MacKinnon, James G.; Nielsen, Morten Ørregaard; Webb, Matthew

Working Paper

Cluster-robust inference: A guide to empirical practice

Queen's Economics Department Working Paper, No. 1456

Provided in Cooperation with:
Queen's University, Department of Economics (QED)

Suggested Citation: MacKinnon, James G.; Nielsen, Morten Ørregaard; Webb, Matthew (2021) : Cluster-robust inference: A guide to empirical practice, Queen's Economics Department Working Paper, No. 1456, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/247198

Queen's Economics Department Working Paper No. 1456

Cluster-Robust Inference: A Guide to Empirical Practice

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Queen's University
and CREATES

Matthew D. Webb
Carleton University

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

5-2021

Cluster-Robust Inference: A Guide to Empirical Practice*

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

Morten Ørregaard Nielsen
Queen's University and CREATES
mon@econ.queensu.ca

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

May 18, 2021

Abstract

Methods for cluster-robust inference are routinely used in economics and many other disciplines. However, it is only recently that theoretical foundations for the use of these methods in many empirically relevant situations have been developed. In this paper, we use these theoretical results to provide a guide to empirical practice. We do not attempt to present a comprehensive survey of the (very large) literature. Instead, we bridge theory and practice by providing a thorough guide on what to do and why, based on recently available econometric theory and simulation evidence. The paper includes an empirical analysis of the effects of the minimum wage on teenagers using individual data, in which we practice what we preach.

Keywords: clustered data, grouped data, cluster-robust variance estimator, CRVE, robust inference, wild cluster bootstrap.

JEL Codes: C12, C15, C21, C23.

*We are grateful to Silvia Gonçalves for helpful discussions and to Serena Ng for suggesting that we write this paper. MacKinnon and Webb thank the Social Sciences and Humanities Research Council of Canada (SSHRC, grant 435-2016-0871) for financial support. Nielsen thanks the Canada Research Chairs program and the SSHRC (grant 435-2017-0131).

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

1 Introduction

Ideally, the observations in a sample would be independent of each other and would each contribute roughly the same amount of information about the parameter(s) of interest. From the earliest days of econometrics, it has been recognized that this ideal situation often does not apply to time-series data. But it has taken much longer for econometricians to realize that it generally does not apply to cross-section data either.

There are many ways in which cross-section data might be dependent, and sometimes it is possible to model this dependence explicitly. For example, there is a large literature on spatial econometrics and statistics, in which each observation is associated with a point in space, and the correlation between any two observations is assumed to depend (usually in a rather simple parametric way) on the distance between them. See, among many others, Anselin (1988), Gelfand, Diggle, Fuentes and Guttorp (2010), and Corrado and Fingleton (2012). However, there are a great many cases in which either the "distance" between any pair of observations cannot be measured, or the correlation between them is not related to distance in any way that can readily be modeled.

A more widely applicable approach, on which we focus in this paper, is to employ cluster-robust inference. This approach has become increasingly popular over the past quarter century and is now used routinely in a great deal of empirical microeconomic work. The idea is to divide the sample into G disjoint clusters. Depending on the nature of the data, the clusters might correspond to classrooms, schools, families, villages, hospitals, firms, industries, years, cities, counties, states, or countries. This list is by no means exhaustive. Any pattern of heteroskedasticity and/or dependence is allowed within each cluster, but it is assumed that the assignment of observations to clusters is known and that there is independence across clusters.

Under these assumptions, it is easy to compute cluster-robust standard errors that can be used to produce asymptotically valid inferences; see Section 2. However, these inferences may not be at all reliable in finite samples. Hypothesis tests may reject far more (or far less) often than they should, and the actual coverage of confidence intervals may differ greatly from their nominal coverage. In consequence, using cluster-robust inference in practice often requires a good deal of care.

There are several recent survey papers on cluster-robust inference, including Cameron and Miller (2015), MacKinnon (2019), Esarey and Menger (2019), and MacKinnon and Webb (2020b). Conley, Gonçalves and Hansen (2018) surveys a broader class of methods for various types of dependent data. Although there will inevitably be some overlap with these papers, our aim is to provide a guide to empirical practice rather than a survey of

the extant literature. We therefore apologize for any missing references and refer the reader to the survey papers just mentioned for more complete lists of references. Our guide is closely based on the econometric theory and simulation evidence that is currently available. When the theory is clear and the evidence is strong, we make definitive recommendations for empirical practice. However, when the theory is less clear or the evidence is weak, our recommendations are more guarded.

