Queen's Economics Department Working Paper No. 1421

When and How to Deal with Clustered Errors in Regression Models

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

Revised version, 5-2020

When and How to Deal With Clustered Errors in Regression Models*

James G. MacKinnon†
Queen's University
jgm@econ.queensu.ca

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

May 14, 2020

Abstract

We discuss when and how to deal with possibly clustered errors in linear regression models. Specifically, we discuss situations in which a regression model may plausibly be treated as having error terms that are arbitrarily correlated within known clusters but uncorrelated across them. The methods we discuss include various covariance matrix estimators, possibly combined with various methods of obtaining critical values, several bootstrap procedures, and randomization inference. Special attention is given to models with few treated clusters and clusters that vary a lot in size, where inference may be problematic. Two empirical examples illustrate the methods we discuss and the concerns we raise, and a simulation experiment illustrates the consequences of over-clustering and under-clustering.

*We thank the Social Sciences and Humanities Research Council of Canada (SSHRC) for financial support. We are grateful to Mehtab Hanzroh for his excellent research assistance. We benefited from the comments of Alfonso Flores-Lagunes, Andreas Hagemann, Azeem Shaikh, Holger Spamann, an anonymous referee, and participants at the CIREQ 2019 Bootstrap Conference and the 2019 Canadian Economics Association Annual Meeting.

†Corresponding author. Address: Department of Economics, 94 University Avenue, Queen's University, Kingston, Ontario K7L 3N6, Canada. Email: jgm@econ.queensu.ca. Tel. 613-533-2293. Fax 613-533-6668.

1 Introduction

When estimating regression models for cross-section data, it used to be common for investigators to assume that the error terms (or disturbances) for any pair of observations are uncorrelated. Although this assumption may seem natural, it is actually very strong, and it has two important implications. First, it means that inference can safely be based on "robust" (that is, heteroskedasticity-robust) covariance matrix estimators. For large samples, these estimators are usually quite reliable, although there can be exceptions when a few observations have high leverage (MacKinnon 2013).

A more profound implication of the assumption that error terms are uncorrelated is that information about the parameters accumulates at a rate proportional to the square root of the sample size. If we estimate the same model on two datasets, one with M observations and one with $N = \phi M$ observations, where $\phi >> 1$, the (true) standard errors for the second set of estimates should be approximately $\phi^{-1/2}$ times those for the first set. Of course, this statement assumes that the investigator does not take advantage of the larger sample size by estimating a more complicated model for the sample of size N than for the one of size M. In practice, models tend to become more complicated as sample sizes increase, so that standard errors are not actually proportional to one over the square root of the sample size.

As we discuss in Section 3, it is much less common than it once was to assume that the error terms of regression models are uncorrelated. Instead, investigators commonly assume that they are "clustered." The sample is divided into clusters (which might be associated, for example, with schools, firms, villages, counties, or states), and the disturbances for observations within each cluster are allowed to be correlated. This requires the use of a covariance matrix estimator that is robust to arbitrary patterns of both heteroskedasticity and intra-cluster correlation; see Section 2.

The use of cluster-robust variance estimators in empirical microeconomics began after such an estimator became available in Stata (Rogers 1993). It became much more widespread after a very influential paper (Bertrand et al. 2004) showed that inferences for difference-in-differences (DiD) estimators based on standard errors that ignore autocorrelation within geographical clusters can be extremely unreliable; see Section 6. Cameron and Miller (2015) is an influential survey. More recent surveys include Conley et al. (2018), Esarey and Menger (2019), and MacKinnon (2019).

Failing to allow for intra-cluster correlation has particularly serious consequences when the sample size is large but the number of clusters is not. Thus one important reason for the increased use of cluster-robust standard errors in recent years is that sample sizes have become larger. When cluster sizes are growing with the sample size N, but the number of clusters is fixed, information about the parameters accumulates at a rate slower than $\sqrt{N}$. However, whether or not there is intra-cluster correlation, heteroskedasticity-robust standard errors are always roughly proportional to $1/\sqrt{N}$. Therefore, as we explain in Section 3, using heteroskedasticity-robust standard errors when there is actually intra-cluster correlation causes errors of inference that become more severe as the sample size increases.

In Section 2, we discuss methods of cluster-robust inference based on t-statistics and Wald statistics. In Section 3, we discuss why it often makes sense to divide the sample into clusters and allow for intra-cluster correlation. In Section 4, we discuss how to cluster. The investigator has to choose the appropriate dimension(s) and level(s) of clustering, and this is

often not easy. In Section 5, we discuss several commonly-encountered cases in which using cluster-robust standard errors in the usual way can lead to very serious errors of inference. We also discuss methods that can be used to obtain more reliable inferences, including the wild cluster bootstrap (Cameron et al. 2008), the wild bootstrap (MacKinnon and Webb 2018), and randomization inference. In Section 6, we discuss two empirical examples that illustrate some of the important issues. Section 7 presents some simple Monte Carlo simulations which demonstrate the consequences of getting the level of clustering correct or incorrect. Finally, Section 8 concludes.

2 Regression models with clustered disturbances

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}') = \boldsymbol{\Omega}, \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $N \times 1$ vectors of observations and disturbances, $\boldsymbol{X}$ is an $N \times K$ matrix of exogenous covariates, and $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector. When the $N \times N$ covariance matrix $\boldsymbol{\Omega}$ is equal to $\sigma^2\mathbf{I}$, the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, and we can make inferences based on the estimated covariance matrix $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, where s^2 is $1/(N-K)$ times the sum of squared residuals. When $\boldsymbol{\Omega}$ is diagonal with diagonal elements that differ, OLS is no longer efficient, but it is still consistent, and we can make inferences by using "robust" standard errors based on a heteroskedasticity-consistent covariance matrix estimator, or HCCME (White 1980).

In many cases, however, as we discuss in Sections 3 and 4, there are very good reasons to believe that $\boldsymbol{\Omega}$ is not a diagonal matrix. Suppose instead that the data can be divided into G clusters, indexed by g, where the g^{th} cluster has N_g observations. Then $\boldsymbol{\Omega}$ is assumed to be block-diagonal, with G diagonal blocks that correspond to the G clusters:

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Omega}_G \end{bmatrix}. \tag{2}$$

Here each of the $\boldsymbol{\Omega}_g$ is an $N_g \times N_g$ positive semidefinite matrix, and every element of the off-diagonal blocks in $\boldsymbol{\Omega}$ is assumed to be zero.

The true covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ in the model given by (1) and (2) is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{3}$$

where the $N_g \times K$ matrix $\boldsymbol{X}_g$ contains the rows of $\boldsymbol{X}$ that belong to the g^{th} cluster. Thus the middle factor is actually the sum of G matrices, each of them $K \times K$.

The matrix (3) can be estimated by using the outer product of the residual vector $\hat{\boldsymbol{u}}_g$ with itself to estimate $\boldsymbol{\Omega}_g$ for all g. This yields a cluster-robust variance estimator, or CRVE.

By far the most widely-used version is

$$\mathrm{CV}_1\colon \quad \frac{G(N-1)}{(G-1)(N-K)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{4}$$

The first factor here is analogous to the factor $N/(N-K)$ used in the conventional heteroskedasticity-robust HC_1 covariance matrix (MacKinnon and White 1985), which replaces the middle matrix in (4) by $\sum_{i=1}^{N}\hat{u}_i^2\boldsymbol{X}_i'\boldsymbol{X}_i$. This factor makes CV_1 larger when either G or N becomes smaller, in order to offset the tendency for OLS residuals to be too small. CV_1 evidently reduces to HC_1 when $G = N$, so that each cluster contains just one observation.

The CRVE (4), like all robust covariance matrix estimators, is a "sandwich" estimator. The filling in the sandwich is supposed to estimate the corresponding filling in (3). However, unlike the individual matrices in the summation in (3), the ones in (4) have rank one, even though they are $K \times K$. Therefore, the individual components of the filling in (4) cannot possibly provide consistent estimators of the corresponding components of the filling in (3). Moreover, unless $G \geq K$, the matrix (4) cannot have full rank. It will have rank at most G in some cases and rank at most $G-1$ in others.

Cluster-robust variance estimators were proposed by Liang and Zeger (1986) and Arellano (1987). They became available in Stata about half a decade later (Rogers 1993). However, econometricians did not study their properties under general assumptions until much later. Bester et al. (2011) showed that, under quite restrictive conditions with N increasing and G fixed, cluster-robust t-statistics for $\beta_j = 0$, where β_j is any element of $\boldsymbol{\beta}$, are asymptotically distributed as $t(G-1)$. This result justifies the use of the $t(G-1)$ distribution for calculating critical values and P values, something that has been the default in Stata since 1993.

More recently, Djogbenou et al. (2019) proved that cluster-robust t-statistics are asymptotically normally distributed under rather weak conditions. These require G to increase with N and allow the N_g to increase as well, but not too fast. There are also limits on how much the cluster sizes can vary. Using a similar framework, Hansen and Lee (2019) proved the asymptotic validity of cluster-robust inference based on the standard normal distribution combined with covariance matrix estimators similar to (4) for a wide variety of linear and nonlinear econometric models, including ones estimated by two-stage least squares, the generalized method of moments (GMM), and maximum likelihood.

Although it is by far the most widely used CRVE, the matrix CV_1 defined in (4) is not the only one. An estimator with somewhat better finite-sample properties, which was proposed by Bell and McCaffrey (2002) and advocated by Imbens and Kolesár (2016), is

$$\mathrm{CV}_2\colon \quad (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{M}_{gg}^{-1/2}\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g'\boldsymbol{M}_{gg}^{-1/2}\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{5}$$

where $\boldsymbol{M}_{gg}^{-1/2}$ is the inverse symmetric square root of the matrix $\boldsymbol{M}_{gg} \equiv \mathbf{I}_{N_g} - \boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g'$, which is the g^{th} diagonal block of the $N \times N$ matrix $\boldsymbol{M}_{\boldsymbol{X}} \equiv \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Instead of multiplying by a scalar factor, CV_2 replaces the residual subvectors $\hat{\boldsymbol{u}}_g$ by rescaled subvectors $\boldsymbol{M}_{gg}^{-1/2}\hat{\boldsymbol{u}}_g$. It reduces to the HC_2 HCCME discussed in MacKinnon and White (1985) when $G = N$. The matrix CV_2 can be calculated efficiently in R using the package `clubSandwich` (Pustejovsky 2017). Although CV_2 seems to yield larger and more accurate standard errors

than CV_1, it is considerably more expensive to compute when the clusters are large, because it requires finding the inverse symmetric square root of $\boldsymbol{M}_{gg}$ for each cluster. For sufficiently large clusters, this can be numerically infeasible (MacKinnon and Webb 2018). Recently, Jackson (2020) proposed an alternative estimator which estimates the "filling" of the CRVE by estimating a common variance and common correlation of residuals within each cluster.

As noted above, the conventional way to make inferences about an individual element of the vector $\boldsymbol{\beta}$, say β_j, is to use the cluster-robust t-statistic

$$t_j = \frac{\hat{\beta}_j - \beta_{j0}}{s_j}, \tag{6}$$

where β_{j0} is the value under the null hypothesis and s_j is the square root of the j^{th} diagonal element of either CV_1 or CV_2. The statistic t_j is then compared with the $t(G-1)$ distribution. A $(1-\alpha)\%$ confidence interval for β_j would be

$$\left[\hat{\beta}_j - s_j C_{t(G-1)}(1-\alpha/2),\ \hat{\beta}_j + s_j C_{t(G-1)}(1-\alpha/2)\right], \tag{7}$$

where $C_{t(G-1)}(1-\alpha/2)$ is the $1-\alpha/2$ quantile of the $t(G-1)$ distribution. When G is small, the latter can be considerably larger than the corresponding quantile of the standard normal distribution. In combination with the fact that cluster-robust standard errors are often much larger than heteroskedasticity-robust ones, this can make the interval (7) much wider than a corresponding "robust" interval.

