{x}_i, \boldsymbol{x}_j] = 0$, unless observations i and j are in the same cluster. Then (11) sets $\widehat{\boldsymbol{B}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{x}_i\, \boldsymbol{x}_j'\hat{u}_i\hat{u}_j \mathbf{1}[i,j \text{ in same cluster}]$, where $\hat{u}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. In multi-way clustering, the key assumption is that $\mathrm{E}[u_i u_j|\boldsymbol{x}_i, \boldsymbol{x}_j] = 0$, unless observations i and j share any cluster dimension. Then the multi-way cluster robust estimate of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ replaces (11) with

$$\widehat{\boldsymbol{B}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{x}_i\, \boldsymbol{x}_j'\hat{u}_i\hat{u}_j \mathbf{1}[i,j \text{ share } \mathbf{any} \text{ cluster}]. \tag{20}$$

This method relies on asymptotics that are in the number of clusters of the dimension with the fewest number of clusters. This method is thus most appropriate when each dimension has many clusters.

Theory for two-way cluster robust estimates of the variance matrix is presented in Cameron, Gelbach, and Miller (2006, 2011), Miglioretti and Heagerty (2006), and Thompson (2006, 2011). See also empirical panel data applications by Acemoglu and Pischke (2003), who clustered at individual level and at region×time level, and by Petersen (2009), who clustered at firm level and at year level. Cameron, Gelbach and Miller (2011) present an extension to multi-way clustering. Like one-way cluster-robust, the method can be applied to estimators other than OLS.

For two-way clustering this robust variance estimator is easy to implement given software that computes the usual one-way cluster-robust estimate. First obtain three different cluster-robust "variance" matrices for the estimator by one-way clustering in, respectively, the first dimension, the second dimension, and by the intersection of the first and second dimensions. Then add the first two variance matrices and, to account for double-counting, subtract the third. Thus

$$\widehat{\mathbb{V}}_{2\text{way}}[\widehat{\boldsymbol{\beta}}] = \widehat{\mathbb{V}}_1[\widehat{\boldsymbol{\beta}}] + \widehat{\mathbb{V}}_2[\widehat{\boldsymbol{\beta}}] - \widehat{\mathbb{V}}_{1\cap 2}[\widehat{\boldsymbol{\beta}}], \tag{21}$$

where the three component variance estimates are computed using (10)-(11) for the three different ways of clustering.

We spell this out in a step-by-step fashion.

1. Identify your two ways of clustering. Make sure you have a variable that identifies each way of clustering. Also create a variable that identifies unique "group 1 by group 2" combinations. For example, suppose you have individual-level data spanning many U.S. states and many years, and you want to cluster on state and on year. You will need a variable for state (e.g. California), a variable for year (e.g. 1990), and a variable for state-by-year (California **and** 1990).
2. Estimate your model, clustering on "group 1". For example, regress y on $\boldsymbol{x}$, clustering on state. Save the variance matrix as $\widehat{\mathbb{V}}_1$.
3. Estimate your model, clustering on "group 2". For example, regress y on $\boldsymbol{x}$, clustering on year. Save the variance matrix as $\widehat{\mathbb{V}}_2$.
4. Estimate your model, clustering on "group 1 by group 2". For example, regress y on $\boldsymbol{x}$, clustering on state-by-year. Save the variance matrix as $\widehat{\mathbb{V}}_{1\cap 2}$.
5. Create a new variance matrix $\widehat{\mathbb{V}}_{2\text{way}} = \widehat{\mathbb{V}}_1 + \widehat{\mathbb{V}}_2 - \widehat{\mathbb{V}}_{1\cap 2}$. This is your new two-way cluster robust variance matrix for $\widehat{\boldsymbol{\beta}}$.
6. Standard errors are the square root of the diagonal elements of this matrix.

If you are interested in only one coefficient, you can also just focus on saving the standard error for this coefficient in steps 2-4 above, and then create $\text{se}_{2\text{way}} = \sqrt{\text{se}_1^2 + \text{se}_2^2 - \text{se}_{1\cap 2}^2}$.

In taking these steps, you should watch out for some potential pitfalls. With perfectly multicollinear regressors, such as inclusion of dummy variables some of which are redundant, a statistical package may automatically drop one or more variables to ensure a nonsingular set of regressors. If the package happens to drop different sets of variables in steps 2, 3, and 4, then the resulting $\widehat{\mathbb{V}}'$s will not be comparable, and adding them together in step 5 will give a nonsense result. To prevent this issue, manually inspect the estimation results in steps 2, 3, and 4 to ensure that each step has the same set of regressors, the same number of observations, etc. The only things that should be different are the reported standard errors and the reported

number of clusters.

B. Implementation

Unlike the standard one-way cluster case, $\widehat{\mathrm{V}}_{2\mathrm{way}}[\widehat{\boldsymbol{\beta}}]$ is not guaranteed to be positive semi-definite, so it is possible that it may compute negative variances. In some applications with fixed effects, $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ may be non positive-definite, but the subcomponent of $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ associated with the regressors of interest may be positive-definite. This may lead to an error message, even though inference is appropriate for the parameters of interest. Our informal observation is that this issue is most likely to arise when clustering is done over the same groups as the fixed effects. Few clusters in one or more dimensions can also lead to $\widehat{\mathrm{V}}_{2\mathrm{way}}[\widehat{\boldsymbol{\beta}}]$ being a non-positive-semidefinite matrix. Cameron, Gelbach and Miller (2011) present an eigendecomposition technique used in the time series HAC literature that zeroes out negative eigenvalues in $\widehat{\mathrm{V}}_{2\mathrm{way}}[\widehat{\boldsymbol{\beta}}]$ to produce a positive semi-definite variance matrix estimate.

The Stata user-written command `cmgreg`, available at the authors' websites, implements multi-way clustering for the OLS estimator with, if needed, the negative eigenvalue adjustment. The Stata add-on command `ivreg2`, due to Baum, Schaffer, and Stillman (2007), implements two-way clustering for OLS, IV and linear GMM estimation. Other researchers have also posted code, available from searching the web.

Cameron, Gelbach, and Miller (2011) apply the two-way method to data from Hersch (1998) that examines the relationship between individual wages and injury risk measured separately at the industry level and the occupation level. The log-wage for 5960 individuals is regressed on these two injury risk measures, with standard errors obtained by two-way clustering on 211 industries and 387 occupations. In that case two-way clustering leads to only a modest change in the standard error of the industry job risk coefficient compared to the standard error with one-way clustering on industry. Since industry job risk is perfectly correlated within industry, by result (6) we need to cluster on industry if there is any within-industry error correlation. By similar logic, the additional need to cluster on occupation depends on the within-occupation correlation of job industry risk, and this correlation need not be high. For the occupation job risk coefficient, the two-way and one-way cluster (on occupation) standard errors are similar. Despite the modest difference in this example, two-way clustering avoids the need to report standard errors for one coefficient clustering in one way and for the second coefficient clustering in the second way.

Cameron, Gelbach, and Miller (2011) also apply the two-way cluster-robust method to data on volume of trade between 98 countries with 3262 unique country pairs. In that case, two-way clustering on each of the countries in the country pair leads to standard errors that are 36% larger than one-way clustering an 230% more than heteroskedastic-robust standard errors. Cameron and Miller (2012) study such dyadic data in further detail. They note that two-way clustering does not pick up all the potential correlations in the data. Instead, more general cluster-robust methods, including one proposed by Fafchamps and Gubert (2007), should be used.

C. Feasible GLS

Similar to one-way clustering, FGLS is more efficient than OLS, provided an appropriate model for $\Omega = \mathrm{E}[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}]$ is specified and is consistently estimated.

The random effects model can be extended to multi-way clustering. For individual i in clusters g and h, the two-way random effects model specifies

$$y_{igh} = \boldsymbol{x}_{igh}'\boldsymbol{\beta} + \alpha_g + \delta_h + \varepsilon_{igh},$$

where the errors α_g, δ_h, and ε_{igh} are each assumed to be i.i.d. distributed with mean 0. For example, Moulton (1986) considered clustering due to grouping of regressors (schooling, age and weeks worked) in a log earnings regression, and estimated a model with common random shock for each year of schooling, for each year of age, and for each number of weeks worked.

The two-way random effects model can be estimated using standard software for (nested) hierarchical linear models; see, for example, Cameron and Trivedi (2009, ch. 9.5.7) for Stata command `xtmixed` (command `mixed` from version 13 on). For estimation of a many-way random effects model, see Davis (2002) who modeled film attendance data clustered by film, theater and time.

The default standard errors after FGLS estimation require that Ω is correctly specified. For two-way and multi-way random effects models, FGLS estimation entails transforming the data in such a way that there is no obvious method for computing a variance matrix estimate that is robust to misspecification of Ω. Instead if there is concern about misspecification of Ω then one needs to consider FGLS with richer models for Ω and assume that these are correctly specified - see Rabe-Hesketh and Skrondal (2012) for richer hierarchical models in Stata - or do less efficient OLS with two-way cluster-robust standard errors.

D. Spatial Correlation

Cluster-robust variance matrix estimates are closely related to spatial-robust variance matrix estimates.

In general, for model (19), $\widehat{\boldsymbol{B}}$ in (20) has the form

$$\widehat{\boldsymbol{B}} = \sum_{i=1}^{N}\sum_{j=1}^{N} w\,(i,j)\boldsymbol{x}_i\boldsymbol{x}_j'\hat{u}_i\hat{u}_j, \tag{22}$$

where $w(i,j)$ are weights. For cluster-robust inference these weights are either 1 (cluster in common) or 0 (no cluster in common). But the weights can also decay from one to zero, as in the case of the HAC variance matrix estimate for time series where $w(i,j)$ decays to zero as $|i-j|$ increases.

For spatial data it is assumed that model errors become less correlated as the spatial distance between observations grows. For example, with state-level data the assumption that model errors are uncorrelated across states may be relaxed to allow correlation that decays to zero as the distance between states gets large. Conley (1999) provides conditions under which (10) and (22) provide a robust variance matrix estimate for the OLS estimator, where the weights $w(i,j)$ decay with the spatial distance. The estimate (which Conley also generalizes to GMM models) is often called a spatial-HAC estimate, rather than spatial-robust, as proofs use mixing conditions (to ensure decay of dependence) as observations grow apart in distance. These conditions are not applicable to clustering due to common shocks which leads to the cluster-robust estimator with independence of observations across clusters.

Driscoll and Kraay (1998) consider panel data with T time periods and N individuals, with errors potentially correlated across individuals (and no spatial dampening), though this correlation across individuals disappears for observations that are more than m time periods apart. Let it denote the typical observation. The Driscoll-Kraay spatial correlation consistent (SCC) variance matrix estimate can be shown to use weight $w(it,js) = 1 - d(it,js)/(m+1)$ in (22), where the summation is now over i,j,s and t, and $d(it,js) = |t-s|$ if $|t-s| \leq m$ and $d(it,js) = 0$ otherwise. This method requires that the number of time periods $T \to \infty$, so is not suitable for short panels, while N may be fixed or $N \to \infty$. The Stata add-on command `xtscc`, due to Hoechle (2007), implements this variance estimator.

An estimator proposed by Thompson (2006) allows for across-cluster (in his example firm) correlation for observations close in time in addition to within-cluster correlation at any time separation. The Thompson estimator can be thought of as using $w(it,js) = \mathbf{1}[i,j$ share a firm, or $d(it,js) \leq m]$. Foote (2007) contrasts the two-way cluster-robust and these other variance matrix estimators in the context of a macroeconomics example. Petersen (2009) contrasts various methods for panel data on financial firms, where there is concern about both within firm correlation (over time) and across firm correlation due to common shocks.

Barrios, Diamond, Imbens, and Kolesár (2012) consider state-year panel data on individuals in states over years with state-level treatment and outcome (earnings) that is correlated spatially across states. This spatial correlation can be ignored if the state-level treatment is randomly assigned. But if the treatment is correlated over states (e.g. adjacent states may be more likely to have similar minimum wage laws) then one can no longer use standard errors clustered at the state level. Instead one should additionally allow for spatial correlation of errors across states. The authors additionally contrast traditional model-based inference with randomization inference.

In practice data can have cluster, spatial and time series aspects, leading to hybrids of cluster-robust, spatial-HAC and time-series HAC estimators. Furthermore, it may be possible to parameterize some of the error correlation. For example for a time series AR(1) error it may be preferable to use $\widehat{\mathrm{E}}[u_t u_s]$ based on an AR(1) model rather than $w(t,s)\hat{u}_t\hat{u}_s$. To date empirical practice has not commonly modeled these combined types of error correlations. This may become more common in the future.

VI. Few Clusters

We return to one-way clustering, and focus on the Wald "t-statistic"

$$w = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}}, \tag{23}$$

where β is one element in the parameter vector $\boldsymbol{\beta}$, and the standard error $s_{\hat{\beta}}$ is the square root of the appropriate diagonal entry in $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$. If $G \to \infty$ then $w \sim N[0,1]$ under $H_0: \beta = \beta_0$. For finite G the distribution of w is unknown, even with normal errors. It is common to use the T distribution with $G-1$ degrees of freedom.

It is not unusual for the number of clusters G to be quite small. Despite few clusters, $\hat{\beta}$ may still be a reasonably precise estimate of β if there are many observations per cluster. But with small G the asymptotics have not kicked in. Then $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ can be downwards-biased.

One should at a minimum use $T(G-1)$ critical values and $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ defined in (10)-(11) with residuals scaled by $\sqrt{c}$ where c is defined in (12) or $c = G/(G-1)$. Most packages rescale the residuals – Stata uses the first choice of c and SAS the second. The use of $T(G-1)$ critical values is less common. Stata uses the $T(G-1)$ distribution after command `regress y x, vce(cluster)`. But the alternative command `xtreg y x, vce(robust)` instead uses standard normal critical values.

Even with both of these adjustments, Wald tests generally over-reject. The extent of over-rejection depends on both how few clusters there are and the particular data and model used. In some cases the over-rejection is mild, in others cases a test with nominal size 0.05 may have true test size of 0.10.

The next subsection outlines the basic problem and discusses how few is "few" clusters. The subsequent three subsections present three approaches to finite-cluster correction – bias-corrected variance, bootstrap with asymptotic refinement, and use of the T distribution

with adjusted degrees-of-freedom. The final subsection considers special cases.