This guide does not discuss regression models with clustered data estimated by instrumental variables (IV). For such models, neither the current state of econometric theory nor the available simulation evidence allows us to make recommendations with any confidence. The number of over-identifying restrictions and the strength of the instruments can greatly affect the reliability of finite-sample IV inference, and dealing with these issues can be even more important than dealing with the finite-sample issues associated with clustering. There is an enormous literature on the topic of weak instruments; see Andrews, Stock and Sun (2019) for a recent survey. That paper suggests that, when the disturbances of a regression model are independent and homoskedastic, it is generally possible to obtain reliable (although perhaps imprecise) inferences even when the instruments are quite weak. However, it also states that this is not the case, in general, when there is heteroskedasticity and/or clustering.

In Section 2, we obtain the variance matrix for the coefficient estimates in a linear regression model with clustered data, along with a cluster-robust variance estimator, or CRVE. Our discussion focuses on the properties of the score vectors for each cluster. Section 3 deals with the important and sometimes controversial issue of when to use cluster-robust inference. It also illustrates how complicated patterns of intra-cluster correlation can arise in the context of a simple factor model. In Section 4, we explain how to obtain asymptotically valid inferences and discuss what determines how reliable, or unreliable, they are likely to be in practice. We also discuss in some detail what an empirical investigator should report in order to convince the reader that their results are reliable.

In Section 5, we describe two methods for bootstrap inference. The pairs cluster bootstrap is very widely applicable, but in many cases the wild cluster bootstrap, which only applies to regression models, is more likely to perform well. In Section 6, we discuss some related inferential procedures. The first of these uses an alternative CRVE, often combined with an alternative critical value estimated from the data, and the second is randomization inference. Section 7 discusses how to choose the correct level at which to cluster, Section 8 presents an empirical example that uses individual data to study the effects of the minimum wage on the labor supply of teenagers, and Section 9 provides a summary guide for empirical practice.

2 Cluster-Robust Variance Estimators

Consider the linear regression model

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

where the data have been divided into G disjoint clusters. Here $\boldsymbol{X}_g$ is an $N_g \times k$ matrix of (for simplicity) exogenous regressors, $\boldsymbol{\beta}$ is a k-vector of coefficients, $\boldsymbol{y}_g$ is an N_g-vector of observations on the regressand, and $\boldsymbol{u}_g$ is an N_g-vector of disturbances (or error terms). Since the g^{th} cluster has N_g observations, the sample size is $N = \sum_{g=1}^{G} N_g$. The $\boldsymbol{X}_g$ may of course be stacked into an $N \times k$ matrix $\boldsymbol{X}$, and likewise the $\boldsymbol{y}_g$ and $\boldsymbol{u}_g$ may be stacked into N-vectors $\boldsymbol{y}$ and $\boldsymbol{u}$, so that (1) can be rewritten as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$.

Under the assumption that the data are actually generated by (1) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}.$$

It follows that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{u}_g = \Big(\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g\Big)^{-1}\sum_{g=1}^{G}\boldsymbol{s}_g, \tag{2}$$

where $\boldsymbol{s}_g = \boldsymbol{X}_g^\top\boldsymbol{u}_g$ denotes the $k \times 1$ score vector corresponding to the g^{th} cluster. For a correctly specified model, $\mathrm{E}(\boldsymbol{s}_g) = \boldsymbol{0}$ for all g. From the rightmost expression in (2), the distribution of the OLS estimator $\hat{\boldsymbol{\beta}}$ depends on $\boldsymbol{u}$ only through the distribution of the score vectors $\boldsymbol{s}_g$. Ideally, the sum of the $\boldsymbol{s}_g$, suitably normalized, would be well approximated by a multivariate normal distribution with mean zero. Asymptotic inference uses the empirical score vectors $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g$, in which the disturbance subvectors $\boldsymbol{u}_g$ are replaced by the residual subvectors $\hat{\boldsymbol{u}}_g$, to estimate the variance matrix of the $\boldsymbol{s}_g$.

Because we can always divide the sample into G clusters in any way we like, (2) is true for any distribution of the disturbance vector $\boldsymbol{u}$. Dividing the sample into clusters only becomes meaningful if we further assume that

$$\mathrm{E}(\boldsymbol{s}_g\boldsymbol{s}_g^\top) = \boldsymbol{\Sigma}_g \quad \text{and} \quad \mathrm{E}(\boldsymbol{s}_g\boldsymbol{s}_{g'}^\top) = \boldsymbol{0}, \quad g, g' = 1, \ldots, G, \quad g' \neq g, \tag{3}$$

where the variance matrix for the g^{th} cluster, $\boldsymbol{\Sigma}_g$, is a $k \times k$ symmetric, positive semidefinite matrix. The second assumption in (3) is the key one. It states that the scores for every cluster are uncorrelated with the scores for every other cluster. In contrast, the first assumption imposes no real limitations on the variance matrix of the scores for each cluster. For now, we will simply assume that (3) holds for some specified division of the observations into clusters.