When there are two or more restrictions to be tested, we can use a Wald test. In order to test the hypothesis that $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, where $\boldsymbol{R}$ is an $r \times K$ matrix and $\boldsymbol{r}$ is an $r \times 1$ vector, we compute the Wald statistic

$$W(\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}\hat{\boldsymbol{V}}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \tag{8}$$

where $\hat{\boldsymbol{V}}$ could be either CV_1 or CV_2. $W(\hat{\boldsymbol{\beta}})$ cannot be computed when $r > G$, and often not when $r = G$. The statistic $W(\hat{\boldsymbol{\beta}})$ is then compared with critical values from either the $\chi^2(r)$ distribution or, preferably, r times the $F(r, G-1)$ distribution. Either version of the Wald test is likely to over-reject severely when r is not much smaller than G. In such cases, it is strongly advised to use the wild cluster bootstrap; see Subsection 5.1.

2.1 Multi-way clustering

The CRVEs given in (4) and (5) are designed to handle arbitrary within-cluster correlation in a single dimension. Cameron et al. (2011) and Thompson (2011) independently proposed extensions to handle clustering in two or more dimensions, although the idea actually dates back to Miglioretti and Heagerty (2006). For example, there might be one set of clusters based on geography and another set based on time. Every observation is assumed to belong to one cluster in each of the two dimensions.

The two 2011 papers proposed covariance matrix estimators but did not derive their asymptotic properties. Recent work has developed the theory of multi-way cluster-robust estimators. Davezies et al. (2020) proposed an alternative multi-way CRVE and proved its asymptotic validity, which is technically challenging. Menzel (2018) proposed a multi-way

bootstrap procedure for inference on sample means. MacKinnon et al. (2020a) compared the properties of two forms of two-way CRVE and showed that several variants of the wild cluster bootstrap (Subsection 5.1) can be combined with a two-way CRVE to obtain more reliable inferences about regression coefficients.

3 When to cluster

The simplest way to model intra-cluster correlation is to assume that there are cluster-specific random effects, say v_g. The i^{th} observation in the g^{th} cluster is then equal to

$$y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\beta} + u_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\beta} + v_g + \epsilon_{gi}, \tag{9}$$

where the v_g are independently distributed with variance σ_v^2 and the ϵ_{gi} are independently distributed with variance σ_ϵ^2. This implies that the variance of u_{gi} is $\sigma_v^2 + \sigma_\epsilon^2$, the correlation between disturbances in different clusters is zero, and the correlation between disturbances within the same cluster is $\rho_u = \sigma_v^2/(\sigma_v^2 + \sigma_\epsilon^2)$.

Although the random-effects model (9) is simple and appealing, it seems to us unrealistic in many cases. By assuming that all of the correlation within each cluster comes from a single cluster-specific effect v_g, which affects all observations equally, it rules out any variation in intra-cluster correlations. More realistically, we might expect there to be several cluster-specific effects for each cluster, and for them to affect different observations differently. It might well also be the case that the v_g, the ϵ_{gi}, or both of them are heteroskedastic, with variances that depend on the regressors.

The random-effects model also assumes that the v_g are uncorrelated with all the regressors. This is often a very strong assumption, and it will lead to inconsistent estimates if it is false. The classic way to solve this problem is to treat the v_g as fixed effects, that is, as constants to be estimated instead of as random effects. Then the original regression (1) becomes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{v} + \boldsymbol{\epsilon}, \tag{10}$$

where $\boldsymbol{D}$ is an $N \times G$ matrix, with $D_{gi} = 1$ for observations that belong to cluster g and $D_{gi} = 0$ otherwise. Of course, one column of $\boldsymbol{D}$ must be omitted if $\boldsymbol{X}$ contains a constant term or the equivalent.

The fixed-effects model (10) is very popular. However, it cannot be used whenever any of the regressors varies only at the cluster level, because that regressor would simply be a linear combination of the columns of $\boldsymbol{D}$. Such a model was estimated by Riddell (1979), which led Kloek (1981) to study models in which a dependent variable measured at the individual level is regressed on data measured at the cluster level. The paper showed that conventional standard errors for OLS estimates are biased downwards, often very seriously so, when the disturbances follow the random-effects model (9) with $\rho_u > 0$. Under the assumptions of Kloek (1981), the appropriate way to make inferences is to use feasible generalized least squares (FGLS) for the random effects model. In some cases, this is numerically equal to OLS, but with a different covariance matrix.

Even when a model includes fixed effects, there is no reason to believe that they account for all of the intra-cluster correlation. Using data from the Current Population Survey

(CPS), Bertrand et al. (2004) performed a number of placebo-law experiments, which are an ingenious way to evaluate the performance of inferential procedures using real data. Every replication uses the same data for the regressand and all but one of the regressors. The only thing that differs across replications is the regressor of interest, a treatment dummy that affects certain clusters in certain years. Since the treatment dummies are generated randomly, valid statistical procedures should reject the null hypothesis about as often as the level of the test.

Even though the log-earnings equations of Bertrand et al. (2004) contained both state and year fixed effects, failing to cluster at the state level led to very severe over-rejection in their experiments. This implies that CPS earnings data do not follow a random-effects model. See also MacKinnon (2016), MacKinnon and Webb (2017a), and Brewer et al. (2018). The over-rejection found in all these papers illustrates the fact, discussed below, that even very small amounts of intra-cluster correlation can have a large effect on the accuracy of inferences when N is large (over 500,000 in this case).

In two influential papers, Moulton (1986, 1990) demonstrated via empirical examples that intra-cluster correlation is widespread and that failing to account for it can lead to standard errors that are much too small. Moreover, Moulton (1986) showed that, in the context of the random-effects model, the square of the ratio of the true standard error to the conventional OLS standard error is

$$1 + \rho_x \rho_u \left(\frac{\text{Var}(N_g)}{\bar{N}_g} + \bar{N}_g - 1 \right), \tag{11}$$

where ρ_x is the intra-cluster correlation of the regressor of interest (after it is projected off all other regressors), and $\bar{N}_g$ is the mean of the N_g. The quantity (11) is sometimes called the "Moulton factor." It is evidently 1 when either ρ_u or ρ_x is 0, increases with both ρ_u and ρ_x, and increases without limit as either $\bar{N}_g$ or the ratio of $\text{Var}(N_g)$ to $\bar{N}_g$ increases.

Although the Moulton factor (11) strictly applies only to the random-effects model (9), it provides useful guidance in many cases. In particular, it makes it clear that the extent of intra-cluster correlation for the regressors is just as important as the extent of intra-cluster correlation for the disturbances, and it shows that the errors of inference we make by not allowing for intra-cluster correlation become more severe as the clusters become larger and/or more variable in size.

Another way to see why cluster sizes matter is to consider the matrices $\boldsymbol{X}_g' \boldsymbol{\Omega}_g \boldsymbol{X}_g$ that appear in (3). We can write the k^{th} diagonal element of one of these matrices as

$$\sum_{i=1}^{N_g} \Omega_{g,ii} X_{gki}^2 + 2 \sum_{i=1}^{N_g} \sum_{j=i+1}^{N_g} \Omega_{g,ij} X_{gki} X_{gkj}, \tag{12}$$

where $\Omega_{g,ij}$ is the ij^{th} element of $\boldsymbol{\Omega}_g$, and X_{gki} is the k^{th} element of the row of $\boldsymbol{X}$ corresponding to the i^{th} observation within the g^{th} cluster. If all the off-diagonal elements of $\boldsymbol{\Omega}_g$ were zero, the second term here would be zero, and expression (12) would be $O(N_g)$. Here we have used the "same-order" or "big O" notation, which is a convenient way to indicate how a quantity changes with the sample size N (or, in this case, the cluster size N_g). Informally, $x = O(N)$ means that x behaves like N as N becomes large. The second term

in (12) would have expectation zero if $\mathrm{E}(X_{gki}X_{gkj}) = 0$. But when there is intra-cluster correlation of both the disturbances and the regressors, the second term is $O(N_g^2)$ and does not vanish. Even if the $\Omega_{g,ij}$ were very small, the second term would ultimately dominate, because $O(N_g^2) > O(N_g)$, so that expression (12) itself would be $O(N_g^2)$.

The implications of this result are profound. When there is no intra-cluster correlation, the covariance matrix (3) is $O(N^{-1})O(N)O(N^{-1}) = O(N^{-1})$, as usual. But when the number of clusters G is fixed, and there is intra-cluster correlation that does not die out as N increases, the covariance matrix is instead $O(N^{-1})O(N^2)O(N^{-1}) = O(1)$, because all of the N_g must be proportional to N. This implies that $\hat{\boldsymbol{\beta}}$ is not a consistent estimator when the sample size goes to infinity with a fixed number of clusters; see Andrews (2005). For this reason, any proof of consistency that allows for arbitrary intra-cluster correlation, such as the one in Djogbenou et al. (2019), requires that G tend to infinity along with N.

The number of clusters G does not have to be proportional to N for $\hat{\boldsymbol{\beta}}$ to be consistent. But when G is increasing more slowly than N, both (3) and the CRVEs (4) and (5) that estimate it consistently will tend to zero at a rate slower than $O(N^{-1})$ as $N \to \infty$. This can make reliable inference more difficult for large samples than for small ones. Because a covariance matrix that is robust only to heteroskedasticity is always $O(N^{-1})$, the errors of inference that we make if we fail to allow for intra-cluster correlation may be very severe when N is large and G is not. The same thing is true if we cluster at too fine a level, for instance, clustering by city rather than by state.

As an example, suppose that there are actually G equal-sized clusters, but in computing the CRVE we mistakenly assume that there are instead $4G$ clusters, each of them 1/4 as large. Then the true covariance matrix (3) will involve G expressions like (12), but the CRVE will involve $4G$ such expressions. For the off-diagonal terms, each of these will have 1/16 as many elements in the double summation. The net effect is that the CRVE will sum over only 1/4 of the off-diagonal elements that it should be summing over. Unless the off-diagonal elements that it misses happen to be very small, it is likely that the CRVE will seriously underestimate the true covariance matrix (3). The Monte Carlo simulations in Section 7 provide an example of this and how it causes Type I errors to increase.

3.1 Fixed effects and clustered standard errors

Unless there are explanatory variables that do not vary within clusters, which is the special case considered by Kloek (1981), investigators have the option of including cluster fixed effects. Because these soak up all the variation of the cluster means, only the variation around those means can be used to identify the parameters of interest. If the cluster means of the disturbances are correlated with some of the regressors, then it is important to remove them by including fixed effects. However, we generally lose efficiency by doing so. Whether or not we should include cluster fixed effects, or other types of fixed effect for that matter, is a modeling decision that depends on the phenomena we are studying.

In contrast, whether or not to use a CRVE, and what sort to use, are inference decisions. In the past, it was often considered sufficient to use heteroskedasticity-robust standard errors whenever a model included cluster fixed effects. This would make sense if the data followed the simple random-effects model (9), which implies that the disturbances are equicorrelated. In that case, the fixed effects will explain the cluster-specific v_g terms, and all that remains

of the disturbances will be the ϵ_{gi}, which by assumption are independent.

However, the assumption of equicorrelated disturbances is a very strong one. If there is either serial or spatial correlation within clusters, then the pattern of intra-cluster correlation must be more complicated than (9) allows. The placebo-law results discussed above suggest that this is true of the CPS earnings data. In such cases, it is necessary to employ cluster-robust standard errors even when a model includes cluster fixed effects. Therefore, especially when cluster sizes are large, we believe that not allowing for intra-cluster correlation is usually a bad idea (but see Subsection 3.3 and Section 5).

3.2 Spatial (auto)correlation

Another cause for concern is that the disturbances may display spatial (auto)correlation, so that they are correlated across as well as within geographic clusters. Barrios et al. (2012) suggested that many regressors are correlated beyond state boundaries and that researchers should investigate this possibility. More recently, Kelly (2019) argued that many empirical models of "long differences" or "persistence" are likely to have disturbances that are spatially correlated. The paper suggested that the procedure proposed in Conley (1999) will offer improved but imperfect inference, provided a large bandwidth is specified. Similarly, Ferman (2019) argued that commonly used datasets such as the CPS and the American Community Survey (ACS) exhibit spatial correlation. The paper suggested that multi-way clustering, with one dimension being the cross section, can correct many issues of spatial correlation. However, this approach is not robust to situations in which there is correlation of errors within both different time periods and different groups. The best approach to dealing with unknown spatial correlation in addition to "conventional" clustering is something that clearly warrants further study.