A. The Basic Problems with Few Clusters

There are two main problems with few clusters. First, OLS leads to "overfitting", with estimated residuals systematically too close to zero compared to the true error terms. This leads to a downwards-biased cluster-robust variance matrix estimate. The second problem is that even with bias-correction, the use of fitted residuals to form the estimate $\widehat{\boldsymbol{B}}$ of $\boldsymbol{B}$ leads to over-rejection (and confidence intervals that are too narrow) if the critical values are from the standard normal or even the $T(G-1)$ distribution.

For the linear regression model with independent homoskedastic normally distributed errors similar problems are easily controlled. An unbiased variance matrix is obtained by estimating the error variance σ^2 by $s^2 = \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}/(N-K)$ rather than $\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}/N$. This is the "fix" in the OLS setting for the first problem. The analogue to the second problem is that the $N[0,1]$ distribution is a poor approximation to the true distribution of the Wald statistic. In the i.i.d. case, the Wald statistic w can be shown to be exactly $T(N-K)$ distributed. For nonnormal homoskedastic errors the $T(N-K)$ is still felt to provide a good approximation, provided N is not too small. Both of these problems arise in the clustered setting, albeit with more complicated manifestations and fixes.

For independent heteroskedastic normally distributed errors there are no exact results. MacKinnon and White (1985) consider several adjustments to the heteroskedastic-consistent variance estimate of White (1980), including one called HC2 that is unbiased in the special case that errors are homoskedastic. Unfortunately if errors are actually heteroskedastic, as expected, the HC2 estimate is no longer unbiased for $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ – an unbiased estimator depends on the unknown pattern of heteroskedasticity and on the design matrix $\boldsymbol{X}$. And there is no way to obtain an exact T distribution result for w, even if errors are normally distributed. Other proposed solutions for testing and forming confidence intervals include using a T distribution with data-determined degrees of freedom, and using a bootstrap with asymptotic refinement.

In the following subsections we consider extensions of these various adjustments to the clustered case, where the problems can become even more pronounced.

Before proceeding we note that there is no specific point at which we need to worry about few clusters. Instead, "more is better". Current consensus appears to be that $G = 50$ is enough for state-year panel data. In particular, Bertrand, Duflo, and Mullainathan (2004, Table 8) find in their simulations that for a policy dummy variable with high within-cluster correlation, a Wald test based on the standard CRVE with critical value of 1.96 had rejection rates of .063, .058, .080, and .115 for number of states (G) equal to, respectively, 50, 20, 10 and 6. The simulations of Cameron, Gelbach and Miller (2008, Table 3), based on a quite different data generating process but again with standard CRVE and critical value of 1.96, had rejection rates of .068, .081, .118, and .208 for G equal to, respectively, 30, 20, 10 and 5. In both cases the rejection rates would also exceed .05 if the critical value was from the $T(G-1)$ distribution.

The preceding results are for balanced clusters. Cameron, Gelbach and Miller (2008, Table 4, column 8) consider unbalanced clusters when $G = 10$. The rejection rate with unbalanced clusters, half of size $N_g = 10$ and half of size 50, is $.183$, appreciably worse than rejection rates of $.126$ and $.115$ for balanced clusters of sizes, respectively, 10 and 100. Recent papers by Carter, Schnepel, and Steigerwald (2013) and Imbens and Kolesar (2012) provide theory that also indicates that the effective number of clusters is reduced when N_g varies across clusters; see also the simulations in MacKinnon and Webb (2013). Similar issues may also arise if the clusters are balanced, but estimation is by weighted LS that places different weights on different clusters. Cheng and Hoekstra (2013) document that weighting

can result in over-rejection in the panel DiD setting of Bertrand, Duflo, and Mullainathan (2004).

To repeat a key message, there is no clear-cut definition of "few". Depending on the situation "few" may range from less than 20 clusters to less than 50 clusters in the balanced case, and even more clusters in the unbalanced case.

B. Solution 1: Bias-Corrected Cluster-Robust Variance Matrix

A weakness of the standard CRVE with residual $\widehat{\boldsymbol{u}}_g$ is that it is biased for $V_{clu}[\widehat{\boldsymbol{\beta}}]$, since $E[\widehat{\boldsymbol{u}}_g\widehat{\boldsymbol{u}}_g'] \neq E[\boldsymbol{u}_g\boldsymbol{u}_g']$. The bias depends on the form of Ω_g but will usually be downwards. Several corrected residuals $\widetilde{\boldsymbol{u}}_g$ to use in place of $\widehat{\boldsymbol{u}}_g$ in (11) have been proposed. The simplest, already mentioned, is to use $\widetilde{\boldsymbol{u}}_g = \sqrt{G/(G-1)}\widehat{\boldsymbol{u}}_g$ or $\widetilde{\boldsymbol{u}}_g = \sqrt{c}\widehat{\boldsymbol{u}}_g$ where c is defined in (12). One should at least use either of these corrections.

Bell and McCaffrey (2002) use

$$\widetilde{\boldsymbol{u}}_g = [\boldsymbol{I}_{N_g} - \boldsymbol{H}_{gg}]^{-1/2}\widehat{\boldsymbol{u}}_g, \tag{24}$$

where $\boldsymbol{H}_{gg} = \boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g'$. This transformed residual leads to unbiased CRVE in the special case that $E[\boldsymbol{u}_g\boldsymbol{u}_g'] = \sigma^2\boldsymbol{I}$. This is a cluster generalization of the HC2 variance estimate of MacKinnon and White (1985), so we refer to it as the CR2VE.

Bell and McCaffrey (2002) also use

$$\widetilde{\boldsymbol{u}}_g = \sqrt{\frac{G-1}{G}}[\boldsymbol{I}_{N_g} - \boldsymbol{H}_{gg}]^{-1}\widehat{\boldsymbol{u}}_g. \tag{25}$$

This transformed residual leads to CRVE that can be shown to equal the delete-one-cluster jackknife estimate of the variance of the OLS estimator. This jackknife correction leads to upwards-biased CRVE if in fact $E[\boldsymbol{u}_g\boldsymbol{u}_g'] = \sigma^2\boldsymbol{I}$. This is a cluster generalization of the HC3 variance estimate of MacKinnon and White (1985), so we refer to it as the CR3VE.

Angrist and Pischke (2009, Chapter 8) and Cameron, Gelbach and Miller (2008) provide a more extensive discussion and cite more of the relevant literature. This literature includes papers that propose corrections for the more general case that $E[\boldsymbol{u}_g\boldsymbol{u}_g'] \neq \sigma^2\boldsymbol{I}$ but has a known parameterization, such as an RE model, and extension to generalized linear models.

Angrist and Lavy (2002) apply the CR2VE correction (24) in an application with $G = 30$ to 40 and find that the correction increases cluster-robust standard errors by between 10 and 50 percent. Cameron, Gelbach and Miller (2008, Table 3) find that the CR3VE correction (24) has rejection rates of .062, .070, .092, and .138 for G equal to, respectively, 30, 20, 10 and 5. These rejection rates are a considerable improvement on .068, .081, .118, and .208 for the standard CRVE, but there is still considerable over-rejection for very small G.

The literature has gravitated to using the CR2VE adjustment rather than the CR3VE adjustment. This reduces but does not eliminate over-rejection when there are few clusters.

C. Solution 2: Cluster Bootstrap with Asymptotic Refinement

In Subsection II.F we introduced the bootstrap as it is usually used, to calculate standard errors that rely on regular asymptotic theory. Here we consider a different use of the bootstrap, one with asymptotic refinement that may lead to improved finite-sample inference.

We consider inference based on $G \to \infty$ (formally, $\sqrt{G}(\hat{\beta} - \beta)$ has a limit normal

distribution). Then a two-sided Wald test of nominal size 0.05, for example, can be shown to have true size $0.05 + O(G^{-1})$ when the usual asymptotic normal approximation is used. For $G \to \infty$ this equals the desired 0.05, but for finite G this differs from 0.05. If an appropriate bootstrap with asymptotic refinement is instead used, the true size is $0.05 + O(G^{-3/2})$. This is closer to the desired 0.05 for large G, as $G^{-3/2} < G^{-1}$. Hopefully this is also the case for small G, something that is established using appropriate Monte Carlo experiments. For a one-sided test or a nonsymmetric two-sided test the rates are instead, respectively, $0.05 + O(G^{-1/2})$ and $0.05 + O(G^{-1})$.

Asymptotic refinement can be achieved by bootstrapping a statistic that is asymptotically pivotal, meaning the asymptotic distribution does not depend on any unknown parameters. The estimator $\hat{\beta}$ is not asymptotically pivotal as its distribution depends on $\mathrm{V}[\hat{\beta}]$ which in turn depends on unknown variance parameters in $\mathrm{V}[\boldsymbol{u}|\boldsymbol{X}]$. The Wald t-statistic defined in (23) is asymptotically pivotal as its asymptotic distribution is $N[0,1]$ which does not depend on unknown parameters.

1. Percentile-t Bootstrap

A percentile-t bootstrap obtains B draws, $w_1^*, \ldots, w_B^*$, from the distribution of the Wald t-statistic as follows. B times do the following:

1. Obtain G clusters $\{(\boldsymbol{y}_1^*, \boldsymbol{X}_1^*), \ldots, (\boldsymbol{y}_G^*, \boldsymbol{X}_G^*)\}$ by one of the cluster bootstrap methods detailed below.
2. Compute the OLS estimate $\hat{\beta}_b^*$ in this resample.
3. Calculate the Wald test statistic $w_b^* = (\hat{\beta}_b^* - \hat{\beta})/s_{\hat{\beta}_b^*}$ where $s_{\hat{\beta}_b^*}$ is the cluster-robust standard error of $\hat{\beta}_b^*$, and $\hat{\beta}$ is the OLS estimate of β from the original sample.

Note that we center on $\hat{\beta}$ and not β_0, as the bootstrap views the sample as the population, i.e., $\beta = \hat{\beta}$, and the B resamples are based on this "population." Note also that the standard error in step 3 needs to be cluster-robust. A good choice of B is $B = 999$; this is more than B for standard error estimation as tests are in the tails of the distribution, and is such that $(B+1)\alpha$ is an integer for common choices of test size α.

Let $w_{(1)}^*, \ldots, w_{(B)}^*$ denote the ordered values of $w_1^*, \ldots, w_B^*$. These ordered values trace out the density of the Wald t-statistic, taking the place of a standard normal or T distribution. For example, the critical values for a 95% nonsymmetric confidence interval or a 5% nonsymmetric Wald test are the lower 2.5 percentile and upper 97.5 percentile of $w_1^*, \ldots, w_B^*$, rather than, for example, the standard normal values of -1.96 and 1.96. The p-value for a symmetric test based on the original sample Wald statistic w equals the proportion of times that $|w| > |w_b^*|,\ b = 1, \ldots, B$.

The simplest cluster resampling method is the pairs cluster resampling introduced in Subsection II.F. Then in step 1. above we form G clusters $\{(\boldsymbol{y}_1^*, \boldsymbol{X}_1^*), \ldots, (\boldsymbol{y}_G^*, \boldsymbol{X}_G^*)\}$ by resampling with replacement G times from the original sample of clusters. This method has the advantage of being applicable to a wide range of estimators, not just OLS. However, for some types of data the pairs cluster bootstrap may not be applicable - see "Bootstrap with Caution" below.

Cameron, Gelbach, and Miller (2008) found that in Monte Carlos with few clusters the pairs cluster bootstrap did not eliminate test over-rejection. The authors proposed using an alternative percentile-t bootstrap, the wild cluster bootstrap, that holds the regressors fixed across bootstrap replications.

2. Wild Cluster Bootstrap

The wild cluster bootstrap resampling method is as follows. First, estimate the main model, imposing (forcing) the null hypothesis H_0 that you wish to test, to give estimate $\widetilde{\boldsymbol{\beta}}_{\mathrm{H_0}}$. For example, for test of statistical significance of a single variable regress y_{ig} on all components of $\boldsymbol{x}_{ig}$ except the variable that has regressor with coefficient zero under the null hypothesis. Form the residual $\tilde{u}_{ig} = y_i - \boldsymbol{x}_{ig}'\widetilde{\boldsymbol{\beta}}_{\mathrm{H_0}}$. Then obtain the b^{th} resample for step 1 above as follows:

1a. Randomly assign cluster g the weight $d_g = -1$ with probability 0.5 and the weight $d_g = 1$ with probability 0.5. All observations in cluster g get the same value of the weight.

1b. Generate new pseudo-residuals $u_{ig}^* = d_g \times \tilde{u}_{ig}$, and hence new outcome variables $y_{ig}^* = \boldsymbol{x}_{ig}'\widetilde{\boldsymbol{\beta}}_{\mathrm{H_0}} + u_{ig}^*$.

Then proceed with step 2 as before, regressing y_{ig}^* on $\boldsymbol{x}_{ig}$, and calculate w_b^* as in step 3. The p-value for the the test based on the original sample Wald statistic w equals the proportion of times that $|w| > |w_b^*|,\ b = 1, \ldots, B$.

For the wild bootstrap, the values $w_1^*, \ldots, w_B^*$ cannot be used directly to form critical values for a confidence interval. Instead they can be used to provide a p-value for testing a hypothesis. To form a confidence interval, one needs to invert a sequence of tests, profiling over a range of candidate null hypotheses $H_0: \beta = \beta_0$. For each of these null hypotheses, obtain the p-value. The 95% confidence interval is the set of values of β_0 for which $p \geq 0.05$. This method is computationally intensive, but conceptually straightforward. As a practical matter, you may want to ensure that you have the same set of bootstrap draws across candidate hypotheses, so as to not introduce additional bootstrapping noise into the determination of where the cutoff is.

In principle it is possible to directly use a bootstrap for bias-reduction, such as to remove bias in standard errors. In practice this is not done, however, as in practice any bias reduction comes at the expense of considerably greater variability. A conservative estimate of the standard error equals the width of a 95% confidence interval, obtained using asymptotic refinement, divided by 2×1.96.

Note that for the wild cluster bootstrap the resamples $\{(\boldsymbol{y}_1^*, \boldsymbol{X}_1), \ldots, (\boldsymbol{y}_G^*, \boldsymbol{X}_G)\}$ have the same $\boldsymbol{X}$ in each resample, whereas for pairs cluster both $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$ vary across the B resamples. The wild cluster bootstrap is an extension of the wild bootstrap proposed for heteroskedastic data. It works essentially because the two-point distribution for forming $\boldsymbol{u}_g^*$ ensures that $\mathrm{E}[\boldsymbol{u}_g^*] = \boldsymbol{0}$ and $\mathrm{V}[\boldsymbol{u}_g^*] = \widetilde{\boldsymbol{u}}_g\widetilde{\boldsymbol{u}}_g'$. There are other two-point distributions that also do so, but Davidson and Flachaire (2008) show that in the heteroskedastic case it is best to use the weights $d_g = \{-1,1\}$, called Rademacher weights.