Although it is often controversial, or at least somewhat debatable, this assumption is almost always made in the literature. The important issue of how to choose the clustering structure, perhaps by testing for the correct level of clustering, will be discussed in detail in Section 7.

It follows immediately from (2) that an estimator of the variance of $\hat{\boldsymbol{\beta}}$ should be based on the usual sandwich formula,

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\Big(\sum_{g=1}^{G} \boldsymbol{\Sigma}_g\Big)(\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{4}$$

The natural way to estimate (4) is to replace the $\boldsymbol{\Sigma}_g$ matrices by their empirical counterparts. If, in addition, we multiply by a correction for degrees of freedom, we obtain the cluster-robust variance estimator, or CRVE,

$$\text{CV}_1\text{:} \qquad \frac{G(N-1)}{(G-1)(N-k)}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\Big(\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top\Big)(\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{5}$$

This is by far the most widely used CRVE in practice, but there are others; see Section 6.1. Observe that, when $G = N$, CV_1 reduces to the familiar HC_1 estimator (MacKinnon and White 1985) that is robust only to heteroskedasticity of unknown form.

In (3), we made assumptions directly about the score vectors. Sometimes it is more illuminating to make assumptions about the disturbances. If $\text{E}(\boldsymbol{u}_g \boldsymbol{u}_{g'}^\top | \boldsymbol{X}) = \boldsymbol{0}$ for all $g' \neq g$, then the second assumption in (3) will hold. It will also hold if the regressors are exogenous and uncorrelated across clusters even when the disturbances are not. Since the score vector $\boldsymbol{s}_g$ can be written as $\sum_{i=1}^{N_g} \boldsymbol{s}_{gi} = \sum_{i=1}^{N_g} \boldsymbol{X}_{gi}^\top u_{gi}$, where $\boldsymbol{X}_{gi}$ is the i^{th} row of $\boldsymbol{X}_g$ and u_{gi} is the i^{th} element of $\boldsymbol{u}_g$, the outer product of the score vector with itself is seen to be

$$\boldsymbol{s}_g \boldsymbol{s}_g^\top = \Big(\sum_{i=1}^{N_g} \boldsymbol{X}_{gi}^\top u_{gi}\Big)\Big(\sum_{i=1}^{N_g} \boldsymbol{X}_{gi}^\top u_{gi}\Big)^\top = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \boldsymbol{X}_{gi}^\top \boldsymbol{X}_{gj} u_{gi} u_{gj} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \boldsymbol{s}_{gi} \boldsymbol{s}_{gj}^\top. \tag{6}$$

When $\text{E}(u_{gi}^2 | \boldsymbol{X}) = \sigma^2$ and $\text{E}(u_{gi} u_{gj} | \boldsymbol{X}) = 0$ for $i \neq j$, then $\text{E}(\boldsymbol{s}_g \boldsymbol{s}_g^\top | \boldsymbol{X}) = \sigma^2 \boldsymbol{X}_g^\top \boldsymbol{X}_g$. In that case, we would replace $\boldsymbol{\Sigma}_g$ with $\sigma^2(\boldsymbol{X}_g^\top \boldsymbol{X}_g)$ in (4) and obtain the classic results that $\text{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 | \boldsymbol{X}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

When the scores are uncorrelated within each cluster, the expectation of $\boldsymbol{s}_{gi} \boldsymbol{s}_{gj}^\top$ is a zero matrix for $i \neq j$. This can happen when either the disturbances or the regressors are uncorrelated. In that case, it holds that $\text{E}(\boldsymbol{s}_g \boldsymbol{s}_g^\top) = \sum_{i=1}^{N_g} \text{E}(\boldsymbol{s}_{gi} \boldsymbol{s}_{gi}^\top)$. In general, the difference between the expectation of the rightmost expression in (6) and this expression is

$$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \text{E}\Big(\boldsymbol{s}_{gi} \boldsymbol{s}_{gj}^\top\Big) - \sum_{i=1}^{N_g} \text{E}\Big(\boldsymbol{s}_{gi} \boldsymbol{s}_{gi}^\top\Big) = \sum_{i=1}^{N_g}\sum_{j\neq i} \text{E}\Big(\boldsymbol{s}_{gi} \boldsymbol{s}_{gj}^\top\Big). \tag{7}$$

The right-hand side of (7), which equals zero whenever there is no intra-cluster correlation, involves a sum over $N_g^2 - N_g$ terms. Therefore, incorrectly assuming that the scores are not correlated within clusters potentially leads to much larger errors of inference when clusters are large than when they are small. For sufficiently large values of N_g, these errors may be large even when all of the $\mathrm{E}(\boldsymbol{s}_{gi}\boldsymbol{s}_{gj}^\top)$ for $i \neq j$ are very small.