3.3 Design-based and sample-based uncertainty

Up to this point, we have implicitly assumed that any sample we may have is infinitesimally small relative to the population. This evidently makes sense when the population is actually very large. But suppose our sample constitutes a substantial fraction of the population in which we are interested. Should our standard errors take account of this fact? If so, it would seem that they must tend to zero as the sample size tends to the size of the population. But cluster-robust standard errors like those based on (4) do not have this property.

Abadie et al. (2020) explores this setting in the absence of clustering. It argues that many samples constitute substantial fractions of the populations of interest and that standard errors should take this into account. However, the paper points out that, even when the sample consists of the entire population, there may still be uncertainty that needs to be accounted for. In particular, if treatment were randomly assigned to some observations but not to others, outcomes would be stochastic because of the random assignment. This is called design-based uncertainty, because it depends on the experimental design. Under the assumption that observations are independent, the paper provides methods for making valid inferences. These generally yield narrower confidence intervals than conventional (heteroskedasticity-robust) procedures, especially when the sample is large relative to the population.

Abadie et al. (2017) extends the design-based approach of Abadie et al. (2020) to the case in which disturbances may be correlated within clusters, as often happens with "natural" experiments that are analyzed using difference-in-differences models. As we discussed above, it is important to take clustering into account when there is a high degree of correlation in an explanatory variable within clusters, as measured by ρ_x in expression (11). In the DiD setting, some states or groups are treated, at least for some observations, while others are not. Thus ρ_x can be quite high, and using a CRVE can greatly improve inference (but see Section 5, especially Subsection 5.3). These situations also occur frequently in the settings of lab and field experiments. Abadie et al. (2017) argue that cluster-robust inference is necessitated by the design-based uncertainty in these settings. Specifically, they argue that, when the assignment to treatment is correlated within clusters, then cluster-robust standard errors that take account of design-based uncertainty are required. Note that these are more complicated than ones based on (4).

Cluster-robust inference is also needed whenever there is a clustered sampling design. In many cases, the population contains a large number of groups of observations, and only some of those groups are included in the sample. For example, many education samples contain observations from all students within some schools but no students from other schools. Clustered sampling also often occurs within larger surveys, such as the CPS, where sampling is typically done by census tract. For each state, all of the households in a given sample may be drawn from a relatively modest proportion of the census tracts within the state. For this sort of survey, the survey design is often quite complex, and it can create both heteroskedasticity and within-state correlation; see Kolenikov (2010), among others.

Abadie et al. (2017) suggested that clustering may be too conservative in settings where there is neither cluster-specific treatment assignment nor a clustered sampling design. In other words, it may be too conservative when there is neither design-based nor sample-based uncertainty. In the latter case, we might have a nationally representative dataset that, unlike the CPS, does not involve sampling by census tract or other geographical subclusters.

The arguments in Abadie et al. (2020, 2017) are interesting and provocative. However, they depend critically on the assumption that the sample is large relative to a finite population in which the investigator is interested. We believe that, in many cases, economists are implicitly interested not in an actual population but in a meta-population from which they imagine the former to have been drawn. For example, if we have data for 50 U.S. states, with no sampling and no experimental design involved, then we can either view those data as non-random quantities, or we can view them as 50 draws from a meta-population of states. If we take the former view, then all we can do is to report some numbers that characterize the population. But if we take the latter view, then we can perform statistical inference in the usual way.

Although the idea of a meta-population may seem odd, economists implicitly make use of it whenever they analyze time-series data (and many other types of data). For instance, if history gives us aggregate inflation data for some country from 1968 to 2019, then we cannot obtain another dataset for the same time period by drawing a new sample. However, we can imagine that the data were generated by some sort of data-generating process (DGP) that characterizes the meta-population of inflation rates and other macro variables, and we can attempt to estimate key features of that DGP.

If we are studying economic history, then we inevitably have to focus on what happened at

a particular time in a particular place. However, if we are using data about what happened at one or more times in one or more places to make general statements about how, for example, certain policies affect certain outcomes, then we have to use what Abadie et al. (2017) calls the "model-based" approach. This approach implicitly involves the idea of a meta-population. The DGP is simply a model like (1) accompanied by a way of obtaining the matrix $\boldsymbol{X}$ and the vector $\boldsymbol{u}$ from random submatrices $\boldsymbol{X}_g$ and subvectors $\boldsymbol{u}_g$ associated with clusters chosen at random from a meta-population of clusters. In general, standard cluster-robust inference is valid within a meta-population framework, and the finite-population arguments of Abadie et al. (2017) do not apply.

4 How to cluster

In order to obtain reliable cluster-robust inferences, the most important decision to be made in many cases is how to divide the sample into clusters. In this section, we attempt to provide some guidance. We take the model-based approach and assume that some level of clustering is appropriate (but see Subsections 3.2 and 3.3). We consider several guiding principles for determining the level of clustering.

The key assumption for cluster-robust inference to be valid is that the disturbances are arbitrarily correlated within clusters but uncorrelated across clusters. It is important to specify the level of clustering in such a way that this assumption is likely to be true, or at least to provide a good approximation.

4.1 Cluster at the coarsest or most aggregate level

When there is more than one level at which to cluster, and the levels are nested, then one should generally cluster at the coarsest feasible level (Cameron and Miller 2015). Suppose, for example, that we can cluster either by city or by state. Clustering by state captures all the within-city correlation, and it also allows for the disturbances to be correlated within states but across cities. In contrast, clustering by city assumes that all of the correlations across cities but within states are zero. If that assumption is false, then standard errors are very likely to be too small.

Of course, there is a downside to clustering at too coarse a level (over-clustering). The smaller is the number of clusters, for a given sample size, the larger is the number of elements of $\boldsymbol{\Omega}$ that implicitly have to be estimated. If we cluster too coarsely, many of these elements are actually zero, and trying to estimate a large number of zeros inevitably makes the CRVE noisier. Even in the ideal case in which the $t(G-1)$ distribution provides a good approximation, making G smaller will cause test power to fall and confidence intervals to become wider, simply because the critical values for the t distribution increase as $G-1$ becomes smaller. However, unless clustering at the highest feasible level means using a value of G that is very small, the loss of power is likely to be modest in comparison with the severe size distortions that can occur from clustering at too low a level (under-clustering); see Section 7. This becomes more true as the sample size becomes larger.

Another problem with over-clustering is that, when G is small, the $t(G-1)$ distribution often does not provide a good approximation. But, except in the most extreme cases, this

problem can generally be overcome by using bootstrap methods; see Subsection 5.1 for a discussion of these methods and Section 7 for simulation evidence.

4.2 Cluster at least at the level of a policy change

When the null hypothesis of interest involves a treatment variable resulting from a policy change, one should always cluster at a level no finer than the one at which the policy was applied. As mentioned in Subsection 3.3, Abadie et al. (2017) suggest that clustering is necessary whenever treatment is assigned at the cluster level. From this perspective, one would want to match the level of clustering done in the analysis to the level at which treatment was assigned. For instance, in a randomized control trial, if treatments were assigned at the village level, then one would want to cluster at the village level. However, based on the arguments of Subsection 4.1, one might choose to cluster at a still coarser level, especially if the number of villages were large and they naturally fell into a reasonable number of larger groups. Ideally, both approaches would yield similar results.

Whether or not there is a policy change, it always makes sense to cluster at a level no finer than the one at which observations were included in the sample. For example, if classrooms were chosen at random for inclusion in the sample, then one would want to cluster at either the classroom level or the school level. But if schools were chosen at random, then one would want to cluster at the school level. In both cases, of course, one might choose to cluster at an even coarser level, such as school districts.

4.3 Cluster at the cross-section level for panel data

When working with panel data and repeated cross-section data, it is important never to cluster below the cross-section level. As shown first in Bertrand et al. (2004), clustering at the level of the cross section allows for arbitrary autocorrelation of the error terms within cross-sectional units. In many contexts, this means that clustering at the state level will result in much more reliable inference than, say, clustering at the state $\times$ year level.

An even more general approach for this sort of data would be to use two-way clustering by cross-sectional unit and time. MacKinnon (2019) provided evidence that this seems to be appropriate in the context of an earnings equation using CPS data that includes both state and year fixed effects.

4.4 There is no golden number of clusters

Early simulation results such as those in Bertrand et al. (2004) and Cameron et al. (2008) concerned models with balanced clusters. This gave a false sense of how well cluster-robust variance estimators perform in finite samples. A rule of thumb emerged that $G \geq 50$ would allow for reliable inference, which was changed (in jest) to $G \geq 42$ in Angrist and Pischke (2008). However, any such rule of thumb can be extremely misleading, because all CRVEs tend to become less reliable as the clusters become more unbalanced; see Imbens and Kolesár (2016), Carter et al. (2017), MacKinnon and Webb (2017a), and Djogbenou et al. (2019). The problems associated with unbalanced clusters are discussed in Subsection 5.2.

It is far more important to get the level of clustering right, as we have discussed in Subsections 4.1, 4.2, and 4.3, than it is to ensure that G is large enough for t-statistics to have their namesake distribution. When we cluster at too fine a level, standard errors will typically be too small by a factor that increases with the sample size. In contrast, when we cluster at the right level, inference based on the $t(G-1)$ distribution may be seriously unreliable, but other methods of inference (notably the bootstrap methods discussed in Subsection 5.1) often provide quite reliable inferences.

4.5 In many settings, over-clustering is mostly harmless

In a model-based context, over-clustering (within reason) tends to be relatively harmless, except in one important special case (Subsection 5.3). Over-clustering can mean either clustering at a coarser level than is actually appropriate or clustering in two dimensions when just one is needed. Simulation results suggest that, in most cases, a moderate amount of over-clustering should have little impact on size (provided the wild cluster bootstrap, discussed in Subsection 5.1, is used) but some impact on power; see Section 7.

Of course, there are limits to the amount of over-clustering that can be handled safely, even when using the wild cluster bootstrap. Although bootstrap P values are often very reliable even when G is quite small (for example, they work remarkably well in the simulations of Section 7 when $G = 10$ and cluster sizes are quite unbalanced), there are extreme cases, discussed in Section 5, where they cannot be relied upon.

4.6 Tests for the level of clustering

It is often difficult to choose the appropriate level of clustering on theoretical grounds, and this choice can be aided by the use of formal statistical tests. Two such tests exist. Ibragimov and Müller (2016) proposed a procedure for testing a null hypothesis of fine clustering against an alternative of coarse clustering. For example, it can test a null of heteroskedasticity against an alternative of city-level clustering, or a null of city-level clustering against an alternative of state-level clustering. However, because the IM test requires the model to be estimated on a cluster-by-cluster basis, it implicitly assumes that there are cluster-level fixed effects. This also means that it cannot be used when the regressor of interest is invariant within each cluster, which will often be the case for models of treatment effects.

MacKinnon et al. (2020b) proposed two closely related procedures which directly test whether the standard error for a single coefficient, or the covariance matrix for two or more coefficients, are based on the correct level of clustering. These "score-variance" tests are based on the difference between the variance of the scores for two nested levels of clustering. Unlike the IM test, they can be used whether or not there are regressors that are invariant within clusters and whether or not there are fixed effects at the level of either fine or coarse clusters. The test statistics are either asymptotically standard normal or, when they are based on two or more coefficients, asymptotically chi-squared. The paper also proposed bootstrap versions of the tests, which sometimes have much better finite-sample properties than the asymptotic versions. This seems to be especially true for models with fixed effects at the coarse-cluster level and regressors that vary at the fine-cluster level.

5 What can go wrong

Although both CV_1 and CV_2, given in expressions (4) and (5), estimate the true covariance matrix (3) consistently under moderately weak conditions (Djogbenou et al. 2019), they do not always provide reliable estimates, even when the sample size is very large. This reflects the fact that all CRVEs differ fundamentally from most other covariance matrix estimators in one important respect. The latter usually converge to the true value as the number of observations, N, tends to infinity. But a CRVE converges to the true value as the number of clusters, G, tends to infinity. Therefore, no matter how large N may be, inference based on cluster-robust standard errors (for the correct set of clusters) can sometimes be problematic, perhaps even seriously misleading, when G is not large.