The wild cluster bootstrap essentially replaces $\boldsymbol{y}_g$ in each resample with one of two values $\boldsymbol{y}_g^* = \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{H_0}} + \widetilde{\boldsymbol{u}}_g$ or $\boldsymbol{y}_g^* = \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{H_0}} - \widetilde{\boldsymbol{u}}_g$. Because this is done across G clusters, there are at most 2^G possible combinations of the data, so there are at most 2^G unique values of $w_1^*, \ldots, w_B^*$. If there are very few clusters then there is no need to actually bootstrap as we can simply enumerate, with separate estimation for each of the 2^G possible datasets.

Webb (2013) expands on these limitations. He shows that there are actually only 2^{G-1} possible t-statistics in absolute value. Thus with $G = 5$ there are at most $2^4 = 16$ possible values of $w_1^*, \ldots, w_B^*$. So if the main test statistic is more extreme than any of these 16 values, for example, then all we know is that the p-value is smaller than $1/16 = 0.0625$. Full enumeration makes this discreteness clear. Bootstrapping without consideration of this issue can lead to inadvertently picking just one point from the interval of equally plausible p-values.

As G gets to be as large as 11 this concern is less of an issue since $2^{10} = 1024$.
Webb (2013) proposes greatly reducing the discreteness of p-values with very low G by instead using a six-point distribution for the weights d_g in step 1b. In this proposed distribution the weights d_g have a $1/6$ chance of taking each value in $\{-\sqrt{1.5}, -\sqrt{1}, -\sqrt{.5}, \sqrt{.5}, \sqrt{1}, \sqrt{1.5}\}$. In his simulations this method outperforms the two-point wild bootstrap for $G < 10$ and is the preferred method to use if $G < 10$.

MacKinnon and Webb (2013) address the issue of unbalanced clusters and find that, even with $G = 50$, tests based on the standard CRVE with $T(G-1)$ critical values can over-reject considerably if the clusters are unbalanced. By contrast, the two-point wild bootstrap with Rademacher weights is generally reliable.

3. Bootstrap with Caution

Regardless of the bootstrap method used, pairs cluster with or without asymptotic refinement or wild cluster bootstrap, an important step when bootstrapping with few clusters is to examine the distribution of bootstrapped values. This is something that should be done whether you are bootstrapping β to obtain a standard error, or bootstrapping t-statistics with refinement to obtain a more accurate p-value. This examination can take the form of looking at four things: (1) basic summary statistics like mean and variance; (2) the sample size to confirm that it is the same as the number of bootstrap replications (no missing values); (3) the largest and smallest handful of values of the distribution; and (4) a histogram of the bootstrapped values.

We detail a few potential problems that this examination can diagnose.

First, if you are using a pairs cluster bootstrap and one cluster is an outlier in some sense, then the resulting histogram may have a big "mass" that sits separately from the rest of the bootstrap distribution – that is, there may be two distinct distributions, one for cases where that cluster is sampled and one for cases where it is not. If this is the case then you know that your results are sensitive to the inclusion of that cluster.

Second, if you are using a pairs cluster bootstrap with dummy right-hand side variables, then in some samples you can get no or very limited variation in treatment. This can lead to zero or near-zero standard errors. For a percentile-t pairs cluster bootstrap, a zero or missing standard error will lead to missing values for w^*, since the standard error is zero or missing. If you naively use the remaining distribution, then there is no reason to expect that you will have valid inference. And if the bootstrapped standard errors are zero plus machine precision noise, rather than exactly zero, very large t-values may result. Then your bootstrap distribution of t-statistics will have really fat tails, and you will not reject anything, even false null hypotheses. No variation or very limited variation in treatment can also result in many of your $\hat{\beta}^*$'s being "perfect fit" $\hat{\beta}^*$'s with limited variability. Then the bootstrap standard deviation of the $\hat{\beta}^*$'s will be too small, and if you use it as your estimated standard error you will over-reject. In this case we suggest using the wild cluster bootstrap.

Third, if your pairs cluster bootstrap samples draw nearly multicollinear samples, you can get huge $\hat{\beta}^*$'s. This can make a bootstrapped standard error seem very large. You need to determine what in the bootstrap samples "causes" the huge $\hat{\beta}^*$'s . If this is some pathological but common draw, then you may need to think about a different type of bootstrap, such as the wild cluster bootstrap, or give up on bootstrapping methods. For an extreme example, consider a DiD model, with first-order "control" fixed effects and an interaction term. Suppose that a bootstrap sample happens to have among its "treatment group" only observations where "post = 1". Then the variables "treated" and "treated*post" are collinear, and an OLS package will drop one of these variables. If it drops the "post" variable, it will report a coefficient on "treated*post", but this coefficient will not be a proper interaction term, it will instead also

include the level effect for the treated group.

Fourth, with less than ten clusters the wild cluster bootstrap should use the six-point version of Webb (2013).

Fifth, in general if you see missing values in your bootstrapped t-statistics, then you need to figure out why. Don't take your bootstrap results at face value until you know what's going on.

D. Solution 3: Improved Critical Values using a T-distribution

The simplest small-sample correction for the Wald statistic is to base inference on a T distribution, rather than the standard normal, with degrees of freedom at most the number of clusters G. Recent research has proposed methods that lead to using degrees of freedom much less than G, especially if clusters are unbalanced.

1. G-L Degrees of Freedom

Some packages, including Stata after command `regress`, use $G-1$ degrees of freedom for t-tests and F −tests based on cluster-robust standard errors. This choice emanates from the complex survey literature; see Bell and McCaffrey (2002) who note, however, that with normal errors this choice still tends to result in test over-rejection so the degrees of freedom should be even less than this.

Even the $T(G-1)$ can make quite a difference. For example with $G=10$ for a two-sided test at level 0.05 the critical value for $T(9)$ is 2.262 rather than 1.960, and if $w=1.960$ the p-value based on $T(9)$ is 0.082 rather than 0.05. In Monte Carlo simulations by Cameron, Gelbach, and Miller (2008) this choice works reasonably well, and at a minimum one should use the $T(G-1)$ distribution rather than the standard normal.

For models that include L regressors that are invariant within cluster, Donald and Lang (2007) provide a rationale for using the $T(G-L)$ distribution. If clusters are balanced and all regressors are invariant within cluster then the OLS estimator in the model $y_{ig}=\boldsymbol{x}_g'\boldsymbol{\beta}+u_{ig}$ is equivalent to OLS estimation in the grouped model $\bar{y}_g=\boldsymbol{x}_g'\boldsymbol{\beta}+\bar{u}_g$. If $\bar{u}_g$ is i.i.d. normally distributed then the Wald statistic is $T(G-L)$ distributed, where $\widehat{\mathbb{V}}[\widehat{\boldsymbol{\beta}}]=s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $s^2=(G-L)^{-1}\sum_g \widehat{\bar{u}_g}^2$. Note that $\bar{u}_g$ is i.i.d. normal in the random effects model if the error components are i.i.d. normal. Usually if there is a time-invariant regressor there is only one, in addition to the intercept, in which case $L=2$.

Donald and Lang extend this approach to inference on $\boldsymbol{\beta}$ in a model that additionally includes regressors $\boldsymbol{z}_{ig}$ that vary within clusters, and allow for unbalanced clusters, leading to $G-L$ for the RE estimator. Wooldridge (2006) presents an expansive exposition of the Donald and Lang approach. He also proposes an alternative approach based on minimum distance estimation. See Wooldridge (2006) and, for a summary, Cameron and Miller (2011).

2. Data-determined Degrees of Freedom

For testing the difference between two means of normally and independently distributed populations with different variances the t test is not exactly T distributed – this is known as the Behrens-Fisher problem. Satterthwaite (1946) proposed an approximation that was extended to regression with clustered errors by Bell and McCaffrey (2002) and developed further by Imbens and Kolesar (2012).

The $T(N-k)$ distribution is the ratio of $N[0,1]$ to independent $\sqrt{[\chi^2(N-K)]/(N-k)}$. For linear regression under i.i.d. normal errors, the derivation of the $T(N-k)$ distribution for the Wald t-statistic uses the result that $(N-K)(s_{\widehat{\beta}}^2/\sigma_{\widehat{\beta}}^2)\sim\chi^2(N-$

K), where $s^2_{\hat{\beta}}$ is the usual unbiased estimate of $\sigma^2_{\hat{\beta}} = \text{V}[\hat{\beta}]$. This result no longer holds for non-i.i.d. errors, even if they are normally distributed. Instead, an approximation uses the $T(v^*)$ distribution where v^* is such that the first two moments of $v^*(s^2_{\hat{\beta}}/\sigma^2_{\hat{\beta}})$ equal the first two moments (v^* and $2v^*$) of the $\chi^2(v^*)$ distribution. Assuming $s^2_{\hat{\beta}}$ is unbiased for $\sigma^2_{\hat{\beta}}$, $\text{E}[v^*(s^2_{\hat{\beta}}/\sigma^2_{\hat{\beta}})] = v^*$. And $\text{V}[v^*(s^2_{\hat{\beta}}/\sigma^2_{\hat{\beta}})] = 2v^*$ if $v^* = 2[(\sigma^2_{\hat{\beta}})^2/\text{V}[s^2_{\hat{\beta}}]]$.

Thus the Wald t-statistic is treated as being $T(v^*)$ distributed where $v^* = 2(\sigma^2_{\hat{\beta}})^2/V[s^2_{\hat{\beta}}]$. Assumptions are needed to obtain an expression for $\text{V}[s^2_{\hat{\beta}}]$. For clustered errors with $\boldsymbol{u} \sim N[\mathbf{0}, \Omega]$ and using the CRVE defined in Subsection II.C, or using CR2VE or CR3VE defined in Subsection VI.B, Bell and McCaffrey (2002) show that the distribution of the Wald t-statistic defined in (23) can be approximated by the $T(v^*)$ distribution where

$$v^* = \frac{(\sum_{j=1}^{G} \lambda_j)^2}{(\sum_{j=1}^{G} \lambda_j^2)}, \tag{26}$$

and λ_j are the eigenvalues of the $G \times G$ matrix $\boldsymbol{G}'\widehat{\Omega}\boldsymbol{G}$, where $\widehat{\Omega}$ is consistent for Ω, the $N \times G$ matrix $\boldsymbol{G}$ has g^{th} column $(\boldsymbol{I}_N - \boldsymbol{H})'_g \boldsymbol{A}_g \boldsymbol{X}_g (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{e}_k$, $(\boldsymbol{I}_N - \boldsymbol{H})_g$ is the $N_g \times N$ submatrix for cluster g of the $N \times N$ matrix $I_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{A}_g = (\boldsymbol{I}_{N_g} - \boldsymbol{H}_{gg})^{-1/2}$ for CR2VE, and $\boldsymbol{e}_k$ is a $K \times 1$ vector of zeroes aside from 1 in the k^{th} position if $\hat{\beta} = \hat{\beta}_k$. Note that v^* needs to be calculated separately, and differs, for each regression coefficient. The method extends to Wald tests based on scalar linear combinations $\boldsymbol{c}'\widehat{\boldsymbol{\beta}}$.

The justification relies on normal errors and knowledge of $\Omega = \text{E}[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}]$. Bell and McCaffrey (2002) perform simulations with balanced clusters ($G = 20$ and $N_g = 10$) and equicorrelated errors within cluster. They calculate v^* assuming $\Omega = \sigma^2\boldsymbol{I}$, even though errors are in fact clustered, and find that their method leads to Wald tests with true size closer to the nominal size than tests based on the conventional CRVE, CRV2E, and CRV3E.

Imbens and Kolesar (2012) additionally consider calculating v^* where $\widehat{\Omega}$ is based on equicorrelated errors within cluster. They follow the Monte Carlo designs of Cameron, Gelbach and Miller (2008), with $G = 5$ and 10 and equicorrelated errors. They find that all finite-sample adjustments perform better than using the standard CRVE with $T(G-1)$ critical values. The best methods use the CR2VE and $T(v^*)$, with slight over-rejection with v^* based on $\widehat{\Omega} = s^2\boldsymbol{I}$ (Bell and McCaffrey) and slight under-rejection with v^* based on $\widehat{\Omega}$ assuming equicorrelated errors (Imbens and Kolesar). For $G = 5$ these methods outperform the two-point wild cluster bootstrap, as expected given the very low G problem discussed in Subsection VI.C. More surprisingly these methods also outperform wild cluster bootstrap when $G = 10$, perhaps because Imbens and Kolesar (2012) may not have imposed the null hypothesis in forming the residuals for this bootstrap.

3. Effective Number of Clusters

Carter, Schnepel and Steigerwald (2013) propose a measure of the effective number of clusters. This measure is

$$G^* = \frac{G}{(1+\delta)}, \tag{27}$$

where $\delta = \frac{1}{G}\sum_{g=1}^{G}\{(\gamma_g - \bar{\gamma})^2/\bar{\gamma}^2\}$, $\gamma_g = \boldsymbol{e}'_k(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'_g\Omega_g\boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}\ \boldsymbol{e}_k$, $\boldsymbol{e}_k$ is a $K \times 1$

vector of zeroes aside from 1 in the k^{th} position if $\hat{\beta} = \hat{\beta}_k$, and $\bar{\gamma} = \frac{1}{G}\sum_{g=1}^{G}\gamma_g$. Note that G^* varies with the regression coefficient considered, and the method extends to Wald tests based on scalar linear combinations $\boldsymbol{c}'\widehat{\boldsymbol{\beta}}$.

The quantity δ measures cluster heterogeneity, which disappears if $\gamma_g = \gamma$ for all g. Given the formula for γ_g, cluster heterogeneity $(\delta \neq 0)$ can arise for many reasons, including variation in N_g, variation in $\boldsymbol{X}_g$ and variation in Ω_g across clusters.

In simulations using standard normal critical values, Carter et al. (2013) find that test size distortion occurs for $G^* < 20$. In application they assume errors are perfectly correlated within cluster, so $\Omega_g = \boldsymbol{l}\boldsymbol{l}'$ where $\boldsymbol{l}$ is an $N_g \times 1$ vector of ones. For data from the Tennessee STAR experiment they find $G^* = 192$ when $G = 318$. For the Hersch (1998) data of Subsection II.B, with very unbalanced clusters, they find for the industry job risk coefficient and with clustering on industry that $G^* = 19$ when $G = 211$.