The famous "Moulton factor" (Moulton 1986) gives the ratio of the true variance of an OLS coefficient, from (4), to the variance based on the classic formula $\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ under the assumption that both the disturbances and the regressor of interest (after other regressors have been partialed out) are equi-correlated within clusters. If the scores were scalars with intra-cluster correlation ρ_s, and the cluster sizes were constant, say $N_g = M$, the Moulton factor would be $1+(M-1)\rho_s$. The second term is proportional to the number of observations per cluster, so the mistakes made by not clustering can be enormous when clusters are large.

3 Why Should We Cluster

We cannot obtain reliable inferences when using clustered data unless we know the actual clustering structure. How to choose the correct level at which to cluster will be discussed in Section 7. Until then, we simply assume that the clustering structure is known, in the sense that assumptions (3) hold.

3.1 Design-Based vs. Model-Based Approach

To discuss whether clustering is needed, Abadie, Athey, Imbens and Wooldridge (2017) distinguish between a "design-based" approach and a "model-based" approach. Following the latter, we assume that every sample is a random outcome, or drawing, from some meta-population. Such a meta-population may be characterized by a data-generating process (DGP), which consists of the model (1) accompanied by a procedure for generating $\boldsymbol{X}_g$ matrices and $\boldsymbol{u}_g$ vectors. The DGP might, for example, generate the $\boldsymbol{X}_g$ and the $\boldsymbol{u}_g$ cluster by cluster from joint distributions at the cluster level. The coefficients of interest are then interpreted as features of that meta-population.

In contrast, Abadie et al. (2017) develop a "design-based" approach in which the investigator is concerned with the characteristics of a finite sample from the meta-population. They call this finite sample a "finite population" and assume that the observed sample constitutes a substantial proportion of it. Under these assumptions, Abadie et al. (2017) show that the correct way to make inferences depends on how treatment was assigned. Unless treatment was assigned at the cluster level, in which case we should cluster at that level,

it may be appropriate to use heteroskedasticity-robust standard errors rather than cluster-robust ones, even in cases where the latter are substantially larger than the former. They also argue that if the sample is "cluster-randomized," so that only some clusters are present within the sample, then we should cluster at the level at which this sampling occurred.

The model-based approach seems entirely appropriate when clusters correspond to objects such as classrooms, schools, hospitals, families, and villages, of which there are potentially a great many. These are also settings in which sampling is likely to occur at the cluster level. Whether it is appropriate for objects such as states, provinces, or countries is perhaps less clear. Consider, for example, the fifty states of the United States. On the one hand, if the investigator believes that state-level data for the United States are random outcomes, then the model-based approach is appropriate. On the other hand, if the investigator believes that the state-level data are the *non-random* populations of the fifty states, and interest is only in descriptive features of those non-random populations, then the design-based approach seems appropriate.

3.2 Placebo Regressions

The validity of alternative standard errors can be assessed by running "placebo regressions." The idea, first suggested in Bertrand, Duflo and Mullainathan (2004), is to start with a model and dataset, then generate a completely artificial regressor at random, add it to the model, and perform a t-test of significance. This is repeated a large number of times, and the rejection frequency is observed. The artificial regressor is often a dummy variable and referred to as a "placebo law" or "placebo treatment." Using a dummy variable is natural because, for any level of intra-cluster correlation of the disturbances, the intra-cluster correlation of the scores is greatest for regressors that do not vary within clusters. However, any artificial regressor that is not completely uncorrelated within clusters can potentially be used. Because any placebo-regression experiment is conditional on just one dataset, the results do not depend on whether the design-based or model-based story is appropriate.

Since a placebo regressor is artificial, we would expect valid significance tests at level α to reject the null close to $\alpha\%$ of the time when the experiment is repeated many times. Using models for log-earnings based on age, education, and other personal characteristics, together with data taken from the Current Population Survey, several papers (Bertrand et al. 2004; MacKinnon 2016; MacKinnon and Webb 2017a; Brewer et al. 2018) find that not clustering, or clustering at below the state level, leads to rejection rates far greater than α. In Section 8.2, we find similar results for the datasets used in our empirical example. Our findings and those of the papers cited above all suggest that, if we fail to use a state-level

CRVE for survey data that samples individuals from multiple states, we will find, with probability much higher than α, that nonsense regressors apparently belong in the model. This seems to be incompatible with the design-based approach, which sometimes (but not always) tells us not to use a CRVE even when doing so leads to larger standard errors.