In this section, we assume that the clusters have been chosen correctly, with no correlation of disturbances across clusters. Nevertheless, there are three situations in which cluster-robust inference may be unreliable. The first is when the number of clusters is small. The second is when cluster sizes, or other features of the clusters, are seriously unbalanced. The third, which can be thought of as a special case of the second, is when the model focuses on the effects of a treatment dummy, and few clusters are "treated." In the first two cases, we recommend using a particular bootstrap method, which we describe in Subsection 5.1. This method can also work well in the third case, but it can sometimes fail disastrously. When it does, randomization inference (Subsection 5.4) or an alternative bootstrap procedure may be able to provide reliable inferences.

5.1 Few clusters

When the number of clusters is reasonably large (say, a few hundred), and each cluster provides roughly the same amount of information, then inference based on cluster-robust standard errors and the $t(G-1)$ distribution is likely to be very reliable. There are also cases in which this type of inference works well even when G is quite small (Bester et al. 2011), but it would usually be unwise to rely on it.

Because the middle factor in any CRVE is a sum over G matrices, each with rank one, it should be obvious that a CRVE may not provide reliable inferences when G is small. Ideally, there would be more clusters than parameters, so that the CRVE could potentially have full rank. This is more important for Wald tests of several restrictions than for t-tests of just one restriction. But it makes sense that G should need to be larger for reliable inference when K (the number of regression coefficients) is large than when it is small. To our knowledge, there have been no simulation studies that focus on the relationship between G and K for possibly large K. However, simulations in Appendix C.2 of Djogbenou et al. (2019) suggest that adding either 4 or 8 additional regressors when $G = 25$ makes tests based on the $t(24)$ distribution noticeably more prone to over-reject.

Carter et al. (2017) proposed the concept of an "effective number of clusters," which they called G^*, and provided a way to compute an approximation to it; a Stata package called `clusteff` is discussed in Lee and Steigerwald (2018). Although the value of G^*, both absolute and relative to G, can provide helpful guidance, there is not currently enough evidence to let us say that, for example, cluster-robust standard errors are reliable whenever $G^* \geq 50$. However, that would surely be a much safer rule of thumb than $G \geq 50$. In our experience,

when G^* is not much less than G and not too small (say, 50 or more), inference based on a cluster-robust t-statistic and the $t(G-1)$ distribution generally seems to work well. In contrast, when G^* is much smaller than G, that type of inference can be very unreliable.

Another approach, which almost always provides more reliable inferences than the $t(G-1)$ distribution, is to rely on bootstrap tests and bootstrap confidence intervals. The basic idea of bootstrap testing is to compare a test statistic with the empirical distribution of a large number of bootstrap test statistics computed from simulated samples. Conceptually, a bootstrap confidence interval is then a set of parameter values for which a bootstrap test does not reject. Accessible introductions to bootstrap methods include MacKinnon (2002), Davidson and MacKinnon (2006a), and Horowitz (2019).

Suppose that we wish to test the restriction $\boldsymbol{a}'\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{a}$ is a known vector. For example, if we are testing the hypothesis that $\beta_k = 0$, the k^{th} element of $\boldsymbol{a}$ would be 1 and all the rest would be 0. We first calculate a cluster-robust t-statistic, say t_a, for the hypothesis that $\boldsymbol{a}'\boldsymbol{\beta} = \mathbf{0}$. We then generate a large number (B) of bootstrap samples indexed by b and use each of them to compute a bootstrap t-statistic t_a^{*b}. Sensible values of B are numbers like 999 and 9,999 (Racine and MacKinnon 2007). Then a symmetric two-tailed test is based on the bootstrap P value

$$p^*(t_a) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\Big(|t_a^{*b}| > |t_a|\Big), \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The more extreme is $|t_a|$, the fewer of the $|t_a^{*b}|$ should exceed it by chance. If, for example, 17 out of 999 do so, then $p^*(t_a) = 0.017$, and we can confidently reject the null hypothesis at the .05 level.

In the context of the model (1), we want to calculate t_a^{*b} from the b^{th} bootstrap sample in precisely the same way as we calculated t_a from the actual sample. How well bootstrap tests perform then depends on how the bootstrap samples used to calculate the t_a^{*b} are generated. In principle, there are many ways to do so. However, both theory and simulation evidence currently favor a particular method, namely, the restricted wild cluster bootstrap that uses the Rademacher distribution.

The wild cluster bootstrap was proposed by Cameron et al. (2008) and studied extensively in MacKinnon and Webb (2017a). Its asymptotic validity was proved by Djogbenou et al. (2019). The equation used to generate the b^{th} bootstrap sample for the restricted wild cluster (WCR) bootstrap is

$$\boldsymbol{y}_g^{*b} = \boldsymbol{X}_g\tilde{\boldsymbol{\beta}} + \boldsymbol{u}_g^{*b} = \boldsymbol{X}_g\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{u}}_g v_g^{*b}, \quad g = 1, \ldots, G, \tag{14}$$

where quantities with g subscripts are scalars, subvectors, or submatrices associated with the g^{th} cluster. In (14), $\tilde{\boldsymbol{\beta}}$ is the vector of least squares estimates of $\boldsymbol{\beta}$ subject to the restriction or restrictions to be tested, and $\tilde{\boldsymbol{u}}_g$ is the restricted residual vector for cluster g. The scalar v_g^{*b} is an auxiliary random variable that multiplies the entire vector $\tilde{\boldsymbol{u}}_g$. It follows the Rademacher distribution, which takes values 1 and -1, each with probability 1/2.

The key idea of the WCR bootstrap, or WCRB, is that the bootstrap disturbances $\boldsymbol{u}_g^{*b}$ for cluster g are generated by multiplying the residual subvector $\tilde{\boldsymbol{u}}_g$ by the scalar random variable v_g^{*b}. This ensures that, asymptotically, the disturbances for cluster g in the bootstrap DGP (14) on average (over actual samples) have covariance matrix $\boldsymbol{\Omega}_g$. In consequence, estimates based on the bootstrap sample $\boldsymbol{y}^{*b}$ have the same asymptotic distribution as ones

based on the actual sample $\boldsymbol{y}$, assuming the restrictions are true. This implies that, if we reject the null hypothesis whenever the bootstrap P value in (13) is less than α, and $\alpha(B+1)$ is an integer, the asymptotic level of the test is α.

Nothing in the above arguments implies that a WCRB test will always perform better in finite samples than a test based on the $t(G-1)$ distribution. However, higher-order theory in Djogbenou et al. (2019) does strongly suggest that this is likely to be the case. It also suggests that the Rademacher distribution (Davidson and Flachaire 2008) is usually the best choice for the auxiliary distribution and that failing to impose the null hypothesis on the bootstrap DGP is a bad idea. In all cases, simulation evidence supports these implications of the theory.

Equation (14) suggests that we need to generate B bootstrap samples of size N and compute a bootstrap t-statistic t_a^{*b} for each of them. This can be computationally challenging when N is large, especially if K is also large. Luckily, there is a way to reduce the computational burden dramatically, especially when G is small. Since the bootstrap DGP (14) satisfies the restrictions, the numerator of the bootstrap t-statistic for $\boldsymbol{a}'\boldsymbol{\beta} = \mathbf{0}$ is

$$\boldsymbol{a}'\hat{\boldsymbol{\beta}}^{*b} = \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}^{*b} = \sum_{g=1}^{G} \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g'\tilde{\boldsymbol{u}}_g v_g^{*b}. \tag{15}$$

The quantities $\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g'\tilde{\boldsymbol{u}}_g$ are scalars that can be calculated after the restricted model has been estimated but before bootstrapping begins. The rightmost expression in (15) is then just the sum over the G clusters of those scalars times the realized Rademacher random variables v_g^{*b}. This expression can be used to calculate $\boldsymbol{a}'\hat{\boldsymbol{\beta}}^{*b}$ extremely rapidly, with computer time that is $O(G)$.

A similar, but much trickier, procedure can be used to calculate the denominator of the bootstrap t-statistic efficiently. The computations for each bootstrap sample are now $O(G^2)$ instead of $O(G)$. However, for large values of N and even moderately large values of G, this can still be very much less expensive than generating a bootstrap sample according to (14) and computing $\boldsymbol{a}'\hat{\boldsymbol{\beta}}^{*b}$ and its cluster-robust standard error in the usual way. The Stata package `boottest` (Roodman et al. 2019) implements this efficient algorithm. Unless N is extremely large (so that the preliminary calculations are time-consuming) or G is greater than several hundred (in which case bootstrapping is probably unnecessary), it is usually very inexpensive to compute the bootstrap P value (13) for $B = 9,999$ or even $B = 99,999$. The R package `clusterSEs` (Esarey 2018) also implements the WCRB but, at time of writing, does not employ the computational tricks used by `boottest`.

Because it is usually inexpensive, we recommend employing the WCRB almost all the time, unless G is quite large. For sufficiently large values of N, this will actually be cheaper than many alternative procedures, such as computing CV_2. When the bootstrap test yields essentially the same conclusion as a test based on the $t(G-1)$ distribution, then we can usually be confident that both results are reasonably reliable. On the other hand, when the bootstrap test provides weaker evidence against the null hypothesis than the asymptotic test (alas, it very rarely provides stronger evidence), then we should usually disregard the latter (but see Subsection 5.3). It is difficult to say whether we should rely on the results of a bootstrap test when they differ sharply from those of the corresponding asymptotic test. This will depend on how well the WCRB is known to perform in similar circumstances.

The WCRB can also be used to form confidence intervals by "inverting" the bootstrap test, and `boottest` does this by default for tests of a single restriction. For details, see Roodman et al. (2019, Section 3.5). These bootstrap confidence intervals can be much more accurate than conventional ones based on cluster-robust standard errors and the $t(G-1)$ distribution (MacKinnon 2015).

One problem with the WCRB is that the number of distinct bootstrap samples with the Rademacher distribution (or any other two-point distribution) is just 2^G. When $G < 10$, this may be too small for p^* to be a reliable estimate. Webb (2014) therefore proposed a six-point distribution which largely solves this problem, because $6^G >> 2^G$. When 2^G is reasonably large but smaller than the chosen value of B, it is better to enumerate all possible bootstrap samples than to draw them at random, and `boottest` does this by default.

Nevertheless, the WCRB often works surprisingly well even for quite small values of G. In fact, Canay et al. (2020) studied situations in which the WCRB using the Rademacher distribution yields exact inferences for large samples even when the number of clusters is fixed and quite small. However, these results require fairly strong homogeneity conditions on the distribution of the covariates across clusters, conditions which emphatically fail to hold in the situations discussed in Subsections 5.2 and 5.3. The analysis of Canay et al. (2020) makes use of a remarkable relationship between the wild cluster bootstrap and randomization inference, a topic to be discussed in Subsection 5.4.

Situations in which inference is particularly difficult are discussed in the next two subsections. Even the WCRB can be unreliable, especially when there are few treated clusters (Subsection 5.3). When there is doubt about its reliability, it would be wise to employ other methods as well. In particular, Imbens and Kolesár (2016) and Bell and McCaffrey (2002) suggested procedures based on CV_2, both of which compute (somewhat different) numbers smaller than $G-1$ to be used as the degrees of freedom for the t distribution. Unfortunately, these procedures are computationally burdensome when N is large (MacKinnon and Webb 2018). A related procedure that is based on CV_1 and is computationally feasible even for very large samples was suggested by Young (2016). Limited simulation evidence suggests that the Imbens-Kolesár and Young procedures do not, in general, perform as well as the WCRB, but they often perform quite well, and they can yield different results in some cases. It is probably safe to accept the inferences from the WCRB when they agree with those from these alternative procedures.

The procedures we have discussed are all based on OLS estimation of (1) using the entire sample. The procedure proposed in Ibragimov and Müller (2010) is instead based on estimating the model on a cluster-by-cluster basis, but this is only feasible if all the regressors of interest vary within each cluster. The procedure of Ibragimov and Müller (2016) overcomes this problem by combining the original clusters into fewer and larger ones, if necessary.

5.2 Unbalanced clusters

The asymptotic validity of inference based on both t-statistics and the wild cluster bootstrap depends on the properties of the score vectors $\boldsymbol{s}_g = \boldsymbol{X}_g'\boldsymbol{u}_g$ (Djogbenou et al. 2019). Ideally, all of them would follow the same multivariate distribution, with covariance matrices $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ the same for all g. Of course, this is rarely the case in practice, and the theory allows for considerable heterogeneity. But when the score vectors are too heterogeneous across clusters,

the conditions for parameter estimates and t-statistics to have their usual asymptotic normal distributions, and for the bootstrap distributions to converge to the asymptotic ones, are no longer satisfied. This suggests that inference will become less reliable as the data become more heterogeneous across clusters.