Carter et al. (2013) do not actually propose using critical values based on the $T(G^*)$ distribution. The key component in obtaining the formula for v^* in the Bell and McCaffrey (2002) approach is determining $\mathrm{V}[s_{\hat{\beta}}^2/\sigma_{\hat{\beta}}^2]$, which equals $\mathrm{E}[(s_{\hat{\beta}}^2 - \sigma_{\hat{\beta}}^2)/\sigma_{\hat{\beta}}^2]$ given $s_{\hat{\beta}}^2$ is unbiased for $\sigma_{\hat{\beta}}^2$. Carter et al. (2013) instead work with $\mathrm{E}[(\tilde{s}_{\hat{\beta}}^2 - \sigma_{\hat{\beta}}^2)/\sigma_{\hat{\beta}}^2]$ where $\tilde{s}_{\hat{\beta}}^2$, defined in their paper, is an approximation to $s_{\hat{\beta}}^2$ that is good for large G (formally $\tilde{s}_{\hat{\beta}}^2/\sigma_{\hat{\beta}}^2 \rightarrow s_{\hat{\beta}}^2/\sigma_{\hat{\beta}}^2$ as $G \rightarrow \infty$). Now $\mathrm{E}[(\tilde{s}_{\hat{\beta}}^2 - \sigma_{\hat{\beta}}^2)/\sigma_{\hat{\beta}}^2] = 2(1+\delta)/G$, see Carter et al. (2013), where δ is defined in (27). This suggests using the $T(G^*)$ distribution as an approximation, and that this approximation will improve as G increases.

E. Special Cases

With few clusters, additional results can be obtained if there are many observations in each group. In DiD studies the few clusters problem arises if few groups are treated, even if G is large. And the few clusters problem is more likely to arise if there is multi-way clustering.

1. Fixed Number of Clusters with Cluster Size Growing

The preceding adjustments to the degrees of freedom of the T distribution are based on the assumption of normal errors. In some settings asymptotic results can be obtained when G is small, provided $N_g \rightarrow \infty$.

Bester, Conley and Hansen (2011), building on Hansen (2007a), give conditions under which the t-test statistic based on (11) is $\sqrt{G/(G-1)}$ times T_{G-1} distributed. Then using $\widetilde{\boldsymbol{u}}_g = \sqrt{G/(G-1)}\widehat{\boldsymbol{u}}_g$ in (11) yields a $T(G-1)$ distributed statistic. In addition to assuming G is fixed while $N_g \rightarrow \infty$, it is assumed that the within group correlation satisfies a mixing condition (this does not happen in all data settings, although it does for time series and spatial correlation), and that homogeneity assumptions are satisfied, including equality of plim $\frac{1}{N_g}\boldsymbol{X}_g'\boldsymbol{X}_g$ for all g.

Let $\hat{\beta}_g$ denote the estimate of parameter β in cluster g, $\overline{\hat{\beta}} = G^{-1}\sum_{g=1}^{G}\hat{\beta}_g$ denote the average of the G estimates, and $s_{\overline{\hat{\beta}}}^2 = (G-1)\sum_{g=1}^{G}(\hat{\beta}_g - \overline{\hat{\beta}})^2$ denote their variance. Suppose that the $\hat{\beta}_g$ are asymptotically normal as $N_g \rightarrow \infty$ with common mean β, and consider test of $H_0: \beta = \beta_0$ based on $t = \sqrt{G}(\hat{\beta}_g - \beta_0)/s_{\overline{\hat{\beta}}}$. Then Ibragimov and Müller (2010) show that tests based on the $T(G-1)$ distribution will be conservative tests (i.e., under-reject) for level $\alpha \leq 0.083$. This approach permits correct inference even with

extremely few clusters, assuming N_g is large. However, the requirement that cluster estimates are asymptotically independent must be met. Thus the method is not directly applicable to a state-year DiD application when there are year fixed effects (or other regressor that varies over time but not states). In that case Ibragimov and Müller propose applying their method after aggregating subsets of states into groups in which some states are treated and some are not.

2. Few Treated Groups

Problems arise if most of the variation in the regressor is concentrated in just a few clusters, even when G is sufficiently large. This occurs if the key regressor is a cluster-specific binary treatment dummy and there are few treated groups.

Conley and Taber (2011) examine a differences-in-differences (DiD) model in which there are few treated groups and an increasing number of control groups. If there are group-time random effects, then the DiD model is inconsistent because the treated groups random effects are not averaged away. If the random effects are normally distributed, then the model of Donald and Lang (2007) applies and inference can use a T distribution based on the number of treated groups. If the group-time shocks are not random, then the T distribution may be a poor approximation. Conley and Taber (2011) then propose a novel method that uses the distribution of the untreated groups to perform inference on the treatment parameter.

Abadie, Diamond and Hainmueller (2010) propose synthetic control methods that provide a data-driven method to select the control group in a DiD study, and that provide inference under random permutations of assignment to treated and untreated groups. The methods are suitable for treatment that effects few observational units.

3. What if you have multi-way clustering and few clusters?

Sometimes we are worried about multi-way clustering, but one or both of the ways has few clusters. Currently we are not aware of an ideal approach to deal with this problem. One potential solution is to try to add sufficient control variables so as to minimize concerns about clustering in one of the ways, and then use a one-way few-clusters cluster robust approach on the other way. Another potential solution is to model one of the ways of clustering in a parametric way, such as with a common shock or an autoregressive error model. Then you can construct a variance estimator that is a hybrid of the parametric model, and cluster robust in the remaining dimension.

VII. Extensions

The preceding material has focused on the OLS (and FGLS) estimator and tests on a single coefficient. The basic results generalize to multiple hypothesis tests, instrumental variables (IV) estimation, nonlinear estimators and generalized method of moments (GMM).

These extensions are incorporated in Stata, though Stata generally computes test p-values and confidence intervals using standard normal and chisquared distributions, rather than T and F distributions. And for nonlinear models stronger assumptions are needed to ensure that the estimator of $\boldsymbol{\beta}$ retains its consistency in the presence of clustering. We provide a brief overview.

A. Cluster-Robust F-tests

Consider Wald joint tests of several restrictions on the regression parameters. Except in the special case of linear restrictions and OLS with i.i.d. normal errors, asymptotic theory yields only a chi-squared distributed statistic, say W, that is $\chi^2(h)$ distributed, where h is the number of (linearly independent) restrictions.

Alternatively we can use the related F statistic, $F = W/h$. This yields the same p-value as the chi-squared test if we treat F as being $F(h,\infty)$ distributed. In the cluster case, a finite-sample adjustment instead treats F as being $F(h, G-1)$ distributed. This is analogous to using the $T(G-1)$ distribution, rather than $N[0,1]$, for a test on a single coefficient.

In Stata, the finite-sample adjustment of using the $T(G-1)$ for a t-test on a single coefficient, and using the $F(h, G-1)$ for an F-test, is only done after OLS regression with command `regress`. Otherwise Stata reports critical values and p-values based on the $N[0,1]$ and $\chi^2(h)$ distributions.

Thus Stata does no finite-cluster correction for tests and confidence intervals following instrumental variables estimation commands, nonlinear model estimation commands, or even after command `regress` in the case of tests and confidence intervals using commands `testnl` and `nlcom`. The discussion in Section VI was limited to inference after OLS regression, but it seems reasonable to believe that for other estimators one should also base inference on the $T(G-1)$ and $F(h, G-1)$ distributions, and even then tests may over-reject when there are few clusters.

Some of the few-cluster methods of Section VI can be extended to tests of more than one restriction following OLS regression. The Wald test can be based on the bias-adjusted variance matrices CR2VE or CR3VE, rather than CRVE. For a bootstrap with asymptotic refinement of a Wald test of $H_0: \boldsymbol{R\beta} = \boldsymbol{r}$, in the b^{th} resample we compute $W_b^* = (\boldsymbol{R\widehat{\beta}}_b^* - \boldsymbol{R\widehat{\beta}})'[\boldsymbol{R}\widehat{\mathbb{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}_b^*]\boldsymbol{R}']^{-1}(\boldsymbol{R\widehat{\beta}}_b^* - \boldsymbol{R\widehat{\beta}})$. Extension of the data-determined degrees of freedom method of Subsection VI.D to tests of more than one restriction requires, at a minimum, extension of Theorem 4 of Bell and McCaffrey (2002) from the case that covers β, where β is a single component of $\boldsymbol{\beta}$, to $\boldsymbol{R\beta}$. An alternative ad hoc approach would be to use the $F\left(h, \overline{v^*}\right)$ distribution where $\overline{v^*}$ is an average (possibly weighted by estimator precision) of v^* defined in (26) computed separately for each exclusion restriction.

For the estimators discussed in the remainder of Section VII, the rank of $\widehat{\mathbb{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}]$ is again the minimum of $G-1$ and the number of parameters (K). This means that at most $G-1$ restrictions can be tested using a Wald test, in addition to the usual requirement that $h \leq K$.

B. Instrumental Variables Estimators

The cluster-robust variance matrix estimate for the OLS estimator extends naturally to the IV estimator, the two-stage least squares (2SLS) estimator and the linear GMM estimator.

The following additional adjustments must be made when errors are clustered. First, a modified version of the Hausman test of endogeneity needs to be used. Second, the usual inference methods when instruments are weak need to be adjusted. Third, tests of over-identifying restrictions after GMM need to be based on an optimal weighting matrix that controls for cluster correlation of the errors.

1. IV and 2SLS

In matrix notation, the OLS estimator in the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ is inconsistent if $\mathrm{E}[\boldsymbol{u}|\boldsymbol{X}] \neq \boldsymbol{0}$. We assume existence of a set of instruments $\boldsymbol{Z}$ that satisfy $\mathrm{E}[\boldsymbol{u}|\boldsymbol{Z}] = \boldsymbol{0}$ and satisfy other conditions, notably $\boldsymbol{Z}$ is of full rank with $\dim[\boldsymbol{Z}] \geq \dim[\boldsymbol{X}]$ and $\mathrm{Cor}[\boldsymbol{Z}, \boldsymbol{X}] \neq \boldsymbol{0}$.

For the clustered case the assumption that errors in different clusters are uncorrelated is now one of uncorrelated errors conditional on the instruments $\boldsymbol{Z}$, rather than uncorrelated errors conditional on the regressors $\boldsymbol{X}$. In the g^{th} cluster the matrix of instruments $\boldsymbol{Z}_g$ is an $N_g \times M$ matrix, where $M \geq K$, and we assume that $\mathrm{E}[\boldsymbol{u}_g|\boldsymbol{Z}_g] = \boldsymbol{0}$ and $\mathrm{Cov}[\boldsymbol{u}_g\boldsymbol{u}_h'|\boldsymbol{Z}_g, \boldsymbol{Z}_h] =$

$\mathbf{0}$ for $g \neq h$.

In the just-identified case, with $\boldsymbol{Z}$ and $\boldsymbol{X}$ having the same dimension, the IV estimator is $\widehat{\boldsymbol{\beta}}_{\text{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}$, and the cluster-robust variance matrix estimate is

$$\widehat{\mathbb{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}_{\text{IV}}] = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{Z}_g'\widehat{\boldsymbol{u}}_g\widehat{\boldsymbol{u}}_g'\boldsymbol{Z}_g\right)(\boldsymbol{X}'\boldsymbol{Z})^{-1}, \tag{28}$$

where $\widehat{\boldsymbol{u}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\beta}}_{\text{IV}}$ are residuals calculated using the consistent IV estimator. We again assume $G \to \infty$. As for OLS, the CRVE may be rank-deficient with rank the minimum of K and $G - 1$.

In the over-identified case with $\boldsymbol{Z}$ having dimension greater than $\boldsymbol{X}$, the 2SLS estimator is the special case of the linear GMM estimator in (29) below with $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$, and the CRVE is that in (30) below with $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ and $\widehat{\boldsymbol{u}}_g$ the 2SLS residuals. In the just-identified case 2SLS is equivalent to IV.

A test for endogeneity of a regressor(s) can be conducted by comparing the OLS estimator to the 2SLS (or IV) estimator that controls for this endogeneity. The two estimators have the same probability limit given exogeneity and different probability limits given endogeneity. This is a classic setting for the Hausman test but, as in the Hausman test for fixed effects discussed in Subsection III.D, the standard version of the Hausman test cannot be used when errors are clustered. Instead partition $= [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$, where $\boldsymbol{X}_1$ is potentially endogenous and $\boldsymbol{X}_2$ is exogenous, and let $\widehat{\boldsymbol{v}}_{ig}$ denote the residuals from first-stage OLS regression of the endogenous regressors on instruments and exogenous regressors. Then estimate by OLS the model

$$y_{ig} = \boldsymbol{x}_{1ig}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2ig}'\boldsymbol{\beta}_2 + \widehat{\boldsymbol{v}}_{1ig}'\boldsymbol{\gamma} + u_{ig}.$$

The regressors $\boldsymbol{x}_1$ are considered endogenous if we reject $H_0: \boldsymbol{\gamma} = \mathbf{0}$ using a Wald test based on a CRVE. In Stata this is implemented using command `estat endogenous`. (Alternatively a pairs cluster bootstrap can be used to estimate the variance of $\widehat{\boldsymbol{\beta}}_{\text{2SLS}} - \widehat{\boldsymbol{\beta}}_{\text{OLS}}$).

2. Weak Instruments

When endogenous regressor(s) are weakly correlated with instrument(s), after partialling out the exogenous regressors in the model, there is great loss of precision. Then the standard error for the coefficient of the endogenous regressor is much higher after IV or 2SLS estimation than after OLS estimation.

Additionally, asymptotic theory takes an unusually long-time to kick in so that even with large samples the IV estimator can still have considerable bias and the Wald statistic is still not close to normally distributed. See, for example, Bound, Jaeger, and Baker (1995), Andrews and Stock (2007), and textbook discussions in Cameron and Trivedi (2005, 2009).

For this second consequence, called the "weak instrument" problem, the econometrics literature has focused on providing theory and guidance in the case of homoskedastic errors. Not all of the proposed methods extend to errors that are correlated within cluster. And the problem may even be greater in the clustered case, as the asymptotics are then in $G \to \infty$ rather than $N \to \infty$, though we are unaware of evidence on this.