The empirical score vectors are $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g$ for clusters $g = 1, \ldots, G$, so a placebo-regressor experiment should lead to over-rejection whenever both the regressor and the residuals display intra-cluster correlation at a level higher than the one at which the standard errors are clustered. For example, suppose there are two potential levels of clustering, fine and coarse, where the fine clusters are nested within the coarse clusters. If the placebo regressor is clustered at the coarse level, we would expect to see over-rejection based on heteroskedasticity-robust standard errors whenever the residuals are clustered at either level, and to see over-rejection based on finely clustered standard errors whenever the residuals are clustered at the coarse level; see Section 8.2 for an example of this.

3.3 Sources of Intra-Cluster Dependence

In principle, intra-cluster correlation could arise in many ways. To fix ideas, consider the following simple, and very standard, factor model that generates intra-cluster dependence of the type assumed in (3). Suppose the disturbance u_{gi} for observation $i = 1, \ldots, N_g$ in cluster $g = 1, \ldots, G$ is generated according to

$$u_{gi} = \lambda_{gi} \varepsilon_g + \varepsilon_{gi}, \tag{8}$$

where $\varepsilon_{gi} \sim \text{iid}(0, \omega^2)$ is an idiosyncratic shock for observation i, $\varepsilon_g \sim \text{iid}(0, 1)$ is a cluster-wide shock for cluster g, i.e. the factor effect, and λ_{gi} is the "factor loading" or weight that determines the extent to which observation i is affected by the cluster-wide shock. We assume that the loadings are non-random, but they could also be random variables. All quantities in cluster g are assumed to be independent of quantities in cluster g' for $g \neq g'$. We note that the random-effects model is the special case in which $\lambda_{gi} = \lambda_g$ is fixed across all i.

As an example, if the clusters denoted classrooms and the outcome were student achievement, then ε_{gi} would be unobserved student-specific characteristics, ε_g would be unobserved teacher input, and λ_{gi} would measure the extent to which the disturbance term for student i is affected by the teacher input. Clearly, the λ_{gi} do not need to be the same for all i. Similar motivating examples based on (8) can easily be given in many fields, including labor economics, health economics, development economics, and financial economics.

To verify that the factor model in (8) generates dependence within the clusters, it suffices to derive the second-order moments of the u_{gi}. We find that $\text{E}(u_{gi}) = 0$ and $\text{Var}(u_{gi}) =$

$\lambda_{gi}^2 + \omega^2$. The cluster dependence is characterized by $\text{Cov}(u_{gi}, u_{gj}) = \lambda_{gi}\lambda_{gj}$, which is zero only when the factor loadings are zero. In the context of the classroom example, the intra-cluster covariances would be zero only if the teacher had no effect on student achievement.

The factor model in (8) is discussed in terms of the disturbances. There are at least two simple cases in which the same model structure, and in particular the same within-cluster correlation structure, applies to the scores. The first is when a regressor is generated by a model similar to (8), but possibly with different parameters. The second is when a regressor only varies at the cluster level, as is often the case for dummy variables.

The model (8) has only one clustering dimension, but it does not apply only to cross-section data. For example, if the observations also had a time dimension, we could replace each of the ε_g by a time-series process at the cluster level. This would yield a pattern of intra-cluster dependency where the correlations within each cross-sectional unit diminish as the observations become further apart in time. In Section 8.2, we generate placebo regressors in this way. For panel data, there is another possibility, which is discussed in Section 3.5. In addition to correlation within cross-sectional units across time periods, there may be correlation within time periods across cross-sectional units.

3.4 Are Cluster Fixed Effects Sufficient?

It is often argued that including cluster fixed effects removes any within-cluster dependence and hence eliminates the need to use a CRVE. However, as Arellano (1987) pointed out, that is in fact only true under very special circumstances. Consider the factor model (8). Including cluster fixed effects will force the intra-cluster sample average to be zero for each cluster. That is, including cluster fixed effects would transform the model into

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g)\varepsilon_g + (\varepsilon_{gi} - \bar{\varepsilon}_g), \tag{9}$$

where the averages are taken across observations within each cluster, so that, for example, $\bar{u}_g = N_g^{-1} \sum_{i=1}^{N_g} u_{gi}$. The intra-cluster covariance for (9) is

$$\text{Cov}(u_{gi} - \bar{u}_g, u_{gj} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gj} - \bar{\lambda}_g), \tag{10}$$

which is zero if and only if λ_{gi} is fixed across all i. In other words, the random-effects model is the *only* model within the class of factor models (8) for which including fixed effects can remove all intra-cluster dependence. Any variation in factor loadings across observations within clusters implies that fixed effects cannot remove all intra-cluster dependence.