In particular, standard theoretical results do not apply when cluster sizes vary too much. Some of the simulations in Djogbenou et al. (2019) have one cluster that contains half the observations. This case may seem extreme, but it is precisely what empirical studies of state laws and corporate governance in the United States encounter when they cluster at the state level, because roughly half of all incorporations are in Delaware (Spamann 2019). In this extreme case, tests based on the $t(G-1)$ distribution over-reject more severely, not less, as G increases. WCRB tests also over-reject more severely as G increases, but to a much lesser extent. In less extreme cases, where the single large cluster becomes a smaller proportion of the sample as G increases, the performance of WCRB tests always improves with G. Once the large cluster becomes small enough (on the order of 20% of the sample in these experiments), the bootstrap tests work very well. But this can require quite a large number of clusters, perhaps on the order of several hundred.

Variation in cluster sizes is not the only sort of heterogeneity that is likely to cause cluster-robust tests to be misleading. Even if all clusters are roughly the same size, the covariance matrices $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ can vary across them for other reasons. Perhaps a few clusters contain a lot more information than the remaining ones, or perhaps the disturbances are heteroskedastic across clusters. In both cases, we would expect all methods to make more serious inferential errors than they would if the model had the same number of homogeneous clusters. However, simulation evidence always suggests that the WCRB is less affected by heterogeneity than $t(G-1)$ tests. This reinforces our earlier recommendation to employ the WCRB almost all the time.

Although we cannot directly observe the $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ matrices, we can observe the N_g and the $\boldsymbol{X}_g'\boldsymbol{X}_g$ matrices, which provide measures of how much information each cluster contains. The effective number of clusters G^* (Carter et al. 2017) is quite sensitive to heterogeneity across the $\boldsymbol{X}_g'\boldsymbol{X}_g$. Thus finding that $G^* \ll G$ provides a useful warning. However, G^* is not sensitive to heteroskedasticity across clusters. To see whether that is a problem, we can calculate the variance of the residuals for each cluster separately.

Because the finite-sample properties of all inferential methods, including the WCRB, depend in complicated ways on the model and dataset, it is difficult to provide a rule of thumb for when it is safe to rely on WCRB P values. These are most likely to be unreliable (often too small, but sometimes too large, as we will see in Subsection 5.3) when G is small, when cluster sizes vary a lot, when the $\boldsymbol{X}_g'\boldsymbol{X}_g$ vary a lot (which is likely to cause $G^* \ll G$), and/or when the variance of the residuals differs sharply across clusters. In such cases, as we recommended above, it is important to employ other methods as well. Of course, we are not suggesting the use of multiple inferential methods as a fishing expedition, but rather as a way of verifying (or casting doubt on) the validity of the WCRB.

5.3 Few Treated Clusters

Many applications of cluster-robust inference involve treatment effects estimated at the cluster level. In what we call the pure treatment case, some schools or villages or experimental

subjects are treated, and others are not. Thus every observation in every treated cluster is treated. In contrast, for difference-in-differences (DiD) models, some jurisdictions are never treated, and others are treated during some, but not all, time periods. In either of these cases, when the number of treated clusters is small, cluster-robust standard errors can be very much too small, even if the total number of clusters is large. Because this situation is commonly encountered, it is worth discussing the issues associated with few treated (or few control) clusters in some detail.

Following MacKinnon and Webb (2017a), consider the pure treatment model

$$y_{gi} = \beta_1 + \beta_2 d_{gi} + u_{gi}, \quad g = 1, \dots, G, \quad i = 1 \dots, N_g, \tag{16}$$

where d_{gi} equals 1 for the first G_1 clusters and 0 for the remaining $G_0 = G - G_1$ clusters. Every observation in the g^{th} cluster is either treated ($d_{gi} = 1$) or not treated ($d_{gi} = 0$). The analysis would be more complicated if we included additional regressors, but it would not change in any fundamental way.

The key problem is that, because the dummy variable d_{gi} must be orthogonal to the residuals, the latter must sum to zero over all the treated observations. This has some unfortunate implications. Suppose that $\boldsymbol{d}_1$ denotes the subvector for cluster 1 of the vector $\boldsymbol{d}$ with typical element d_{gi}. Then, in the extreme case in which only cluster 1 is treated, $\boldsymbol{d}_1'\hat{\boldsymbol{u}}_1 = 0$. In other words, the residuals for cluster 1 must sum to zero. This implies that $\boldsymbol{d}_1'\hat{\boldsymbol{u}}_1\hat{\boldsymbol{u}}_1'\boldsymbol{d}_1 = 0$. But that quantity is supposed to estimate the element $\boldsymbol{d}_1'\boldsymbol{\Omega}_1\boldsymbol{d}_1$ of the matrix $\boldsymbol{X}_1'\boldsymbol{\Omega}_1\boldsymbol{X}_1$. Because most of the information about β_2 in (16) is coming from cluster 1, which is the only treated cluster, $\boldsymbol{d}_1'\boldsymbol{\Omega}_1\boldsymbol{d}_1$ needs to be estimated accurately if we are to obtain a reliable standard error for $\hat{\beta}_2$. But since $\boldsymbol{d}_1'\hat{\boldsymbol{u}}_1\hat{\boldsymbol{u}}_1'\boldsymbol{d}_1 = 0$, the cluster-robust ($\text{CV}_1$) standard error for $\hat{\beta}_2$ is actually very much too small. This causes the cluster-robust t-statistic to be very much too large. When $G_1 = 1$, it would not be unusual for the t-statistic to be five times as large as it should be. In this extreme case, the CV_2 covariance matrix cannot even be computed (MacKinnon and Webb 2018).

When there are two treated clusters, $\boldsymbol{d}_1'\hat{\boldsymbol{u}}_1 + \boldsymbol{d}_2'\hat{\boldsymbol{u}}_2 = 0$. This implies that $\boldsymbol{d}_1'\hat{\boldsymbol{u}}_1\hat{\boldsymbol{u}}_1'\boldsymbol{d}_1$ and $\boldsymbol{d}_2'\hat{\boldsymbol{u}}_2\hat{\boldsymbol{u}}_2'\boldsymbol{d}_2$, although both non-zero, underestimate $\boldsymbol{d}_1'\boldsymbol{\Omega}_1\boldsymbol{d}_1$ and $\boldsymbol{d}_2'\boldsymbol{\Omega}_2\boldsymbol{d}_2$ severely. The problem diminishes as the number of treated clusters increases; see MacKinnon and Webb (2017a, Appendix A.3). Although this argument is for the pure treatment model, essentially the same argument applies to DiD models. Whenever only cluster 1 is treated, $\boldsymbol{d}_1'\hat{\boldsymbol{u}}_1 = 0$, because the residuals for the treated observations sum to zero, and the residuals for the control observations are multiplied by elements of $\boldsymbol{d}_1$ that equal 0.

Unfortunately, bootstrapping does not solve the problem. As MacKinnon and Webb (2017a, Section 6) explains, the WCRB always under-rejects very severely when $G_1 = 1$. In simulation experiments with 400,000 replications (MacKinnon and Webb 2017b), there are often no rejections at all for tests at the .05 level. There is also typically very severe under-rejection for $G_1 = 2$. The bootstrap fails in this case because the absolute values of the actual and bootstrap test statistics are strongly positively correlated. When $|t_a|$ is large, the $|t_a^{*b}|$ tend to be large as well, so that the bootstrap P value in (13) is not likely to be small. The extent of the under-rejection depends on G, G_1, the cluster sizes, and the numbers of treated observations within the treated clusters. For given values of G and G_1, the problem tends to be most severe when the number of treated observations is small.

In order to avoid this problem, MacKinnon and Webb (2018) suggested using the ordinary wild bootstrap, which uses one auxiliary random variable per observation instead of one per cluster. Surprisingly, even though the distribution of the $\hat{\boldsymbol{\beta}}^{*b}$ does not coincide asymptotically with the distribution of $\hat{\boldsymbol{\beta}}$, tests based on the ordinary wild bootstrap are asymptotically valid (Djogbenou et al. 2019). Moreover, in some circumstances, these tests can perform very well, even when $G_1 \leq 2$. In the pure treatment case, the key requirement is that all clusters be the same size. However, there are also many cases in which ordinary wild bootstrap tests (based on cluster-robust t-statistics) either over-reject or under-reject systematically. In our view, these tests are worth trying when G_1 is very small and the WCRB does not reject. However, one should not rely on their results unless simulation evidence suggests that they perform well for the case at hand. Empiricists may want to conduct their own simulations to assess the validity of inference procedures given the structure of their data.

A very different approach, which can be used when there is just one treated cluster but quite a few control clusters, is the method of "synthetic controls" proposed in Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010). The idea is to compare the outcomes for the treated cluster with a weighted average of outcomes for the control clusters, the weights being chosen so that the synthetic control resembles the treated cluster in the pre-treatment period. In particular, if we are interested in a treatment that started at a certain date, the weights may be chosen to make the pre-treatment outcomes for the synthetic control as close as possible to the pre-treatment outcomes for the treated cluster.

Several other procedures have been suggested to deal with the problems of inference for unbalanced and few treated clusters. Some of the most interesting ones are based on randomization inference, and these will be discussed in the next subsection. We have already mentioned the procedures of Bell and McCaffrey (2002), Imbens and Kolesár (2016), and Young (2016), which employ either an alternative CRVE, a method of estimating degrees of freedom, or both. These all work better than simply using CV_1 and the $t(G-1)$ distribution; simulation results for these procedures can be found in the appendix of MacKinnon and Webb (2018). Donald and Lang (2007) proposed a two-step estimator in which the data are collapsed into pre-treatment and post-treatment means for the treatment and control clusters. The procedure of Ibragimov and Müller (2016), mentioned in Subsection 5.1, allows for inference so long as there are at least two treated and two untreated clusters, although power will be much higher with at least four of each. Finally, Ferman and Pinto (2019) suggested a DiD procedure that works for few treated and many control groups when there is heteroskedasticity of known form across clusters.

5.4 Randomization Inference

Randomization inference, or RI, refers to a family of procedures that do not involve comparing a test statistic with either an asymptotic or a bootstrap distribution. Instead, an actual parameter estimate (or an actual test statistic) is compared with a set of hypothetical estimates (or hypothetical test statistics) obtained by re-randomizing. There is a large literature on RI, dating back to Fisher (1935). For modern treatments, see Lehmann and Romano (2005, Chapters 5 and 15) and Imbens and Rubin (2015, Chapter 5). Randomization tests can work extraordinarily well in some cases, even when G_1 is very small. An interesting application of randomization inference is Young (2019).

There is more than one way to perform randomization tests in the context of DiD models and treatment models like (16). The simplest approach is to consider all possible assignments of treatment to clusters. Then each re-randomization involves pretending that a particular set of clusters was actually treated. The values of the dependent variable do not change across re-randomizations, but the values of the treatment dummy do change. For example, suppose there are $G = 15$ clusters, of which $G_1 = 2$ are treated. Then there are $(15 \cdot 14)/2 = 105$ ways in which treatment could have been assigned. One of them corresponds to the actual sample, and the other 104 correspond to re-randomizations. If the estimate $\hat{\beta}_2$ is sufficiently extreme compared with the 104 estimates associated with the re-randomizations, then it seems reasonable to reject the null hypothesis of no treatment effect.

An RI procedure based on parameter estimates that is similar, but not identical, to the one just outlined was suggested for DiD models by Conley and Taber (2011). MacKinnon and Webb (2020) pointed out that randomization tests could also be based on cluster-robust t-statistics and studied the properties of both procedures. When G is sufficiently large, all clusters are identical (in terms of size, error variance, within-cluster correlation, etc.), and treatment is assigned at random, both forms of RI test work extremely well under the null hypothesis, but the one based on coefficient estimates has more power. However, when clusters are not identical, and the treatment status of each cluster is not hidden from the investigator, both tests can either over-reject or under-reject under the null. The one based on t-statistics generally performs better for $G_1 > 1$, especially when the treated clusters are larger or smaller than the controls. The Stata package `ritest`, described in Hess (2017), can be used to perform these and other randomization tests.