We begin with case of a single endogenous regressor. A standard diagnostic for detecting weak instruments is to estimate by OLS the first-stage regression of the endogenous regressor on the exogenous regressors and the additional instrument(s). Then calculate the F-statistic for the joint significance of the instruments; in the case of a just-identified model

there is only one instrument to test so the F-statistic is the square of the t-statistic. With clustered errors, this F-statistic needs to be based on a cluster-robust variance matrix estimate. It is common practice to interpret the cluster-robust F-statistic in the same way as when errors are i.i.d., using the tables of Stock and Yogo (2005) or the popular rule-of-thumb, due to Staiger and Stock (1997), that there may be a weak instrument problem if $F < 10$. But it should be noted that these diagnostics for weak instruments were developed for the simpler case of i.i.d. errors. Note also that the first stage cluster-robust F-statistic can only be computed if the number of instruments is less than the number of clusters.

With more than one endogenous variable and i.i.d. errors the F-statistic generalizes to the Cragg-Donald minimum eigenvalue statistic, and one can again use the tables of Stock and Yogo (2005). For clustered errors generalizations of the Cragg-Donald minimum eigenvalue statistic have been proposed, see Kleibergen and Paap (2008), but it is again not clear whether these statistics can be compared to the Stock and Yogo tables when errors are clustered.

Now consider statistical inference that is valid even if instruments are weak, again beginning with the case of a single endogenous regressor. Among the several testing methods that have been proposed given i.i.d. errors, the Anderson-Rubin method can be generalized to the setting of clustered errors. Consider the model $y_{ig} = \beta x_{ig} + u_{ig}$, where the regressor x is endogenous and the first-stage equation is $x_{ig} = \boldsymbol{z}_{ig}'\boldsymbol{\pi} + v_{ig}$. (If there are additional exogenous regressors $\boldsymbol{x}_2$, as is usually the case, the method still works if the variables y, x and $\boldsymbol{z}$ are defined after partialling out $\boldsymbol{x}_2$.) The two equations imply that $y_{ig} - \beta^* x_{ig} = \boldsymbol{z}_{ig}'\boldsymbol{\pi}(\beta - \beta^*) + w_{ig}$, where $w_{ig} = u_{ig} + v_{ig}(\beta - \beta^*)$. So a test of $\beta = \beta^*$ is equivalent to a Wald test of $\boldsymbol{\gamma} = \mathbf{0}$ in the model $y_{ig} - \beta^* x_{ig} = \boldsymbol{z}_{ig}'\boldsymbol{\gamma} + w_{ig}$. With clustered errors the test is based on cluster-robust standard errors.

Additionally, a weak instrument 95% confidence interval for β can be constructed by regressing $y_{ig} - \beta^* x_{ig}$ on $\boldsymbol{z}_{ig}$ for a range of values of β^* and including in the confidence interval for β only those values of β^* for which we did not reject $H_0: \boldsymbol{\gamma} = \mathbf{0}$ when testing at 5%. As in the i.i.d. case, this can yield confidence intervals that are unbounded or empty, and the method loses power when the model is overidentified.

When there is more than one endogenous regressor this method can also be used, but it can only perform a joint F-test on the coefficients of all endogenous regressors rather than separate tests for each of the endogenous regressors.

Chernozhukov and Hansen (2008) provide a simple presentation of the method and an empirical example. Finlay and Magnusson (2009) provide this and other extensions, and provide a command `ivtest` for Stata. We speculate that if additionally there are few clusters, then some of the adjustments discussed in Section VI would help.

Baum, Schaffer and Stillman (2007) provide a comprehensive discussion of various methods for IV, 2SLS, limited information maximum likelihood (LIML), k-class, continuous updating and GMM estimation in linear models, and present methods using their `ivreg2` Stata command. They include weak instruments methods for errors that are i.i.d., heteroskedastic or within-cluster correlated errors.

3. Linear GMM

For over-identified models the linear GMM estimator is more efficient than the 2SLS estimator if $\mathrm{E}[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{Z}] \neq \sigma^2\boldsymbol{I}$. Then

$$\widehat{\boldsymbol{\beta}}_{\mathrm{GMM}} = (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{y}), \tag{29}$$

where $\boldsymbol{W}$ is a full rank $K \times K$ weighting matrix. The CRVE for GMM is

$$\widehat{\mathbf{V}}_{\text{clu}}[\widehat{\boldsymbol{\beta}}_{\text{GMM}}] = (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\left(\sum_{g=1}^{G}\boldsymbol{Z}_g'\widehat{\boldsymbol{u}}_g\widehat{\boldsymbol{u}}_g'\boldsymbol{Z}_g\right)\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}, \tag{30}$$

where $\widehat{\boldsymbol{u}}_g$ are residuals calculated using the GMM estimator.

For clustered errors, the efficient two-step GMM estimator uses $\boldsymbol{W} = (\sum_{g=1}^{G}\boldsymbol{Z}_g'\widehat{\boldsymbol{u}}_g\widehat{\boldsymbol{u}}_g'\boldsymbol{Z}_g)^{-1}$, where $\widehat{\boldsymbol{u}}_g$ are 2SLS residuals. Implementation of this estimator requires that the number of clusters exceeds the number of instruments, since otherwise $\sum_{g=1}^{G}\boldsymbol{Z}_g'\widehat{\boldsymbol{u}}_g\widehat{\boldsymbol{u}}_g'\boldsymbol{Z}_g$ is not invertible. Here $\boldsymbol{Z}$ contains both the exogenous regressors in the structural equation and the additional instruments required to enable identification of the endogenous regressors. When this condition is not met, Baum, Schaffer and Stillman (2007) propose doing two-step GMM after first partialling out the instruments $\boldsymbol{z}$ from the dependent variable y, the endogenous variables in the initial model $y_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta} + u_{ig}$, and any additional instruments that are not also exogenous regressors in this model.

The over-identifying restrictions (OIR) test, also called a Hansen test or a Sargan test, is a limited test of instrument validity that can be used when there are more instruments than necessary. When errors are clustered the OIR tests must be computed following the cluster version of two-step GMM estimation; see Hoxby and Paserman (1998).

Just as GLS is more efficient than OLS, specifying a model for $\Omega = \text{E}[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{Z}]$ can lead to more efficient estimation than GMM. Given a model for Ω, and conditional moment condition $\text{E}[\boldsymbol{u}|\boldsymbol{Z}] = \boldsymbol{0}$, a more efficient estimator is based on the unconditional moment condition $\text{E}[\boldsymbol{Z}'\Omega^{-1}\boldsymbol{u}] = \boldsymbol{0}$. Then we minimize $(\boldsymbol{Z}'\widehat{\Omega}^{-1}\boldsymbol{u})'(\boldsymbol{Z}'\widehat{\Omega}^{-1}\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\widehat{\Omega}^{-1}\boldsymbol{u})$, where $\widehat{\Omega}$ is consistent for Ω. Furthermore the CRVE can be robustified against misspecification of Ω, similar to the case of FGLS. In practice such FGLS-type improvements to GMM are seldom used, even in simpler settings that the clustered setting. An exception is Shore-Sheppard (1996) who considers the impact of equicorrelated instruments and group-specific shocks in a model similar to that of Moulton. One reason may be that this option is not provided in Stata command `ivregress`. In the special case of a random effects model for Ω, command `xtivreg` can be used along with a pairs cluster bootstrap used to guard against misspecification of Ω.

C. Nonlinear Models

For nonlinear models there are several ways to handle clustering. We provide a brief summary; see Cameron and Miller (2011) for further details.

For concreteness we focus on logit regression. Recall that in the cross-section case y_i takes value 0 or 1 and the logit model specifies that $\text{E}[y_i|\boldsymbol{x}_i] = \Pr[y_i = 1|\boldsymbol{x}_i] = \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})$, where $\Lambda(z) = e^z/(1+e^z)$.

1. Different Models for Clustering

The simplest approach is a pooled approach that assumes that clustering does not change the functional form for the conditional probability of a single observation. Thus, for the logit model, whatever the nature of clustering, it is assumed that

$$\text{E}[y_{ig}|\boldsymbol{x}_{ig}] = \Pr[y_{ig} = 1|\boldsymbol{x}_{ig}] = \Lambda(\boldsymbol{x}_{ig}'\boldsymbol{\beta}). \tag{31}$$

This is called a population-averaged approach, as $\Lambda(\boldsymbol{x}_{ig}'\boldsymbol{\beta})$ is obtained after averaging out any

within-cluster correlation. Inference needs to control for within-cluster correlation, however, and more efficient estimation may be possible.

The generalized estimating equations (GEE) approach, due to Liang and Zeger (1986) and widely used in biostatistics, introduces within-cluster correlation into the class of generalized linear models (GLM), a class that includes the logit model. One possible model for within-cluster correlation is equicorrelation, with $\text{Cor}\,[y_{ig}, y_{jg}|\boldsymbol{x}_{ig}, \boldsymbol{x}_{jg}] = \rho$. The Stata command `xtgee y x, family(binomial) link(logit) corr(exchangeable)` estimates the population-averaged logit model and provides the CRVE assuming the equicorrelation model for within-cluster correlation is correctly specified. The option `vce(robust)` provides a CRVE that is robust to misspecification of the model for within-cluster correlation. Command `xtgee` includes a range of models for the within-error correlation. The method is a nonlinear analog of FGLS given in Subsection II.D, and asymptotic theory requires $G \to \infty$.

A further extension is nonlinear GMM. For example, with endogenous regressors and instruments $\boldsymbol{z}$ that satisfy $\text{E}\,[y_{ig} - \exp(\boldsymbol{x}_{ig}'\boldsymbol{\beta})|\boldsymbol{z}_{ig}] = 0$, a nonlinear GMM estimator minimizes $\boldsymbol{h}(\boldsymbol{\beta})'\boldsymbol{W}\boldsymbol{h}(\boldsymbol{\beta})$ where $\boldsymbol{h}(\boldsymbol{\beta}) = \sum_g \sum_i \boldsymbol{z}_{ig}(y_{ig} - \exp(\boldsymbol{x}_{ig}'\boldsymbol{\beta}))$. Other choices of $\boldsymbol{h}(\boldsymbol{\beta})$ that allow for intracluster correlation may lead to more efficient estimation, analogous to the linear GMM example discussed at the end of Subsection VII.B. Given a choice of $\boldsymbol{h}(\boldsymbol{\beta})$, the two-step nonlinear GMM estimator at the second step uses weighting matrix $\boldsymbol{W}$ that is the inverse of a consistent estimator of $\text{V}[\boldsymbol{h}(\boldsymbol{\beta})]$, and one can then use the minimized objection function for an overidentifying restrictions test.

Now suppose we consider a random effects logit model with normally distributed random effect, so

$$\Pr[y_{ig} = 1|\alpha_g, \boldsymbol{x}_{ig}] = \Lambda(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta}), \tag{32}$$

where $\alpha_g \sim N[0, \sigma_\alpha^2]$. If α_g is known, the N_g observations in cluster g are independent with joint density

$$f(y_{1g}, \ldots, y_{N_g g}|\alpha_g, \boldsymbol{X}_g) = \prod_{i=1}^{N_g} \Lambda(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta})^{y_{ig}}[1 - \Lambda(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta})]^{1-y_{ig}}.$$

Since α_g is unknown it is integrated out, leading to joint density

$$f(y_{1g}, \ldots, y_{N_g g}|\boldsymbol{X}_g) = \int \left(\prod_{i=1}^{N_g} \Lambda(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta})^{y_{ig}}[1 - \Lambda(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta})]^{1-y_{ig}} \right) h(\alpha_g|\sigma_\alpha^2) d\alpha_g,$$

where $h(\alpha_g|\sigma_\alpha^2)$ is the $N[0, \sigma_\alpha^2]$ density. There is no closed form solution for this integral, but it is only one-dimensional so numerical approximation (such as Gaussian quadrature) can be used. The consequent MLE can be obtained in Stata using the command `xtlogit y x, re`. Note that in this RE logit model (31) no longer holds, so $\boldsymbol{\beta}$ in the model (32) is scaled differently from $\boldsymbol{\beta}$ in (31). Furthermore $\boldsymbol{\beta}$ in (32) is inconsistent if the distribution for α_g is misspecified, so there is little point in using option `vce(robust)` after command `xtlogit, re`.

It is important to realize that in nonlinear models such as logit, the population-averaged and random effects approaches lead to quite different estimates of $\boldsymbol{\beta}$ that are not comparable since $\boldsymbol{\beta}$ means different things in the different models. The resulting estimated average marginal effects may be similar, however, just as they are in standard cross-section logit and

probit models.

With few clusters, Wald statistics are likely to over-reject as in the linear case, even if we scale the CRVE's given in this section by $G/(G-1)$ as is typically done; see (12) for the linear case. McCaffrey, Bell, and Botts (2001) consider bias-correction of the CRVE in generalized linear models. Asymptotic refinement using a pairs cluster bootstrap as in Subsection VI.C is possible. The wild bootstrap given in Subsection VI.D is no longer possible in a nonlinear model, aside from nonlinear least squares, since it requires additively separable errors. Instead one can use the score wild bootstrap proposed by Klein and Santos (2012) for nonlinear models, including maximum likelihood and GMM models. The idea in their paper is to estimate the model once, generate scores for all observations, and then perform a bootstrap in the wild-cluster style, perturbing the scores by bootstrap weights at each step. For each bootstrap replication the perturbed scores are used to build a test statistic, and the resulting distribution of this test statistic can be used for inference. They find that this method performs well in small samples, and can greatly ease computational burden because the nonlinear model need only be estimated once. The conservative test of Ibragimov and Müller (2010) can be used if $N_g \to \infty$.

2. Fixed Effects

A cluster-specific fixed effects version of the logit model treats the unobserved parameter α_g in (32) as being correlated with the regressors $\boldsymbol{x}_{ig}$. In that case both the population-averaged and random effects logit estimators are inconsistent for $\boldsymbol{\beta}$.

Instead we need a fixed effects logit estimator. In general there is an incidental parameters problem if asymptotics are that N_g is fixed while $G \to \infty$, as there only N_g observations for each α_g, and inconsistent estimation of α_g spills over to inconsistent estimation of $\boldsymbol{\beta}$. Remarkably for the logit model it is nonetheless possible to consistently estimate $\boldsymbol{\beta}$. The logit fixed effects estimator is obtained in Stata using the command `xtlogit y x, fe`. Note, however, that the marginal effect in model (32) is $\partial \Pr[y_{ig} = 1|\alpha_g, \boldsymbol{x}_{ig}]/\partial x_{ijk} = \Lambda(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta})(1 - \Lambda(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta}))\beta_k$. Unlike the linear FE model this depends on the unknown α_g. So the marginal effects cannot be computed, though the ratio of the marginal effects of the k^{th} and l^{th} regressor equals β_k/β_l which can be consistently estimated.