Furthermore, (10) strongly suggests that, whether or not a regression model includes cluster fixed effects, the scores will tend to be clustered whenever within-cluster dependence

can be approximated by a factor model like (8). Including fixed effects will no doubt reduce the intra-cluster correlations, but rarely will it eliminate them. Because even very small intra-cluster correlations can have a large effect on standard errors when the clusters are large (see the discussion at the end of Section 2), it is generally very unwise to assume that cluster fixed effects make it unnecessary to use a CRVE.

3.5 Two-Way Clustering

Up to this point, we have assumed that there is clustering in only one dimension. However, there could well be clustering in two or more dimensions. With data that have both a spatial and a temporal dimension, there may be clustering by jurisdiction and also by time period. In finance, there is often clustering by firm and by year. Thus, instead of (1), we might have

$$\boldsymbol{y}_{gh} = \boldsymbol{X}_{gh}\boldsymbol{\beta} + \boldsymbol{u}_{gh}, \quad g = 1, \ldots, G, \quad h = 1, \ldots, H, \tag{11}$$

where the vectors $\boldsymbol{y}_{gh}$ and $\boldsymbol{u}_{gh}$ and the matrix $\boldsymbol{X}_{gh}$ contain, respectively, the rows of $\boldsymbol{y}$, $\boldsymbol{u}$, and $\boldsymbol{X}$ that correspond to both the g^{th} cluster in the first clustering dimension and the h^{th} cluster in the second one. The GH clusters into which the data are divided in (11) represent the intersection of the two clustering dimensions.

If there are N_g observations in the g^{th} cluster for the first dimension, N_h observations in the h^{th} cluster for the second dimension, and N_{gh} observations in the gh^{th} cluster for the intersection, the number of observations in the entire sample is $N = \sum_{g=1}^{G} N_g = \sum_{h=1}^{H} N_h = \sum_{g=1}^{G}\sum_{h=1}^{H} N_{gh}$, where N_{gh} might equal 0 for some values of g and h. The scores for the clusters in the first dimension are $\boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g$, for the clusters in the second dimension $\boldsymbol{s}_h = \boldsymbol{X}_h^\top \boldsymbol{u}_h$, and for the intersections $\boldsymbol{s}_{gh} = \boldsymbol{X}_{gh}^\top \boldsymbol{u}_{gh}$. If, by analogy with (3), we assume that

$$\boldsymbol{\Sigma}_g = \mathrm{E}(\boldsymbol{s}_g \boldsymbol{s}_g^\top),\ \boldsymbol{\Sigma}_h = \mathrm{E}(\boldsymbol{s}_h \boldsymbol{s}_h^\top),\ \boldsymbol{\Sigma}_{gh} = \mathrm{E}(\boldsymbol{s}_{gh} \boldsymbol{s}_{gh}^\top),\ \mathrm{E}(\boldsymbol{s}_{gh} \boldsymbol{s}_{g'h'}^\top) = 0 \text{ for } g \neq g', h \neq h', \tag{12}$$

then the variance matrix of the scores is seen to be

$$\boldsymbol{\Sigma} = \sum_{g=1}^{G} \boldsymbol{\Sigma}_g + \sum_{h=1}^{H} \boldsymbol{\Sigma}_h - \sum_{g=1}^{G}\sum_{h=1}^{H} \boldsymbol{\Sigma}_{gh}. \tag{13}$$

The last condition in (12) means that the scores are assumed to be independent whenever they do not share a cluster along either dimension. The third term in (13) must be subtracted in order to avoid double counting. It is important to distinguish between two-way clustering and clustering by the intersection of the two dimensions. If we assumed the latter instead of the former, then all three terms on the right-hand side of (13) would be equal, and consequently $\boldsymbol{\Sigma} = \sum_{g=1}^{G}\sum_{h=1}^{H} \boldsymbol{\Sigma}_{gh}$. Thus these assumptions are radically different.