Spamann (2019) proposed an RI test similar to the one based on t-statistics discussed in MacKinnon and Webb (2020), but modified slightly to take account of Delaware's unique status. In simulation experiments with 51 clusters where half the sample belonged to one cluster, this test generally performed better than using either the $t(G-1)$ distribution or the wild cluster bootstrap. However, not surprisingly, its performance in this extreme case was far from perfect.

Any procedure based on randomization inference runs into difficulties when the number of possible re-randomizations is small. For the procedures discussed so far, this number is $G!/(G_1!G_0!)-1$. The problem arises because it is hard to make precise probability statements when comparing an estimate or test statistic with a discrete distribution that has few mass points. MacKinnon and Webb (2019) studied various methods that can be used in this case, including one called wild bootstrap randomization inference, or WBRI. The idea is to generate many bootstrap samples for the actual sample and for each re-randomization. There are two variants. WBRI-β computes a P value based on the position of $\hat{\beta}$ in the sorted list of coefficient estimates, and WBRI-t computes a P value based on the position of the t-statistic for $\beta = 0$ in the sorted list of t-statistics. For an application, see Subsection 6.2.

Several other RI procedures have also been suggested. Canay et al. (2017) proposed a randomization test based on the cluster-level estimators of Ibragimov and Müller (2010). It applies to cases where the number of clusters is fairly small but the number of observations is large. The randomization involves permuting the signs of cluster-level test statistics, which requires a symmetry assumption. Since the statistics can only be computed for clusters that include both treated and untreated observations, the investigator may often need to merge clusters. This test can be substantially more powerful than the t-test proposed in Ibragimov

and Müller (2010), but G cannot be too small. It seems to be advisable to have $G \geq 8$, after clusters have been merged if necessary.

Hagemann (2019a) developed RI tests that can be used even when G is quite small, and Hagemann (2019b) extended them to be valid even when there is substantial heterogeneity across clusters. Unlike the test of Canay et al. (2017), Hagemann's tests do not require cluster-level estimation. However, G_1 and G_0, the numbers of treated and control clusters, should both be no less than 4. Thus, like the tests that involve cluster-level estimation, these tests cannot handle the example given earlier in this section, where $G_1 = 2$ and $G_0 = 13$. Toulis (2019) proposed a broad family of tests based on residual re-randomization. Some of the tests within this family apply to models with both one-way and two-way clustering. Since the one-way test is closely related to the wild cluster bootstrap, it seems likely to have similar properties.

6 Empirical Examples

We now present two empirical examples to highlight some of the issues discussed above, such as having few clusters, having few treated or control clusters, testing the appropriate level of clustering, and/or misspecifying that level. We do not present these examples either to confirm or overturn the results reported originally, but rather to illustrate what can happen in various situations. The analysis makes use of Stata and a few Stata packages, notably the `boottest` package discussed in Roodman et al. (2019).

6.1 Experimental Evidence on Female Voting Behavior

Our first example illustrates the consequences of unbalanced clusters. Giné and Mansuri (2018) examines the effect on women's voting behavior of informing women about the voting process and the importance of voting. The paper estimates a number of regression models. The one that we focus on (for which the paper presents only partial results) is

$$Y_i = \beta_1 T_{1,T} + \beta_2 T_{2,T} + \beta_3 T_{1,U} + \beta_4 T_{2,U} + \boldsymbol{X}_i \boldsymbol{\gamma} + \epsilon_i. \tag{17}$$

Here $T_{j,T}$ for $j = 1, 2$ is an indicator variable for individuals in the treated region who are in target group j, and $T_{j,U}$ is similarly defined for individuals in the untreated region. The parameters of interest are β_1 through β_4. β_1 and β_2 capture the direct effects of each treatment, while β_3 and β_4 capture their spillover effects. The vector $\boldsymbol{X}_i$ contains a constant and 18 control variables, including fixed effects at the village level.

Every observation is associated with one of 67 neighborhoods and one of 9 villages. In the paper, standard errors are clustered by neighborhood, but we also cluster by village. Even though there are 67 neighborhoods, the number of effective clusters G^* (Carter et al. 2017) calculated using `clusteff` is between 14.2 and 15.7, depending on the coefficient for which G^* is calculated. Thus, at the neighborhood level, the clusters are apparently very unbalanced. In contrast, when clustering by village, there are 9 actual clusters and between 8.6 and 8.8 effective clusters, so that the clusters are well balanced. Only 10 of the 67 neighborhoods are controls, which must be partly responsible for the small value of G^* when clustering by neighborhood.

Table 1: Female Voting Example

Variable	Coefficient	Conventional and Bootstrap P values			
		Neighborhood		Village	
		$t(66)$	WCRB	$t(8)$	WCRB
$T_{1,T}$	$\hat{\beta}_1 = -0.0788$	0.0427	0.0916	0.0703	0.0759
$T_{2,T}$	$\hat{\beta}_2 = -0.1161$	0.0039	0.0188	0.0210	0.0113
$T_{1,U}$	$\hat{\beta}_3 = -0.1315$	0.0040	0.0094	0.0217	0.0291
$T_{2,U}$	$\hat{\beta}_4 = -0.0955$	0.0692	0.0949	0.1321	0.1363

Score-Variance Tests

Levels of Clustering	Test Stat.	$\chi^2(10)$ P	Boot P
None vs. Neighborhood	261.917	0.0000	0.0924
None vs. Village	340.280	0.0000	0.0001
Neighborhood vs. Village	61.139	0.0000	0.2719

These results are for the analysis in Panel A of Table 9 of Giné and Mansuri (2018). In the top panel, columns 3 and 5 report P values based on the $t(G-1)$ distribution, and columns 4 and 6 report WCRB P values calculated using 99,999 replications. Although the original paper reported P values equivalent to the ones in column 3, it did not report coefficient estimates. There are 2637 observations, 67 neighborhood clusters, and 9 village clusters. The bootstrap DGPs use Rademacher weights when clustering by neighborhood and Webb (6-point) weights when clustering by village. The bottom panel presents score-variance test statistics for all four coefficients jointly, along with asymptotic and bootstrap P values, the latter based on 9,999 replications.

The top panel of Table 1 shows OLS estimates for both levels of clustering. These differ slightly from the ones reported in Giné and Mansuri (2018), because we estimate (17) using OLS without weighting, while the original paper uses weights proportional to the inverse probability of assignment to treatment. We avoid the use of weights for computational convenience. Although `boottest` can handle any type of weighting supported by Stata, `clusteff` cannot currently do so.

Not surprisingly, the $t(G-1)$ P values in the top panel of Table 1 always increase when we move from neighborhood-level to village-level clustering. Because there are only 9 village clusters, and because the effective number of clusters at the neighborhood level is so much smaller than the actual number, these P values are probably not reliable. This is a case where bootstrapping is essential. With neighborhood-level clustering, the WCRB P values are always larger, often much larger, than the ones based on the $t(66)$ distribution. With village-level clustering, however, the WCRB P values may be either larger or smaller than the $t(8)$ ones. They may also be either larger or smaller than the bootstrap P values based on neighborhood-level clustering. Whatever the level of clustering, $\hat{\beta}_2$ and $\hat{\beta}_3$ appear to be significant at the .05 level, but the other two coefficients do not.

In the bottom panel of Table 1, we test for the level of clustering, jointly for all four coefficients of interest, using the score-variance test proposed in MacKinnon et al. (2020b). We perform three different tests, of no clustering versus neighborhood, no clustering versus village, and neighborhood versus village. All of the test statistics are very large, and the

asymptotic P values based on the $\chi^2(10)$ distribution all equal zero to at least eight digits. However, the bootstrap P values are never exactly zero. In two out of three cases, they do not allow us to reject the null hypothesis at the 5% level. Even for the test of no clustering against village-level clustering, there is apparently one bootstrap sample out of $9,999$ where the bootstrap test statistic is larger than the actual test statistic of 340.28.

Large discrepancies between bootstrap and asymptotic P values for score-variance tests are also observed for the empirical example in MacKinnon et al. (2020b). Like regression (17), that model involves fixed effects at the coarse-cluster level and key regressors which vary only at the fine-cluster level. This suggests that it may be unwise to rely on asymptotic P values for score-variance tests in such models. Since simulation results suggest that bootstrap tests work well for the empirical example in MacKinnon et al. (2020b), we conjecture that the bootstrap P values in Table 1 are also reliable.

The bootstrap score-variance tests in Table 1 make it clear that it would be a serious mistake to assume that there is no clustering. However, they do not provide a clear choice between clustering at the neighborhood level and clustering at the village level. Nevertheless, although we cannot reject neighborhood clustering against village clustering, the fact that no clustering is rejected much more strongly against the latter than against the former suggests that it is probably appropriate to cluster at the village level.

6.2 Public Procurement Auctions in Italy

Our second example is from Branzoli and Decarolis (2015), which exploits a shift in the type of auction used for public-works projects in one Italian city to study how various aspects of the contracts changed for these projects. Specifically, Turin shifted from average-bid to first-price auctions in 2003 while other municipalities did not do so. In their main estimates, the authors estimate a difference-in-differences regression for each of four outcome variables using two different samples, one at the city level and one at the county level. These estimates are found in their Table 2, Panel A.

We focus on just two of these outcome variables for the city-level sample. The outcome variables that we study are the percentage of the value of the contract that was subcontracted (`%-subc`) and whether or not there was a consortium (`Consortium`). We rescaled the latter to equal either 0 or 100 so that the coefficients on the auction-design dummy variable would be of comparable magnitude for each of the two regressions. In their main estimates, the authors cluster by municipality-year. They also perform a robustness check in which they cluster by municipality. In the first case, there are 101 clusters, of which only 4 are treated. In the second case, there are only 15 clusters, of which just 1 is treated. Thus, in this example, the number of treated clusters is either worryingly small or extremely small.

For each of the two outcomes and two levels of clustering, we calculate P values and 95% confidence intervals using the $t(G-1)$ distribution and four different bootstrap methods. In addition to two variants of the wild cluster bootstrap, WCRB and WCUB, we use two variants of the ordinary wild bootstrap, based on restricted (WRB) and unrestricted (WUB) estimates, respectively; see the discussion in Subsection 5.3. The top panel of Table 2 presents the results for `%-subc`, and the bottom panel presents the ones for `Consortium`.

For `%-subc`, there seems to be moderately strong evidence that moving from average-bid to first-price auctions reduces the percentage of contract values that are subcontracted, on

Table 2: Public Procurement Example

Results for `%-subc`, estimate is -9.5661, $N = 1468$				
Method	P value (M-Y)	95% interval (M-Y)	P value (M)	95% interval (M)
$t(G-1)$	0.0000	$[-13.37, -5.77]$	0.0000	$[-11.74, -7.39]$
WCRB	0.0080	$[-15.50, -3.74]$	0.0145	$[-19.33, -2.88]$
WCUB	0.0023	$[-20.77, \ 1.64]$	0.0000	$[-11.51, -7.63]$
WRB	0.0012	$[-14.86, -4.35]$	0.0003	$[-13.88, -5.33]$
WUB	0.0012	$[-20.03, \ 0.90]$	0.0002	$[-18.08, -1.05]$
Results for `Consortium`, estimate is 10.0000, $N = 1461$				
Method	P value (M-Y)	95% interval (M-Y)	P value (M)	95% interval (M)
$t(G-1)$	0.0096	$[\ 2.43, 17.57]$	0.0001	$[\ 6.22, 13.78]$
WCRB	0.1482	$[-3.70, 20.52]$	0.2220	$[-9.54, 24.19]$
WCUB	0.1323	$[-18.07, 38.07]$	0.0000	$[\ 6.13, 13.87]$
WRB	0.0960	$[-2.36, 22.34]$	0.1734	$[-5.27, 25.17]$
WUB	0.0947	$[-14.74, 34.74]$	0.1707	$[-20.55, 40.55]$

These results are for Table 2, Panel A of Branzoli and Decarolis (2015). There are 101 municipality-year (M-Y) clusters and 9 municipality (M) clusters. With clustering by municipality-year, all bootstrap methods use the Rademacher distribution with 99,999 bootstrap samples. With clustering by municipality, WRB and WUB use $B = 99,999$, but WCRB and WCUB enumerate all 32,768 possible bootstrap samples. Note that the estimate reported as 10.0000 is not a mistake; the actual value is 9.9999646.

average. All P values are less than 0.05. However, the length of the confidence intervals varies quite a bit, and the ones for the WCU and WU bootstraps include zero when clustering by municipality-year.