The logit model is one of few nonlinear models for which fixed effects estimation is possible when N_g is small. The other models are Poisson with $\mathrm{E}[y_{ig}|\boldsymbol{X}_g, \alpha_g] = \exp(\alpha_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta})$, and nonlinear models with $\mathrm{E}[y_{ig}|\boldsymbol{X}_g, \alpha_g] = \alpha_g + m(\boldsymbol{x}_{ig}'\boldsymbol{\beta})$, where $m(\cdot)$ is a specified function.

The natural approach to introduce cluster-specific effects in a nonlinear model is to include a full set of cluster dummies as additional regressors. This leads to inconsistent estimation of $\boldsymbol{\beta}$ in all models except the linear model (estimated by OLS) and the Poisson regression model, unless $N_g \to \infty$. There is a growing literature on bias-corrected estimation in such cases; see, for example, Fernández-Val (2009). This paper also cites several simulation studies that gauge the extent of bias of dummy variable estimators for moderate N_g, such as $N_g = 20$.

Yoon and Galvao (2013) consider fixed effects in panel quantile regression models with correlation within cluster and provide methods under the assumption that both the number of individuals and the number of time periods go to infinity.

D. Cluster-randomized Experiments

Increasingly researchers are gathering their own data, often in the form of field or laboratory experiments. When analyzing data from these experiments they will want to account for the clustered nature of the data. And so when designing these experiments, they should also account for clustering. Fitzsimons, Malde, Mesnard, and Vera-Hernández (2012) use a wild cluster bootstrap in an experiment with 12 treated and 12 control clusters.

Traditional guidance for computing power analyses and minimum detectable effects (see e.g. Duflo, Glennerster and Kremer, 2007, pp. 3918-3922, and Hemming and Marsh (2013)) are based on assumptions of either independent errors or, in a clustered setting, a random effects common-shock model. Ideally one would account for more general forms of clustering in these calculations (the types of clustering that motivate cluster-robust variance estimation), but this can be difficult to do ex ante. If you have a data set that is similar to the one you will be analyzing later, then you can assign a placebo treatment, and compute the ratio of cluster-robust standard errors to default standard errors. This can provide a sense of how to adjust the traditional measures used in design of experiments.

VIII. Empirical Example

In this section we illustrate the most common applications of cluster-robust inference. There are two examples. The first is a Moulton-type setting that uses individual-level cross section data with clustering on state. The second is the Bertrand et al. (2004) example of DiD in a state-year panel with clustering on state and potentially with state fixed effects.

The micro data are from the March CPS, downloaded from IPUMS-CPS (King et al., 2010). We use data covering individuals who worked 40 or more weeks during the prior year, and whose usual hours per week in that year was 30 or more. The hourly wage is constructed as annual earnings divided by annual hours (usual hours per week times number of weeks worked), deflated to real 1999 dollars, and observations with real wage in the range ($2, $100) are kept.

The cross-section example uses individual-level data for 2012. The panel example uses data aggregated to the state-year level for 1977 to 2012. In both cases we estimate log-wage regressions and perform inference on a generated regressor that has zero coefficient. Specifically, we test $H_0: \beta = 0$ using $w = \hat{\beta}/s_{\hat{\beta}}$. For each example we present results for a single data set, before presenting a Monte Carlo experiment that focuses on inference when there are few clusters.

We contrast various ways to compute standard errors and perform Wald tests. Even when using a single statistical package, different ways to estimate the same model may lead to different empirical results due to calculation of different degrees of freedom, especially in models with fixed effects, and due to uses of different distributions in computing p-values and critical values. To make this dependence clear we provide the particular Stata command used to obtain the results given below; similar issues arise if alternative statistical packages are employed.

The data and accompanying Stata code (version 13) are available at our websites.

A. Individual-level Cross-section Data: One Sample

In our first application we use data on 65,685 individuals from the year 2012. The model is

$$y_{ig} = \beta d_g + \boldsymbol{z}_{ig}'\boldsymbol{\gamma} + u_{ig}, \tag{33}$$

where y_{ig} is log-wage, d_g is a randomly generated dummy "policy" variable, equal to one for one-half of the states and zero for the other half, and $\boldsymbol{z}_{ig}$ is a set of individual-level controls (age, age squared, and education in years). Estimation is by OLS and by FGLS controlling for state-level random effects.

The policy variable d_g is often referred to as a "placebo" treatment, and should be statistically significant in only 5% of tests performed at significance level 0.05.

Table 1 reports the estimated coefficient of the policy variable, along with its standard error computed in several different ways, when there are 51 clusters (states).

OLS results given in the first column of Table 1 are obtained using Stata command `regress`. The default standard error is misleadingly small ($se = 0.0042$), leading to the dummy variable being very highly statistically significant ($t = -0.0226/0.0042 = -5.42$) even though it was randomly generated independently of log-wage. The White heteroskedastic-robust standard error, from `regress` option `vce(robust)`, is similar in magnitude. From Subsection IV.A White standard errors should not be used if N_g is small, but here N_g is large. The cluster-robust standard error ($se = 0.0229$) using option `vce(cluster state)` is 5.5 times larger and leads to the more sensible result that the regressor is statistically insignificant ($t = -0.99$). In results not presented in Table 1, the cluster-robust standard errors of the other regressors - age, age squared and education - were, respectively, 1.2, 1.2 and 2.3 times the default. So ignoring clustering again understates the standard errors. As expected, a pairs cluster bootstrap (without asymptotic refinement) using option `vce(boot, cluster(state))`, yields very similar cluster-robust standard error.

Note that formula (6) suggests that the cluster-robust standard errors are 4.9 times the default ($\sqrt{1 + (1 \times 0.018 \times (65685/51 - 1)} = 4.9$), close to the observed multiple of 5.5. Formula (6) may work especially well in this example as taking the natural logarithm of wage leads to model error that is close to homoskedastic and equicorrelation is a good error model for individuals clustered in regions.

FGLS estimates for a random effects model with error process defined in (16) are given in the second column of Table 1. These were obtained using command `xtreg, re` after `xtset state`. The cluster-robust standard error defined in (15), and computed using option `vce(robust)`, is 0.0214/0.0199 = 1.08 times larger than the default. The pairs cluster bootstrap, implemented using option `vce(boot)` yields a similar cluster-robust standard error.

In principle FGLS can be more efficient than OLS. In this example, there is a modest gain in efficiency with the cluster-robust standard error equal to 0.0214 for FGLS compared to 0.0229 for OLS.

Finally, to illustrate the potential pitfalls of pairs cluster bootstrapping for standard errors when there are few clusters, discussed in Subsection VI.C, we examine a modification with only six states broken into treated (AZ, LA, MD) and control (DE, PA, UT). For these six states, we estimate a model similar to that in Table 1. Then $\hat{\beta} = 0.0373$ with default $se = 0.0128$. We then perform a pairs cluster bootstrap with 999 replications. The bootstrap $se = 0.0622$ is similar to the cluster-robust $se = 0.0577$. However, several problems arise. First, 28 replications cannot be estimated, presumably due to no variation in treatment in the bootstrap samples. Second, a kernel density estimate of the bootstrapped $\hat{\beta}s$ reveals that their distribution is very multi-modal and has limited density near the middle of the distribution. Considering these results, we would not feel comfortable using the pairs cluster bootstrap in this dataset with these few clusters. Better is to base inference on a wild cluster bootstrap.

This example highlights the need to use cluster-robust standard errors even when model errors are only weakly correlated within cluster (the intraclass correlation of the residuals in

this application is 0.018), if the regressor is substantially correlated within cluster (here perfectly correlated within cluster) and/or cluster sizes are large (ranging here from 519 to 5866).

B. Individual-level Cross-section Data: Monte Carlo

We next perform a Monte Carlo exercise to investigate the performance of various cluster-robust methods as the number of clusters becomes small. The analysis is based on the same cross-section regression as in the previous subsection.

In each replication, we generate a dataset by sampling (with replacement) states and all their associated observations. For quicker computation of the Monte Carlo simulation, within each state we use only a 20% subsample of individuals, so there are on average approximately 260 observations per cluster.

We explore the effect of the number of clusters G by performing varying simulations with G equal to 6, 10, 20 or 50. Given a sample of states, we assign a dummy "policy" variable to one-half of the states. We run OLS regressions of log-wage on the policy variable and the same controls as used for the Table 1 regressions.

In these simulations we perform tests of the null hypothesis that the slope coefficient of the policy variable is zero. Table 2 presents rejection rates that with millions of replications should equal 0.05, since we are testing a true hypothesis at a nominal 5% level. For $G = 6$ and 10 we perform 4,000 simulations, so we expect that 95% of these simulations will yield estimated test size in the range $(0.043,\ 0.057)$ if the true test size is 0.05. For larger G there are 1,000 simulations and the 95% simulation interval is instead $(0.036,\ 0.064)$.

We begin with lengthy discussion of the many clusters case. These results are given in the final column ($G = 50$) of Table 2. Rows 1-9 report sizes for Wald tests based on $t = \hat{\beta}/se$ where se is computed in various ways, while rows 10-15 report sizes for tests using various bootstraps with an asymptotic refinement. Basic Stata commands yield the standard errors in rows 1-4 and 9, while the remaining rows require additional coding.

Row 1 presents the size of tests using heteroskedastic-robust standard errors, obtained using Stata command using `regress, vce(robust)`. Ignoring clustering leads to great over-rejection due to considerable under-estimation of the standard error. Using formula (6) for this 20% subsample yields a standard error inflation factor of $\sqrt{1 + (1 \times 0.018 \times (0.20 \times 65685/51 - 1)} = 2.38$. So $t = 1.96$ using the heteroskedastic-robust standard error is really $t = 1.96/2.38 = 0.82$. And, using standard normal critical values, an apparent $p = 0.05$ is really $p = 0.41$ since Pr[|z| > 0.82] = 0.41. This crude approximation is fairly close to $p = 0.498$ obtained in this simulation.

Results using cluster-robust standard errors, presented in rows 2-4 and obtained from `regress, vce(cluster state)`, show that even with 50 clusters the choice of distribution to use in obtaining critical value makes a difference. The rejection rate is closer to 0.05 when $T(G-1)$ critical values are used than when $N[0,1]$ critical values are used. Using $T(G-2)$ in row 4, suggested by the study of Donald and Lang (2007), leads to slight further improvement, but there is still over-rejection.

Results using the bias adjustments CR2 and CR3 discussed in Subsection VI.B, along with $T(G-1)$ critical values, are presented in rows 5-6. Bias adjustment leads to further decrease in the rejection rates, towards the desired 0.05.

Rows 7 and 8 use critical values from the T distribution with the data-determined degrees-of-freedom of Subsection VI.D, equal to 17 on average when $G = 50$ (see rows 14 and 17). This leads to further improvement in the Monte Carlo rejection rate.

Bootstrap standard errors obtained from a standard pairs cluster bootstrap, implemented using command `regress, vce(boot, cluster(state))` are used in

row 9. For $G = 50$ the rejection rate is essentially the same as that in row 3, as expected since this bootstrap has no asymptotic refinement.

Rows 10-15 implement the various percentile-t bootstraps with asymptotic refinement presented in Subsection VI.C. Only 399 bootstraps are used here as any consequent bootstrap simulation error averages out over the many Monte Carlo replications. But if these bootstraps were used just once, as in an empirical study, a percentile-t bootstrap should use at least 999 replications. Row 10 can be computed using the `bootstrap:` command, see our posted code, while rows 11-15 require additional coding. For $G = 50$ the various bootstraps give similar results, with rejection rates of a bit more than 0.06.

Rows 16-19 give the effective number of clusters. The Imbens and Kolesar (2013) measure v^* in (26), denoted IK, and the Carter, Schnepel and Steigerwald (2013) measure G^* in (27), denoted CSS, are both equal to 17 on average when $G = 50$. For the IK measure, across the 1,000 simulations, the 5th percentile is 9.6 and the 95th percentile is 29.5.

We next examine settings with fewer clusters than $G = 50$. Then most methods lead to rejection rates even further away from the nominal test size of 0.05.

Consider the case $G = 6$. Rows 2-4 and 8 compute the same Wald test statistic but use different degrees of freedom in computing p-values. This makes an enormous difference when G is small, as the critical value for a Wald test at level 0.05 rises from 2.571 to 2.776 and 3.182 for, respectively, the $T(5)$, $T(4)$ and $T(3)$ distributions, and from row 16 the IK degrees of freedom averages 3.3 across the simulations. The CSS degrees of freedom is larger than the IK as, from Subsection VI.D, it involves an approximation that only disappears as G becomes large.

Using a bias-corrected CRVE also makes a big difference. It appears from rows 6 and 7 that it is best to use the CR3 bias-correction with $T(G-1)$ critical values, and the CR2 bias-correction with $T(v^*)$ critical values where v^* is the Imbens and Kolesar (2013) calculated degrees of freedom.

A downside to using cluster-robust standard errors is that they provide an estimate of the standard deviation of $\hat{\beta}$ that is more variable than the default or heteroskedastic-robust standard errors. This introduces a potential bias – variance tradeoff. To see whether this increased variability is an issue we performed 1000 Monte Carlo replications using the full cross-section micro dataset, resampling the 50 states with replacement. The standard deviation of the cluster-robust standard error across the 1,000 replications equaled 12.3% of the mean cluster-robust standard error, while the standard deviation of the heteroskedastic-robust standard error equaled 4.5% of its mean. So while the CRVE is less biased than heteroskedastic-robust (or default), it is also more variable. But the increased variability is relatively small, especially compared to the very large bias that can arise if clustering is not controlled for.

Rows 10-15 present various bootstraps with asymptotic refinement. From row 10, the pairs cluster bootstrap performs extremely poorly for $G \leq 10$.

Results for the wild cluster bootstrap using a Rademacher 2 point distribution are presented in rows 11-13. From Subsection VI.B this bootstrap yields only $2^6 = 64$ possible datasets when $G = 6$, and hence at most 64 unique values for w^*. This leads to indeterminacy for the test p-value. Suppose the p-value is in the range [a,b]. Then H_0 is rejected in row 11 if a < 0.05, in row 12 if (a+b)/2 < 0.05, and in row 13 if b < 0.05. The indeterminacy leads to substantially different results for G as low as six, though not for $G \geq 10$.

The wild cluster bootstrap using the Webb 6 point distribution, see row 14, does not have this complication when $G = 6$. And it yields essentially the same results as those using the Rademacher 2 point distribution when $G \geq 10$.

Comparing row 15 to row 12, imposing the null hypothesis in performing the wild

bootstrap does not change the rejection rate very much in this set of simulations when $G \geq 10$, although it appears to matter when $G = 6$. By comparison we have found a more substantial difference when simulating from the d.g.p. of Cameron et al. (2008).