An estimate of the variance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\boldsymbol{\Sigma}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \quad \hat{\boldsymbol{\Sigma}} = \sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top + \sum_{h=1}^{H} \hat{\boldsymbol{s}}_h \hat{\boldsymbol{s}}_h^\top - \sum_{g=1}^{G} \sum_{h=1}^{H} \hat{\boldsymbol{s}}_{gh} \hat{\boldsymbol{s}}_{gh}^\top. \tag{14}$$

Here $\hat{\boldsymbol{\Sigma}}$ is an estimate of (13), with the empirical scores defined in the usual way; for example, $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g$. In practice, each of the matrices on the right-hand side of the second equation in (14) is usually multiplied by a scalar factor, like the one in (5), designed to correct for degrees of freedom. Because the third term is subtracted, the matrix $\hat{\boldsymbol{\Sigma}}$ may not always be positive definite. This problem can be avoided by omitting the third term, which it is asymptotically valid to do under some assumptions (MacKinnon, Nielsen and Webb 2021b). Another possibility is to use an eigenvalue decomposition (Cameron, Gelbach and Miller 2011), although this merely forces the variance matrix to be positive semidefinite.

The idea of two-way clustering can, of course, be generalized to three-way clustering, four-way clustering, and so on. However, the algebra rapidly becomes daunting. If there were three clustering dimensions, for example, the analog of (13) would have seven terms.

Two-way clustering seems to have been suggested first in Miglioretti and Heagerty (2006) and rediscovered independently by Cameron, Gelbach and Miller (2011) and Thompson (2011). Although two-way clustering has been widely used in empirical work, the asymptotic theory to justify it is much more challenging than the theory for the one-way case, and this theory is still under active development (Chiang et al. 2020; Davezies et al. 2021; Chiang et al. 2021; MacKinnon et al. 2021b; Menzel 2021). We expect this area to advance rapidly over the next few years. In view of this, and because of the technical difficulties involved, we will focus mainly on one-way clustering in the remainder of the paper.

4 Asymptotic Inference

Inference in econometrics is often based on asymptotic approximations. By letting the sample size become arbitrarily large, one can often obtain a tractable (asymptotic) distribution for the statistic of interest, and then hope that this provides a good approximation to the exact distribution. With clustered data, there is more than one natural way to let the sample size become large, because we can make various assumptions about what happens to G and the N_g as we let N tend to infinity. Which assumptions it is appropriate to use will depend on the characteristics of the sample and the (unknown) DGP.

For the regression model (1), inference is commonly based on the t-statistic,

$$t_a = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{\left(\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a}\right)^{1/2}}, \tag{15}$$

where the hypothesis to be tested is $\boldsymbol{a}^\top \boldsymbol{\beta} = \boldsymbol{a}^\top \boldsymbol{\beta}_0$, with $\boldsymbol{a}$ a known k-vector. Here $\hat{\boldsymbol{V}}$ may denote the CV_1 CRVE in (5) or perhaps some other CRVE (Section 6.1). In many cases, just one element of $\boldsymbol{a}$, say the j^{th}, equals 1, and the remaining elements equal 0, so that (15) is simply $\hat{\beta}_j - \beta_{j0}$ divided by its standard error. When there are $r > 1$ linear restrictions, which can be written as $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$ with $\boldsymbol{R}$ an $r \times k$ matrix, inference can be based on the Wald statistic,

$$W = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \left(\boldsymbol{R}\hat{\boldsymbol{V}}\boldsymbol{R}^\top\right)^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}). \tag{16}$$

Of course, when $r = 1$, the t-statistic (15) is just the signed square root of a particular Wald statistic with $\boldsymbol{R} = \boldsymbol{a}^\top$ and $\boldsymbol{r} = \boldsymbol{a}^\top \boldsymbol{\beta}_0$.

4.1 Assumptions for Asymptotic Inference

In order for inferences based on the statistics (15) and (16) to be asymptotically valid, two key asymptotic results must hold. First, a central limit theorem (CLT) must apply to the sum of the score vectors $\boldsymbol{s}_g$ in (2). In the limit, after appropriate normalization, the vector $\sum_{g=1}^G \boldsymbol{s}_g$ needs to follow a multivariate normal distribution with variance matrix $\sum_{g=1}^G \boldsymbol{\Sigma}_g$. Second, again after appropriate normalization, a law of large numbers (LLN) must apply to the matrix $\sum_{g=1}^G \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top$ in the middle of the variance matrix estimator (5), so that it converges to $\sum_{g=1}^G \boldsymbol{\Sigma}_g$. We refer to "appropriate normalization" here rather than specifying the normalization factors explicitly because, with clustered data, the issue of normalization is a very tricky one; see Section 4.1.2. For asymptotic inference to be reliable, we need both the CLT and the LLN to provide good approximations.