For `Consortium`, the evidence of any effect is much weaker. Although both conventional P values are less than 0.05, all but one of the bootstrap P values is not. The P value of 0.0000 for WCUB with clustering at the municipality level should not be believed, because MacKinnon and Webb (2017a) showed (analytically) that this method usually over-rejects to an extreme extent when there is just one treated cluster.

Because of the small number of treated clusters, it may seem attractive to employ randomization inference, or RI (Subsection 5.4) instead of bootstrap methods. In fact, in a robustness check, Branzoli and Decarolis (2015) uses the procedure of Conley and Taber (2011) to construct RI-based confidence intervals.

If the 101 municipality-year clusters were independent, and treatment were assigned at random to four of them, then randomization inference would probably work well. There would be $101!/(97!\,4!) - 1$, or 4,082,924 possible re-randomizations. We could obtain an RI P value either by comparing the observed coefficient estimate to the empirical distribution of a large number (say, 9,999) of estimates obtained by re-randomization, or by comparing the observed t-statistic to the empirical distribution of a large number of t-statistics obtained in the same way; see MacKinnon and Webb (2020).

In our view, however, it is extremely difficult to justify either clustering by municipality-year or randomization inference that treats municipality-years as independent. Even though all the regressions include year and municipality fixed effects, we would expect there to

be correlation across years for every municipality. But when we attempt to use RI at the municipality level, we run into a serious problem: There are only 14 possible ways in which to re-randomize. Even if the coefficient (or t-statistic) for Turin turned out to more extreme than all the others, that would happen by chance with probability $1/15 = 0.0667$. This would not allow us to reject at the .05 level the null hypothesis that changing the auction rules had no effect.

In fact, the evidence from RI at the municipality level is much weaker than this. In each case, we can rank all 15 estimates, that is, the actual one for Turin and the 14 others based on re-randomization. For `%-subc`, the estimate of -9.5661 is only the third-largest (in absolute value). However, the t-statistic of -9.5143 is the largest one. Thus, for `%-subc`, we can reject the null hypothesis at the .10 level if we rely on t-statistics, but not if we rely on coefficient estimates. For `Consortium`, the estimate of 10.0000 is the second-largest, but it is quite a bit smaller than the largest re-randomized estimate, which is -18.1503. In this case, the t-statistic of 2.6411 is only the fourth-largest. Thus, for `Consortium`, we cannot reject the null hypothesis at any conventional level using either RI method.

For situations in which the number of possible re-randomizations is small, MacKinnon and Webb (2019) proposed a method called wild bootstrap randomization inference, or WBRI; see Subsection 5.4. In this case, when we generate a total of $15 \times 700 = 10,500$ samples, we obtain P values that differ somewhat from the ones in Table 2. For `Consortium`, they are 0.130 (for WBRI-β) and 0.104 (for WBRI-t), so that we cannot reject at the .10 level. For `%-subc`, they are 0.075 (for WBRI-β) and 0.022 (for WBRI-t), so that we may or may not be able to reject at the .05 level.

7 Monte Carlo Simulations

As we stressed in Section 4, choosing the correct level at which to cluster is extremely important. However, we are not aware of any simulation experiments that focus on the consequences of over-clustering and under-clustering. We therefore perform a limited set of Monte Carlo simulations designed to do so. In our experiments, clustering may occur at one of three levels. There are 60 zones, each with 100 observations, which are grouped into 20 cities and 10 states. Seven cities each contain one zone, six cities each contain three zones, and seven cities each contain five zones. Each state contains two cities. Thus states have between 200 and 1000 observations.

The model is

$$y_{sczi} = \beta_1 + \beta_2 x_{sczi} + u_{sczi}, \tag{18}$$

where the regressor x_{sczi} is equal to $x_{1s} \sim \mathrm{N}(0,1)$ plus $x_{2i} \sim \mathrm{N}(0,1)$. As the notation implies, x_{1s} takes the same value for every observation in state s, and the x_{2i} are independent across observations. Thus the correlation of the regressor between any pair of observations in the same state (or zone, or city) is 1/2.

The disturbances may or may not be correlated within zones, cities, and states. For all s, c, z, and i,

$$u_{sczi} = \phi_s v_{1s} + \phi_c v_{2c} + \phi_z v_{3z} + \epsilon_i,$$

where v_{1s}, v_{2c}, v_{3z}, and ϵ_i are distributed as $\mathrm{N}(0,1)$, one of ϕ_s, ϕ_c, and ϕ_z is equal to ϕ, and

Table 3: CRVE Rejection Percentages for Three Levels of Clustering

	Zone $t(59)$			City $t(19)$			State $t(9)$		
ϕ/DGP	zone	city^{u}	state^{u}	zone^{o}	city	state^{u}	zone^{o}	city^{o}	state
0.00	5.24	5.28	5.23	6.26	6.28	6.24	7.17	7.21	7.15
0.05	5.29	8.66	12.98	6.42	6.98	11.22	7.43	8.39	9.08
0.10	5.48	15.65	26.81	6.74	7.71	17.94	8.08	9.80	10.44
0.15	5.50	21.62	36.79	6.98	8.10	22.16	8.52	10.49	10.90
0.20	5.55	25.80	43.02	7.15	8.30	24.60	8.79	10.93	11.18
0.30	5.70	30.34	49.61	7.48	8.47	27.08	9.37	11.26	11.51
0.40	5.65	32.53	52.55	7.51	8.59	28.13	9.48	11.44	11.59
0.50	5.75	33.78	54.19	7.65	8.57	28.67	9.69	11.45	11.67

The table shows rejection percentages for tests at the 5% level based on Monte Carlo experiments with 400,000 replications for standard errors clustered at three levels. Column headings indicate the actual level of clustering in the DGP. An "o" superscript indicates that standard errors are over-clustered, and a "u" superscript indicates that they are under-clustered.

the others are equal to 0. Thus if, for example, there is clustering at the city level, u_{sczi} is equal to ϕ times a city-level random component v_{2c} plus an individual component ϵ_i.

In the experiments, we vary ϕ between 0 and 0.5, and we also vary the level at which the disturbances are clustered. We then test the null hypothesis that $\beta_2 = 0$. Tables 3 and 4 report 5% rejection rates, as percentages, for experiments with 400,000 replications. In Table 3, the rejection rates are based on cluster-robust t-statistics and the $t(G-1)$ distribution. In Table 4, they are based on the WCR bootstrap using the Rademacher distribution and 399 bootstrap replications.

When all correlation is within zones and standard errors are clustered at the zone level, inference based on the $t(G-1)$ distribution is very good, and inference based on the WCRB appears to be perfect; see the first column of results in Tables 3 and 4, respectively. This is not surprising, of course, because 60 is a fairly large number of clusters, and all zones are the same size. The results become much more interesting when there is correlation at the city or state levels.

Both tables show that there is severe over-rejection whenever we under-cluster. This happens in columns 2, 3, and 6, where the DGP is marked with a "u". The over-rejection increases with ϕ, initially very rapidly. The most severe over-rejection occurs in column 3, where the DGP has state-level clustering but standard errors are calculated at the zone level. Thus the standard errors are calculated two levels below the correct one. The over-rejection is less severe, but still very substantial, when the standard errors are calculated at the city level, only one level below the correct one.

The effects of under-clustering necessarily become worse as the sample size increases. We repeated the above experiments with 1000 observations per zone instead of 100. For large values of ϕ, the results did not change very much. For example, the rejection rate for $\phi = 0.50$ in the third column of Table 4 increased from 52.46 to 55.43. For smaller values, however, the rejection frequencies increased much more. For example, the rate for $\phi = 0.10$ in the third column of Table 4 increased from 25.69 to 48.73. Thus the effects of ignoring small amounts of correlation by clustering at too fine a level become more serious as the

Table 4: WCR Bootstrap Rejection Percentages for Three Levels of Clustering

	Zone WCRB			City WCRB			State WCRB		
ϕ/DGP	zone	city^u	state^u	zone^o	city	state^u	zone^o	city^o	state
0.00	4.99	5.02	4.95	5.10	5.11	5.07	5.28	5.38	5.30
0.05	4.98	8.19	12.39	5.12	5.21	8.25	5.35	5.50	5.62
0.10	5.03	14.79	25.69	5.21	5.27	12.51	5.46	5.71	5.85
0.15	5.00	20.44	35.50	5.15	5.31	15.21	5.57	5.83	5.88
0.20	4.98	24.35	41.48	5.16	5.31	16.83	5.60	5.93	5.97
0.30	5.03	28.66	48.00	5.26	5.37	18.52	5.76	6.05	6.02
0.40	4.98	30.77	50.84	5.21	5.40	19.32	5.74	6.06	6.13
0.50	5.02	31.95	52.46	5.27	5.35	19.80	5.84	6.01	6.17

See notes to Table 3. The WCRB uses the Rademacher distribution with 399 bootstrap replications.

sample size increases, just as the analysis of Section 3 suggested.

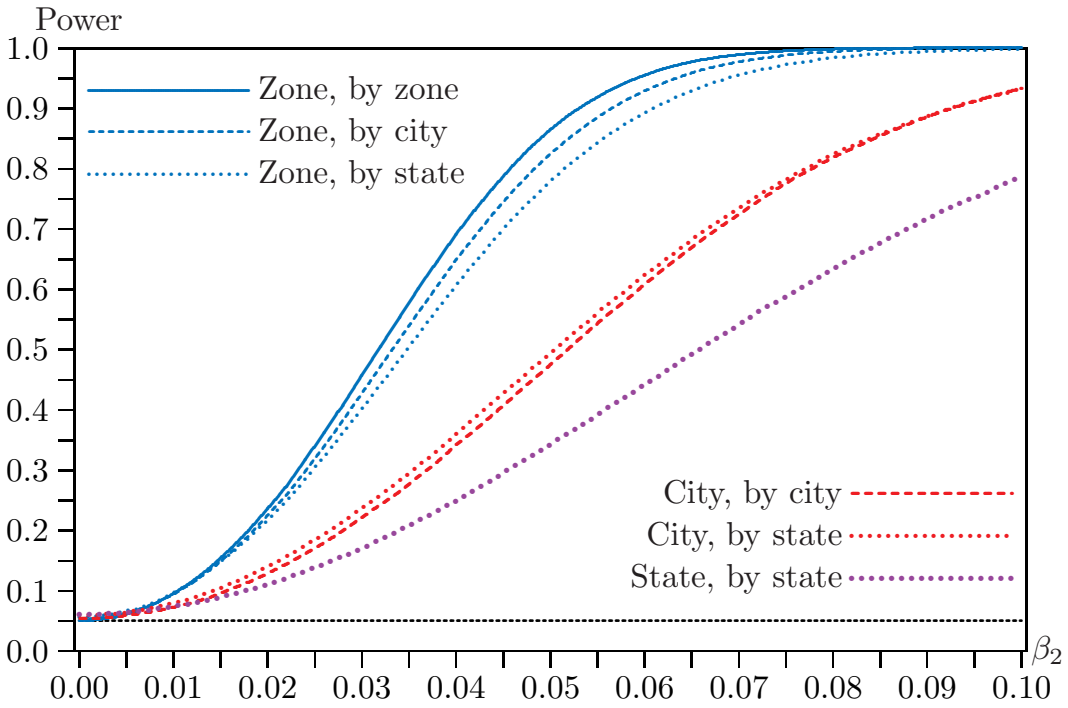
The consequences of over-clustering are much less severe than those of under-clustering, especially when the WCR bootstrap is used. Inference based on the $t(G-1)$ distribution is somewhat unreliable when clustering by city and quite unreliable when clustering by state. In contrast, over-clustering causes very little size distortion when inference is based on the WCRB. This may be seen in columns 4, 7, and 8 of Table 4, where the DGP is marked with an "o". Even though there are only 10 clusters, and they vary considerably in size, the combination of state-level clustering and the WCRB always works quite well, with rejection rates never much greater than 6%.