In summary, the various wild cluster bootstraps lead to test size that is closer to 0.05 than using a standard Wald test with cluster-robust standard errors and $T(G-1)$ critical values. But the test size still exceeds 0.05 and the bias adjustments in rows 6 and 7 appear to do better.

This example illustrates that at a minimum one should use a cluster-robust Wald test with $T(G-1)$ critical values. Especially when there are few clusters it is better still to use bias-adjusted cluster-robust standard errors or to use a wild cluster bootstrap. In this Monte Carlo experiment, with few clusters the test size was closest to the nominal size using a Wald test with cluster-robust standard errors computed using the CR2 correction and $T(v^*)$ critical values, or using the larger CR2 correction with $T(G-1)$ critical values.

C. State–Year Panel Data: One Sample

We next turn to a panel difference-in-difference application motivated by Bertrand et al. (2004). The underlying data are again individual-level data from the CPS, but now obtained for each of the years 1977 to 2012.

In applications where the policy regressor of interest is only observed at the state-year level, it is common to first aggregate the individual-level data to the state-year level before OLS regression. Several methods are used; we use the following method.

The model estimated for 51 states from 1977 to 2012 is

$$\tilde{y}_{ts} = \alpha_s + \delta_t + \beta \times d_{ts} + u_{ts}, \tag{34}$$

where $\tilde{y}_{ts}$ is the average log-wage in year t and state s (after partialling out individual level covariates), α_s and δ_t are state and year dummies, and d_{ts} is a random "policy" variable that turns on and stays on for the last 18 years for one half of the states. Here $G = 51$, $T = 36$ and $N = 1836$.

The individual level covariates (age, age squared, and years of education) are partialled out using a two-step estimation procedure presented in Hansen (2007b). Define D_{ts} to be state-by-year dummies. First we OLS regress log wage (y_{its}) on state-by-year dummies D_{ts} and on the individual level covariates. And second $\tilde{y}_{ts}$ in equation (34) equals the estimated coefficients of the D_{ts} dummies.

To speed up bootstraps, and to facilitate computation of the CR2 residual adjustment, we additionally partial out the state fixed effects and year fixed effects in (34) by the standard Frisch-Waugh method. We separately regress $\tilde{y}_{ts}$ and d_{ts} on the state dummies and the year dummies. Then β is estimated by regressing the residuals of $\tilde{y}_{ts}$ on the residuals of d_{ts}, with no constant. As noted below, regression using residuals leads to slightly different standard errors due to different degrees of freedom used in calculating the CRVE.

Table 3 presents results for the policy dummy regressor which should have coefficient zero since it is randomly assigned.

We begin with model 1, OLS controlling for state and year fixed effects. Using default or White-robust standard errors (rows 1-2) leads to a standard error of 0.0037 that is much smaller than the cluster-robust standard error of 0.0119 (row 3), where clustering is on state. Similar standard errors are obtained using the CR2 correction (rows 4 and 5) and bootstrap without asymptotic refinement (row 6). Note that from rows 10 and 11 the IK and CSS degrees of freedom are calculated to be, respectively, $G-1$ and G, an artifact of having balanced cluster sizes and a single regressor that is symmetric across clusters.

The inclusion of state and year fixed effects complicates computation of the degrees of freedom (*df*) adjustment used in computing the CRVE. The row 3 column 1 results are obtained from regression of residual on residual without intercept, using command `regress, noconstant vce(cluster(state))`. Then from formula (12) $df = \frac{G}{G-1} \times \frac{GT}{GT-1}$. If instead we directly regressed log-wage on the state and year fixed effects and the *K* regressors, again using `regress, vce(cluster(state))`, then $df = \frac{G}{G-1} \times \frac{GT}{GT-T-G-1}$ and the cluster-robust standard error equals 0.0122 rather than 0.0119 . And if instead we directly estimate the log-wage equation using `xtreg, vce(robust)` after `xtset state` then $df = \frac{G}{G-1}$ and the cluster-robust standard error equals 0.0120. In this example with large *T* and *G* these adjustments make little difference, but for small *G* or *T* one should use $df = \frac{G}{G-1}$ as explained in Subsection III.B.

The corresponding p-values for tests of the null hypothesis that $\beta = 0,$ following OLS regression, are given in column 4 of Table 3. Default and heteroskedastic-robust standard errors lead to erroneously large t-statistics (of 0.0156 / 0.0037 = 4.22), so $p = 0.000$ and the null hypothesis is incorrectly rejected. Using various standard errors that control for clustering (rows 3-6) leads to $p \simeq 0.20$ so that the null is not rejected. Rows 7-9 report p-values from several percentile-t bootstraps that again lead to rejection of $H_0: \beta = 0$.

For illustrative purposes we also compute standard errors allowing for two-way clustering, see Section V, with clustering on both state and year. These are computed using the user-generated Stata add-on program cgmreg.ado. Clustering on year is necessary if both the regressor and the model errors are correlated across states in a given year. For this application, the result (s.e. = 0.01167) is very similar to that from clustering on state alone (s.e. = 0.01185). In some other panel applications the two-way cluster robust standard errors can be substantially larger than those from clustering on state alone.

The main lesson from the model 1 OLS results is that even after inclusion of state fixed effects one needs to use cluster-robust standard errors that cluster on state. The inclusion of state fixed effects did not eliminate the within-state correlation of the error. In this example the correct cluster-robust standard errors are 3.6 times larger than the default.

Model 2 again uses OLS estimation, but drops the state fixed effects from the model (34). Dropping these fixed effects leads to much less precise estimation as the cluster-robust standard error (row 3) increases from 0.0119 to 0.0226 and this cluster-robust standard error is now 0.0226 / 0.0062 = 3.6 times the default, comparable to a ratio of 3.6 when state fixed effects were included. Note that inclusion of state fixed effects (model 1) did soak up some of the within-state error correlation, as expected, but there still remained substantial within-cluster correlation of the error so that cluster-robust standard errors need to be used.

For model 2 the comparisons of the various standard errors and p-values are qualitatively similar to those for model 1, so are not discussed further.

Model 3 estimates the same model as model 1, except that the state and year fixed effects are directly estimated, and estimation is now by FGLS allowing for an AR(1) process for the errors. Since there are 36 years of data the bias correction of Hansen (2007b), see Subsection III.C, will make little difference and is not used here. Estimation uses Stata command `xtreg, pa corr(ar 1)` after `xtset state`.

Comparing rows 1 and 3, again even with inclusion of state fixed effects one should obtain standard errors that cluster on state, using `xtreg, pa` option `vce(robust))`. The difference is not as pronounced as for OLS, with FGLS cluster-robust standard error that is 0.0084/0.0062 = 1.4 times the default.

FGLS estimation has led to substantial gain in efficiency, with cluster-robust standard error (row 3) for FGLS of 0.0084 compared to 0.0119 for OLS.

This example illustrates that even with state fixed effects included in a state year panel inference should be based on cluster-robust standard errors. Furthermore there can be substantial efficiency gains to estimating by FGLS rather than OLS.

D. State–Year Panel Data: Monte Carlo

We next perform a Monte Carlo exercise to investigate the performance of various cluster-robust methods as the number of clusters becomes small. The analysis is based on the same state-year panel regression as in the previous subsection, with each state-year observation based on log-wages after partialling out the individual level covariates.

In each simulation, we draw a random set of G states (with replacement), where G takes values 6, 10, 20, 20 and 50. When a state is drawn, we take all years of data for that state. We then assign our DiD "policy" variable to half the states, with the policy turned on mid-way through the time period. In these simulations we perform tests of the null hypothesis that the slope coefficient of the policy variable is zero. As in Table 2, for $G = 6 \text{ and } 10$ we perform 4,000 simulations, so we expect that 95% of these simulations will yield estimated test size in the range $(0.043,\ 0.057)$. For larger G there are 1,000 simulations and the 95% simulation interval is instead $(0.036,\ 0.064)$.

We begin with the last column of Table 4, with $G = 50$ states. All tests aside from that based on default standard errors (row 1) have rejection rates that are not appreciably different from 0.05, once we allow for simulation error.

As the number of clusters decreases it becomes clear that one should use the $T(G-1)$ or $T(G-2)$ distribution for critical values, and even this leads to mild over-rejection with low G. The pairs cluster percentile-t bootstrap fails with few clusters, with rejection rate of only 0.005 when $G = 6$. For low G, the wild cluster percentile-t bootstrap has similar results with either 2-point or 6-point weights, with very slight over-rejection.

XI. Concluding Thoughts

It is important to aim for correct statistical inference, many empirical applications feature the potential for errors to be correlated within clusters, and we need to make sure our inference accounts for this. Often this is straightforward to do using traditional cluster-robust variance estimators - but sometimes things can be tricky. The leading difficulties are (1) determining how to define the clusters, and (2) dealing with few clusters; but other complications can arise as well. When faced with these difficulties, there is no simple hard and fast rule regarding how to proceed. You need to think carefully about the potential for correlations in your model errors, and how that interacts with correlations in your covariates. In this essay we have aimed to present the current leading set of tools available to practitioners to deal with clustering issues.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association* 105(490): 493-505.

Acemoglu, Daron, and Jörn-Steffen Pischke. 2003. "Minimum Wages and On-the-job Training." *Research in Labor Economics* 22: 159-202.

Andrews, Donald W. K. and James H. Stock. 2007. "Inference with Weak Instruments." In *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Vol. III, ed. Richard Blundell, Whitney K. Newey, and T. Persson, 122-173, Cambridge: Cambridge University Press.

Angrist, Joshua D., and Victor Lavy. 2002. "The Effect of High School Matriculation Awards: Evidence from Randomized Trials." *American Economic Review* 99 : 1384-1414.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arellano, Manuel 1987. "Computing Robust Standard Errors for Within-Group Estimators." *Oxford Bulletin of Economics and Statistic* 49(4): 431-434.

Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, and Michal Kolesár. 2012. "Clustering, Spatial Correlations and Randomization Inference." *Journal of the American Statistical Association* 107(498): 578–591.

Baum, Christopher F., Mark E. Schaffer, Steven Stillman. 2007. "Enhanced Routines for Instrumental Variables/GMM Estimation and Testing." *The Stata Journal* 7(4): 465-506.

Bell, Robert M., and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28(2): 169-179.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?." *Quarterly Journal of Economics* 119(1): 249-275.

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen. 2011. "Inference with Dependent Data Using Cluster Covariance Estimators." *Journal of Econometrics* 165(2): 137-151.

Bhattacharya, Debopam. 2005. "Asymptotic Inference from Multi-stage Samples." *Journal of Econometrics* 126: 145-171.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430): 443-450.

Brewer, Mike, Thomas F. Crossley, and Robert Joyce. 2013. "Inference with Differences-in-Differences Revisited." Unpublished.

Cameron, A. Colin, Jonah G. Gelbach, and Douglas L. Miller. 2006. "Robust Inference with Multi-Way Clustering." NBER Technical Working Paper 0327.

———. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics.* 90(3): 414-427.

———. 2011. "Robust Inference with Multi-Way Clustering." *Journal Business and Economic Statistics* 29(2): 238-249.

Cameron, A. Colin, and Douglas L. Miller. 2011. "Robust Inference with Clustered Data." In *Handbook of Empirical Economics and Finance.* ed. Aman Ullah and David E. Giles, 1-28. Boca Raton: CRC Press.

———. 2012. "Robust Inference with Dyadic Data: with Applications to Country-pair International Trade." University of California - Davis. Unpublished.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

———. 2009. *Microeconometrics using Stata*. College Station, TX: Stata Press.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2013. "Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity." University of California - Santa Barbara. Unpublished.

Cheng, Cheng, and Mark Hoekstra. 2013. "Pitfalls in Weighted Least Squares Estimation: A Practitioner's Guide." Texas A&M University. Unpublished.

Chernozhukov, Victor, and Christian Hansen. 2008. "The Reduced Form: A Simple Approach to Inference with Weak Instruments." *Economics Letters* 100(1): 68-71.

Conley, Timothy G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92(1): 1-45.

Conley, Timothy G., and Christopher R. Taber. 2011. "Inference with 'Difference in Differences' with a Small Number of Policy Changes." *Review of Economics and Statistics* 93(1): 113-125.

Davidson, Russell, and Emmanuel Flachaire. 2008. "The Wild Bootstrap, Tamed at Last." *Journal of Econometrics* 146(1): 162–169.

Davis, Peter. 2002. "Estimating Multi-Way Error Components Models with Unbalanced Data Structures." *Journal of Econometrics* 106(1): 67-95.

Donald, Stephen G., and Kevin Lang. 2007. "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics* 89(2): 221-233.

Driscoll, John C., and Aart C. Kraay. 1998. "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data." *Review of Economics and Statistics* 80(4): 549-560.

Duflo, Esther, Rachel Glennerster and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, Vol. 4, ed. Dani Rodrik and Mark Rosenzweig, 3895-3962. Amsterdam: North-Holland.

Fafchamps, Marcel, and Flore Gubert. 2007. "The Formation of Risk Sharing Networks." *Journal of Development Economics* 83(2): 326-350.

Fernández-Val, Iván. 2009. "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models." *Journal of Econometrics* 150(1): 70-85.

Finlay, Keith, and Leandro M. Magnusson. 2009. "Implementing Weak Instrument Robust Tests for a General Class of Instrumental-Variables Models." *Stata Journal* 9(3): 398-421.

Fitzsimons, Emla, Bansi Malde, Alice Mesnard, and Marcos Vera-Hernández. 2012. "Household Responses to Information on Child Nutrition: Experimental Evidence from Malawi." IFS Working Paper W12/07.

Foote, Christopher L. 2007. "Space and Time in Macroeconomic Panel Data: Young Workers and State-Level Unemployment Revisited." Working Paper 07-10, Federal Reserve Bank of Boston.

Greenwald, Bruce C. 1983. "A General Analysis of Bias in the Estimated Standard Errors of Least Squares Coefficients." *Journal of Econometrics* 22(3): 323-338.

Hansen, Christian. 2007a. "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large." *Journal of Econometrics* 141(2): 597-620.

Hansen, Christian. 2007b. "Generalized Least Squares Inference in Panel and Multi-level Models with Serial Correlation and Fixed Effects." *Journal of Econometrics* 141(2): 597-620.

Hausman, Jerry, and Guido Kuersteiner. 2008. "Difference in Difference Meets Generalized Least Squares: Higher Order Properties of Hypotheses Tests." *Journal of Econometrics* 144(2): 371-391.