The results in Table 4 suggest that, provided we use the WCR bootstrap, the cost of over-clustering is quite modest when the null hypothesis is true. But what if that null hypothesis is false? Figure 1 shows six power functions for tests of $\beta_2 = 0$ in the model (18). For the three blue curves, clustering is actually at the zone level. The solid curve shows power when the CRVE correctly clusters at the zone level, the dashed curve shows power when it over-clusters at the city level, and the dotted curve shows power when it over-clusters at the state level. There is evidently some power loss due to over-clustering. This is modest when we cluster by city, but it is about twice as large when we cluster by state.

For the two red curves, clustering is actually at the city level. Power is very much less than it was with clustering at the zone level, because there are far more non-zero off-diagonal elements in the $\mathbf{\Omega}$ matrix. Oddly, except for very large values of β_2, power seems to be slightly higher when the CRVE over-clusters at the state level than when it correctly clusters at the city level. This apparent gain in power is spurious, of course. It arises because the test over-rejects a bit more in the former case than in the latter. The figure does not attempt to show size-adjusted power because there is no way to size-adjust these tests in practice; see Davidson and MacKinnon (2006b).

For the purple curve, clustering is actually at the state level. The only valid way for the CRVE to cluster is also at the state level, so only one curve is shown. The power loss, relative to clustering at the city level, is quite severe. Once again, this is not at all surprising. With only 10 states, there are a great many non-zero off-diagonal elements in the $\mathbf{\Omega}$ matrix. Thus the sample contains much less information than it did with clustering at a finer level.

Figure 1: Power of WCR Bootstrap Tests



Notes: In these experiments, $N = 6000$ and $\phi = 0.2$. The WCRB uses the Rademacher distribution with 399 bootstrap replications. The label "X, by y" means that the disturbances are actually clustered at level X, and the CRVE is clustered at level y.

8 Conclusion

Disturbances (error terms) that are correlated within clusters can cause severe problems for inference, especially when the sample size is large and the number of clusters is not. The standard approach is to use t-statistics based on cluster-robust standard errors from a CRVE, usually CV_1 given in (4), together with the $t(G-1)$ distribution. This approach generally works well when there is a large number of clusters that are roughly balanced in terms of cluster sizes and the features of the key regressors and the disturbances. However, when the number of clusters is small, or the clusters are seriously unbalanced, the standard approach can yield very unreliable inferences.

The restricted wild cluster bootstrap (WCRB) works well in a far wider set of circumstances than the standard approach. Accordingly, we advocate using it as the default method of statistical inference. Computational tricks pioneered in the `boottest` package for Stata make computing the WCRB very fast and easy in most cases. There are still situations in which the WCRB will not be particularly reliable, most notably when there are very few clusters (say, six or less), when clusters are very heterogeneous, or when there are very few treated (or control) clusters. In such cases, it is important to check whether the WCRB seems to be reliable. One approach is to compare WCRB P values or confidence intervals with those from alternative procedures such as randomization inference (Subsection 5.4) or

the procedures discussed near the end of Subsection 5.1. If the results of the WCRB broadly agree with those from other procedures, then it seems reasonable to accept the former. A second approach is to conduct a Monte Carlo experiment that mimics the model and dataset on hand. This approach can be particularly useful in unusual settings where the clusters are severely unbalanced in size or in some other way.

The validity of the asymptotic approximations that underlie cluster-robust inference is driven by the number of clusters rather than the sample size. Unfortunately, and contrary to popular belief, there is no "golden number" of clusters beyond which CRVE-based inference becomes reliable. The requisite number of clusters depends on many factors. These include how much the cluster sizes vary, how unbalanced the clusters are in other respects, and, in many cases, how many clusters are treated, how many clusters are not treated, and the numbers of treated observations in each of the treated clusters. Because it conditions on all these aspects of the sample, which no "golden number" could ever do, our recommendation is to use the restricted wild cluster bootstrap essentially all the time.

In Section 4, we suggested a few guidelines for how to cluster. In general, one should cluster at the coarsest level possible. The simulation experiments in Section 7 suggest that there can be large size distortions from under-clustering compared with much smaller power losses from over-clustering. This is especially true in large samples. When studying the effects of policy changes, one should always cluster at least at the level of the policy change, and perhaps at a more aggregate level. When working with panel data, it is desirable to cluster by the cross-section dimension (perhaps in addition to other dimensions) in order to capture any serial correlation.

Much more is known about cluster-robust inference than was the case even ten years ago. Nevertheless, there are still many unanswered questions. We do not know how to control for both spatial correlation and conventional within-cluster correlation at the same time. We do not know how to obtain reliable inferences when there are few treated clusters and the treated clusters are atypical. Although testing procedures exist, we do not really know how to determine the right level at which to cluster, especially when over-clustering means having few clusters or few treated clusters. Finally, except in certain special cases, we do not know how to conduct inference when there is a very small number of clusters, say, six or less.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? NBER Working Papers 24003, National Bureau of Economic Research, Inc.

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2020). Sampling-based vs. design-based uncertainty in regression analysis. *Econometrica 88*, 265 – 296.

Abadie, A., A. Diamond, and J. 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.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review 93*(1), 112–132.

Andrews, D. W. K. (2005). Cross-section regression with common shocks. *Econometrica 73*(5), 1551–1585.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion* (1 ed.). Princeton University Press.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics 49*(4), 431–434.

Barrios, T., R. Diamond, G. W. Imbens, and M. Kolesár (2012). Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association 107*(498), 578–591.

Bell, R. M. and D. F. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology 28*(2), 169–181.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*(1), 137–151.

Branzoli, N. and F. Decarolis (2015). Entry and subcontracting in public procurement auctions. *Management Science 61*(12), 2945–2962.

Brewer, M., T. F. Crossley, and R. Joyce (2018). Inference with difference-in-differences revisited. *Journal of Econometric Methods 7*(1), 1–16.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics 29*(2), 238–249.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica 85*(3), 1013–1030.

Canay, I. A., A. Santos, and A. Shaikh (2020). The wild bootstrap with a 'small' number of 'large' clusters. *The Review of Economics and Statistics 102*, to appear.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017). Asymptotic behavior of a t-test robust to cluster heterogeneity. *The Review of Economics and Statistics 99*(4), 698–709.

Conley, T. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics 92*(1), 1–45.

Conley, T. G., S. Gonçalves, and C. B. Hansen (2018). Inference with dependent data in accounting and finance applications. *Journal of Accounting Research 56*(4), 1139–1203.

Conley, T. G. and C. R. Taber (2011). Inference with "Difference in Differences" with a small number of policy changes. *The Review of Economics and Statistics 93*(1), 113–125.

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2020). Empirical process results for exchangeable arrays. ArXiv e-prints, CREST.

Davidson, R. and E. Flachaire (2008). The wild bootstrap, tamed at last. *Journal of Econometrics 146*(1), 162–169.

Davidson, R. and J. G. MacKinnon (2006a). Bootstrap methods in econometrics. In T. C. Mills and K. D. Patterson (Eds.), *Palgrave Handbook of Econometrics: Volume 1 Econometric Theory*, pp. 812–838. Palgrave Macmillan.

Davidson, R. and J. G. MacKinnon (2006b). The power of bootstrap and asymptotic tests. *Journal of Econometrics 133*(2), 421–441.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*(2), 393–412.

Donald, S. G. and K. Lang (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics 89*(2), 221–233.

Esarey, J. (2018). clusterSEs: Calculate cluster-robust p-values and confidence intervals. Technical report.

Esarey, J. and A. Menger (2019). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods 7*(3), 541–559.

Ferman, B. (2019). Inference in differences-in-differences: How much should we trust in independent clusters? MPRA Paper 93746, University Library of Munich, Germany.

Ferman, B. and C. Pinto (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *The Review of Economics and Statistics 101*, 452–467.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Giné, X. and G. Mansuri (2018). Together we will: Experimental evidence on female voting behavior in Pakistan. *American Economic Journal: Applied Economics 10*(1), 207–35.

Hagemann, A. (2019a). Placebo inference on treatment effects when the number of clusters is small. *Journal of Econometrics 213*(1), 190–209.

Hagemann, A. (2019b). Permutation inference with a finite number of heterogeneous clusters. ArXiv e-prints 1907.01049 [econ.EM].

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of Econometrics 210*(2), 268–290.

Hess, S. (2017). Randomization inference with Stata: A guide and software. *Stata Journal 17*(3), 630–651.

Horowitz, J. L. (2019). Bootstrap methods in econometrics. *Annual Review of Economics 11*(1), 193–224.

Ibragimov, R. and U. K. Müller (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics 28*(4), 453–468.

Ibragimov, R. and U. K. Müller (2016). Inference with few heterogeneous clusters. *The Review of Economics and Statistics 98*(1), 83–96.

Imbens, G. W. and M. Kolesár (2016). Robust standard errors in small samples: Some practical advice. *The Review of Economics and Statistics 98*(4), 701–712.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical*

Sciences. New York: Cambridge University Press.

Jackson, J. E. (2020). Corrected standard errors with clustered data. *Political Analysis 28*, to appear.

Kelly, M. (2019, 06). The standard errors of persistence. Technical report.

Kloek, T. (1981). Ols estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica 49*(1), 205–207.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *The Stata Journal 10*(2), 165–199.

Lee, C. H. and D. G. Steigerwald (2018). Inference for clustered data. *Stata Journal 18*(2), 447–460.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (Third ed.). New York: Springer.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics 35*(4), 615–645.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–461. Springer.

MacKinnon, J. G. (2015). Wild cluster bootstrap confidence intervals. *L'Actualité Économique 91*, 11–33.

MacKinnon, J. G. (2016). Inference with large clustered datasets. *L'Actualité Économique 92*, 649–665.

MacKinnon, J. G. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics 52*(3), 851–881.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2020a). Wild bootstrap and asymptotic inference with multiway clustering. *Journal of Business & Economic Statistics 38*, to appear.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2020b). Testing for the appropriate level of clustering in linear regression models. QED Working Paper 1428, Queen's University, Department of Economics.

MacKinnon, J. G. and M. D. Webb (2017a). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics 32*(2), 233–254.

MacKinnon, J. G. and M. D. Webb (2017b). Pitfalls when estimating treatment effects using clustered data. *The Political Methodologist 24*(2), 20–31.

MacKinnon, J. G. and M. D. Webb (2018). The wild bootstrap for few (treated) clusters. *Econometrics Journal 21*(2), 114–135.

MacKinnon, J. G. and M. D. Webb (2019). Wild bootstrap randomization inference for few treated clusters. In K. P. Huynh, D. T. Jacho-Chávez, and G. Tripathi (Eds.), *The Econometrics of Complex Survey Data: Theory and Applications*, Volume 39 of *Advances*

in Econometrics, Chapter 3, pp. 61–85. Emerald Group.

MacKinnon, J. G. and M. D. Webb (2020). Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics* (to appear).

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*(3), 305–325.

Menzel, K. (2018). Bootstrap with cluster-dependence in two or more dimensions. ArXiv e-prints, New York University.

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, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics 32*(3), 385–397.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics 72*(2), 334–338.

Pustejovsky, J. (2017). clubsandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. Technical report.

Racine, J. S. and J. G. MacKinnon (2007). Simulation-based tests that can use any number of simulations. *Communications in Statistics–Simulation and Computation 36*, 357–365.

Riddell, W. C. (1979). The empirical foundations of the Phillips curve: Evidence from Canadian wage contract data. *Econometrica 47*(1), 1–24.

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin 13*, 19–23.

Roodman, D., J. G. MacKinnon, M. Ø. Nielsen, and M. D. Webb (2019). Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal 19*(1), 4–60.

Spamann, H. (2019). On inference when using state corporate laws for identification. Discussion Paper 1024, Harvard Law School.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics 99*, 1–10.

Toulis, P. (2019). Life after bootstrap: Residual randomization inference in regression models. Technical report, University of Chicago.

Webb, M. D. (2014). Reworking wild bootstrap based inference for clustered errors. QED Working Paper 1315, Queen's University, Department of Economics.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica 48*(4), 817–838.

Young, A. (2016). Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections. Working paper, London School of Economics.

Young, A. (2019). Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics 134*(2), 557–598.