Hemming, Karla, and Jen Marsh. 2013. "A Menu-driven Facility for Sample-size Calculations in Cluster Randomized Controlled Trails." *Stata Journal* 13(1): 114-135.

Hersch, Joni. 1998. “Compensating Wage Differentials for Gender-Specific Job Injury Rates.” *American Economic Review* 88(3): 598-607.
Hoechle, Daniel. 2007. “Robust Standard Errors for Panel Regressions with Cross–sectional Dependence.” *Stata Journal* 7(3): 281-312.
Hoxby, Caroline, and M. Daniele Paserman. 1998. “Overidentification Tests with Group Data.” NBER Technical Working Paper 0223.
Ibragimov, Rustam, and Ulrich K. Müller. 2010. “T-Statistic Based Correlation and Heterogeneity Robust Inference.” *Journal of Business and Economic Statistics* 28(4): 453-468.
Imbens, Guido W., and Michal Kolesar. 2012. “Robust Standard Errors in Small Samples: Some Practical Advice.” NBER Working Paper 18478.
Inoue, Atsushi, and Gary Solon. 2006. “A Portmanteau Test for Serially Correlated Errors in Fixed Effects Models.” *Econometric Theory* 22(5): 835-851.
Kézdi, Gábor. 2004. “Robust Standard Error Estimation in Fixed-Effects Panel Models.” *Hungarian Statistical Review* Special Number 9: 95-116.
King, Miriam, Steven Ruggles, J. Trent Alexander, Sarah Flood, Katie Genadek, Matthew B. Schroeder, Brandon Trampe, and Rebecca Vick. 2010. *Integrated Public Use Microdata Series, Current Population Survey: Version 3.0.* [Machine-readable database]. Minneapolis: University of Minnesota.
Kish, Leslie. 1965. *Survey Sampling.* New York: John Wiley.
Kish, Leslie, and Martin R. Frankel. 1974. “Inference from Complex Surveys with Discussion.” *Journal Royal Statistical Society B* 36(1): 1-37.
Kleibergen, F. and R. Paap. 2006. “Generalized Reduced Rank Tests iusing the Singular Value Decomposition.” *Journal of Econometrics* 133(1): 97-128.
Klein, Patrick, and Andres Santos. 2012. “A Score Based Approach to Wild Bootstrap Inference.” *Journal of Econometric Methods*:1(1): 23-41.
Kloek, T. 1981. “OLS Estimation in a Model where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated.” *Econometrica* 49(1): 205-07.
Liang, Kung-Yee, and Scott L. Zeger. 1986. “Longitudinal Data Analysis Using Generalized Linear Models.” *Biometrika* 73(1): 13-22.
MacKinnon, James. G., and Halbert White. 1985. “Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties.” *Journal of Econometrics* 29(3): 305-325.
MacKinnon, James, and Matthew D. Webb. 2013. “Wild Bootstrap Inference for Wildly Different Cluster Sizes.” Queens Economics Department Working Paper No. 1314.
McCaffrey, Daniel F., Bell, Robert M., and Carsten H. Botts. 2001. “Generalizations of Bias Reduced Linearization.” *Proceedings of the Survey Research Methods Section*, American Statistical Association.
Miglioretti, D. L., and P. J. Heagerty. 2006. “Marginal Modeling of Nonnested Multilevel Data using Standard Software.” *American Journal of Epidemiology* 165(4): 453-463.
Moulton, Brent R. 1986. “Random Group Effects and the Precision of Regression Estimates.” *Journal of Econometrics* 32: 385-397.
Moulton, Brent R. 1990. “An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units.” *Review of Economics and Statistics* 72(3): 334-38.
Newey, Whitney K., and Kenneth D. West. 1987. “A Simple, Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica* 55(3): 703-708.
Petersen, Mitchell A. 2009. “Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches.” *Review of Financial Studies* 22(1): 435-480.

Pfeffermann, Daniel, and Gaf Nathan. 1981. “Regression Analysis of Data from a Cluster Sample.” *Journal American Statistical Association* 76(375): 681-689.
Rabe-Hesketh, Sophia, and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata, Volumes I and II*, Third Edition. College Station, TX: Stata Press.
Rogers, William H. 1993. “Regression Standard Errors in Clustered Samples.” *Stata Technical Bulletin* 13: 19-23.
Satterthwaite, F. E. 1946. “An Approximate Distribution of Estimates of Variance Components.” *Biometrics Bulletin* 2(6): 110-114.
Schaffer, Mark E., and Stillman, Steven. 2010. “xtoverid: Stata Module to Calculate Tests of Overidentifying Restrictions after xtreg, xtivreg, xtivreg2 and xthtaylor.” http://ideas.repec.org/c/boc/bocode/s456779.html
Scott, A. J., and D. Holt. 1982. “The Effect of Two-Stage Sampling on Ordinary Least Squares Methods.” *Journal American Statistical Association* 77(380): 848-854.
Shah, Bbabubhai V., M. M. Holt and Ralph E. Folsom. 1977. “Inference About Regression Models from Sample Survey Data.” *Bulletin of the International Statistical Institute Proceedings of the 41st Session* 47(3): 43-57.
Shore-Sheppard, L. 1996. “The Precision of Instrumental Variables Estimates with Grouped Data.” Princeton University Industrial Relations Section Working Paper 374.
Solon, Gary, Steven J. Haider, and Jeffrey Wooldridge. 2013. “What Are We Weighting For?” NBER Working Paper 18859.
Staiger, Douglas, and James H. Stock. 1997. “Instrumental Variables Regression with Weak Instruments.” *Econometrica* 65: 557-586.
Stock, James H., and Mark W. Watson. 2008. “Heteroskedasticity-robust Standard Errors for Fixed Effects Panel Data Regression.” *Econometrica* 76(1): 155-174.
Stock, James H. and M. Yogo. 2005. “Testing for Weak Instruments in Linear IV Regressions.” In I*dentification and Inference for Econometric Models*. Ed. Donald W. K. Andrews and James H. Stock, 80-108. Cambridge: Cambridge University Press.
Thompson, Samuel. 2006. “Simple Formulas for Standard Errors that Cluster by Both Firm and Time.” SSRN paper. http://ssrn.com/abstract=914002.
Thompson, Samuel. 2011. “Simple Formulas for Standard Errors that Cluster by Both Firm and Time.” *Journal of Financial Economics* 99(1): 1-10.
Webb, Matthew D. 2013. “Reworking Wild Bootstrap Based Inference for Clustered Errors.” Queens Economics Department Working Paper 1315.
White, Halbert. 1980. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica* 48(4): 817-838.
White, Halbert. 1984. *Asymptotic Theory for Econometricians.* San Diego: Academic Press.
Wooldridge, Jeffrey M. 2003. “Cluster-Sample Methods in Applied Econometrics.” *American Economic Review* 93(2): 133-138.
Wooldridge, Jeffrey M. 2006. “Cluster-Sample Methods in Applied Econometrics: An Extended Analysis.” Michigan State University. Unpublished.
Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
Yoon, Jungmo, and Antonio Galvao. 2013. “Robust Inference for Panel Quantile Regression Models with Individual Effects and Serial Correlation.” Unpublished.

Table 1 - Cross-section individual level data

Impacts of clustering and estimator choices on estimated coefficients and standard errors

	Estimation Method	
	OLS	FGLS (RE)
Slope coefficient	0.0108	0.0314
Standard Errors		
Default	0.0042	0.0199
Heteroscedastic Robust	0.0042	-
Cluster Robust (cluster on State)	0.0229	0.0214
Pairs cluster bootstrap	0.0224	0.0216
Number observations	65685	65685
Number clusters (states)	51	51
Cluster size range	519 to 5866	519 to 5866
Intraclass correlation	0.018	-

Notes: March 2012 CPS data, from IPUMS download. Default standard errors for OLS assume errors are iid; default standard errors for FGLS assume the Random Effects model is correctly specified. The Bootstrap uses 399 replications. A fixed effect model is not possible, since the regressor is invariant within states.

Table 2 - Cross-section individual level data

Monte Carlo rejection rates of true null hypothesis (slope = 0) with different number of clusters and different rejection methods

Nominal 5% rejection rates

	Wald test method	Numbers of Clusters				
		6	10	20	30	50
	Different standard errors and critical values					
1	White Robust, T(N-k) for critical value	0.439	0.457	0.471	0.462	0.498
2	Cluster on state, T(N-k) for critical value	0.215	0.147	0.104	0.083	0.078
3	Cluster on state, T(G-1) for critical value	0.125	0.103	0.082	0.069	0.075
4	Cluster on state, T(G-2) for critical value	0.105	0.099	0.076	0.069	0.075
5	Cluster on state, CR2 bias correction, T(G-1) for critical value	0.082	0.070	0.062	0.060	0.065
6	Cluster on state, CR3 bias correction, T(G-1) for critical value	0.048	0.050	0.050	0.052	0.061
7	Cluster on state, CR2 bias correction, T(IK DOF) for critical value	0.052	0.050	0.047	0.047	0.054
8	Cluster on state, T(CSS effective # clusters)	0.114	0.079	0.057	0.056	0.061
9	Pairs cluster bootstrap for standard error, T(G-1) for critical value	0.082	0.072	0.069	0.067	0.074
	Bootstrap Percentile-T methods					
10	Pairs cluster bootstrap	0.009	0.031	0.046	0.051	0.061
11	Wild cluster bootstrap, Rademacher 2 point distribution, low-p-value	0.097	0.065	0.062	0.051	0.060
12	Wild cluster bootstrap, Rademacher 2 point distribution, mid-p-value	0.068	0.065	0.062	0.051	0.060
13	Wild cluster bootstrap, Rademacher 2 point distribution, high-p-value	0.041	0.064	0.062	0.051	0.060
14	Wild cluster bootstrap, Webb 6 point distribution	0.079	0.067	0.061	0.051	0.061
15	Wild cluster bootstrap, Rademacher 2 pt, do not impose null hypothesis	0.086	0.063	0.050	0.053	0.056
16	IK effective DOF (mean)	3.3	5.6	9.4	12.3	16.9
17	IK effective DOF (5th percentile)	2.7	3.7	4.9	6.3	9.6
18	IK effective DOF (95th percentile)	3.8	7.2	14.5	20.8	29.5
19	CSS effective # clusters (mean)	4.7	6.6	9.9	12.7	17
20	Average number of observations	1554	2618	5210	7803	13055

Notes: March 2012 CPS data, 20% sample from IPUMS download. For 6 and 10 clusters, 4000 Monte Carlo replications. For 20-50 clusters, 1000 Monte Carlo replications. The Bootstraps use 399 replications. "IK effective DOF" from Imbens and Kolesar (2013), and "CSS effective # clusters" from Carter, Schnepel and Steigerwald (2013), see Subsection VI.D. Row 11 uses lowest p-value from interval, when Wild percentile-T bootstrapped p-values are not point identified due to few clusters. Row 12 uses mid-range of interval, and row 13 uses largest p-value of interval.

Table 3 - State-year panel data with differences-in-differences estimation
Impacts of clustering and estimation choices on estimated coefficients, standard errors, and p-values

		Standard Errors			p-values		
	Model:	1	2	3	1	2	3
	Estimation Method:	OLS-FE	OLS-no FE	FGLS AR(1)	OLS-FE	OLS-no FE	FGLS AR(1)
	Slope coefficient	0.0156	0.0040	-0.0042			
	Standard Errors						
1	Default standard errors, T(N-k) for critical value	0.0037	0.0062	0.0062	0.000	0.521	0.494
2	White Robust, T(N-k) for critical value	0.0037	0.0055	na	0.000	0.470	na
3	Cluster on state, T(G-1) for critical value	0.0119	0.0226	0.0084	0.195	0.861	0.617
4	Cluster on state, CR2 bias correction, T(G-1) for critical value	0.0118	0.0226	na	0.195	0.861	na
5	Cluster on state, CR2 bias correction, T(IK DOF) for critical value	0.0118	0.0226	na	0.195	0.861	na
6	Pairs cluster bootstrap for standard error, T(G-1) for critical value	0.0118	0.0221	0.0086	0.191	0.857	0.624
	Bootstrap Percentile-T methods						
7	Pairs cluster bootstrap	na	na		0.162	0.878	
8	Wild cluster bootstrap, Rademacher 2 point distribution	na	na		0.742	0.968	
9	Wild cluster bootstrap, Webb 6 point distribution	na	na		0.722	0.942	
10	Imbens-Kolesar effective DOF	50	50				
11	C-S-S effective # clusters	51	51				
	Number observations	1836	1836	1836			
	Number clusters (states)	51	51	51			

Notes: March 1997-2012 CPS data, from IPUMS download. Models 1 and 3 include state and year fixed effects, and a "fake policy" dummy variable that turns on in 1995 for a random subset of half of the states. Model 2 includes year fixed effects but not state fixed effects. The Bootstraps use 999 replications. Model 3 uses FGLS, assuming an AR(1) error within each state. "IK effective DOF" from Imbens and Kolesar (2013), and "CSS effective # clusters" from Carter, Schnepel and Steigerwald (2013), see Subsection VI.D.

Table 4 - State-year panel data with differences-in-differences estimation
Monte Carlo rejection rates of true null hypothesis (slope = 0) with different # clusters and different rejection methods
Nominal 5% rejection rates

	Estimation Method	Numbers of Clusters				
		6	10	20	30	50
	Wald Tests					
1	Default standard errors, T(N-k) for critical value	0.589	0.570	0.545	0.526	0.556
2	Cluster on state, T(N-k) for critical value	0.149	0.098	0.065	0.044	0.061
3	Cluster on state, T(G-1) for critical value	0.075	0.066	0.052	0.039	0.058
4	Cluster on state, T(G-2) for critical value	0.059	0.063	0.052	0.038	0.058
5	Pairs cluster bootstrap for standard error, T(G-1) for critical value	0.056	0.060	0.050	0.036	0.057
	Bootstrap Percentile-T methods					
6	Pairs cluster bootstrap	0.005	0.019	0.051	0.044	0.069
7	Wild cluster bootstrap, Rademacher 2 point distribution	0.050	0.059	0.050	0.036	0.055
8	Wild cluster bootstrap, Webb 6 point distribution	0.056	0.059	0.048	0.037	0.058

Notes: March 1997-2012 CPS data, from IPUMS download. Models include state and year fixed effects, and a "fake policy" dummy variable that turns on in 1995 for a random subset of half of the states. For 6 and 10 clusters, 4000 Monte Carlo replications. For 20-50 clusters, 1000 Monte Carlo replications. The Bootstraps use 399 replications.