The simplest assumption about how the sample size goes to infinity is that every cluster has a fixed number of observations, say M. Then $N = MG$, and both N and G go to infinity at the same rate. Thus the appropriate normalizing factor for the parameter estimates is either $\sqrt{G}$ or $\sqrt{N}$. In this case, it is not difficult to show that $\sqrt{G}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically multivariate normal with variance matrix equal to the probability limit of G times the right-hand side of (4). Moreover, the latter can be estimated consistently by G times the CV_1 matrix (5). The first proof for this case of which we are aware is in White (1984, Chapter 6); see also Hansen (2007).

In actual samples, clusters often vary greatly in size, so the assumption that every cluster is the same size is usually untenable. This assumption can be relaxed by allowing cluster

sizes to vary in different ways. Two strands of literature have evolved in the development of asymptotic theory for clustered data. These can be described as "small number of large clusters" and "large number of clusters," respectively.

4.1.1 Small Number of Large Clusters

Rather than assuming that G is proportional to N, a few authors have instead assumed that G remains fixed (i.e., is "small") as $N \to \infty$, while the size of the clusters diverges (i.e., is "large"). Notably, Bester, Conley and Hansen (2011) proved that the t-statistic (15) follows the $t(G-1)$ distribution asymptotically and that $1/r$ times the Wald statistic (16) follows the $F(r, G-1)$ distribution asymptotically. These results are proven under some very strong assumptions, however. In particular, all clusters are assumed to be the same size, say M again. In addition, the pattern of dependence within each cluster is assumed to be such that a CLT applies to the normalized score vectors $M^{-1/2}\boldsymbol{s}_g$ for all $g = 1, \ldots, G$, as $M \to \infty$.

This second assumption is crucial, as it limits the amount of dependence within each cluster and requires it to diminish quite rapidly as $M \to \infty$. Although Bester et al. (2011) discusses a particular model for which this requirement holds, it rules out the most common model of intra-cluster correlation, namely, the random-effects model discussed in Section 3.3. It also rules out the standard factor model, even in the presence of cluster fixed effects, as discussed in Section 3.4. For these models, no CLT can possibly apply to the vector $M^{-1/2}\boldsymbol{s}_g$.

To see why a CLT cannot apply to $M^{-1/2}\boldsymbol{s}_g$ under the factor model in (8), suppose that $\boldsymbol{x}_g = 1$. In this case,

$$\operatorname{Var}(M^{-1/2}\boldsymbol{s}_g) = \frac{1}{M}\sum_{i,j=1}^{M}\operatorname{Cov}(u_{gi}, u_{gj}) = \frac{1}{M}\sum_{i=1}^{M}\lambda_{gi}^2 + \frac{2}{M}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\lambda_{gi}\lambda_{gj}. \tag{17}$$

Because of the double summation, the second term on the right-hand side of (17) clearly does not converge as $M \to \infty$ unless additional, and very strong, assumptions are made.

Another, quite different, approach to inference when G is fixed was developed in Ibragimov and Müller (2010). The parameter of interest is a scalar, say β, which can be thought of as one element of $\boldsymbol{\beta}$. The key idea is to estimate β separately for each of the G clusters. This yields estimates $\hat{\beta}_g$ for $g = 1, \ldots, G$. Inference is then based on the average, say $\bar{\beta}$, and standard error, say $s_{\hat{\beta}}$, of the $\hat{\beta}_g$. Ibragimov and Müller (2010) shows that the test statistic $\sqrt{G}(\bar{\beta} - \beta_0)/s_{\hat{\beta}}$ is approximately distributed as $t(G-1)$ when all clusters are large and a CLT applies to $N_g^{-1/2}\boldsymbol{s}_g$ for each g. As we saw above, this assumption cannot hold even for the simple random-effects or factor models.

A practical problem with this procedure is that β may not be estimable for at least some clusters. For models of treatment effects at the cluster level, this will actually be the case for

every cluster. For difference-in-differences (DiD) models with clustering at the jurisdiction level, it will be the case for every jurisdiction that is never treated. Ibragimov and Müller (2016) suggests a way to surmount this problem by combining clusters into larger ones that allow β to be estimated for each of them. Even when β itself can be estimated for each cluster, the full model may not be estimable. This can happen, for example, when there are fixed effects for categorical variables, and not all types of observations are in each category. When this occurs, the interpretation of β may differ across clusters.

Although the estimates and test statistics proposed in Ibragimov and Müller (2010, 2016) differ from the more conventional ones studied in Bester et al. (2011), both approaches lead to t-statistics that follow the $t(G-1)$ distribution asymptotically. This distribution has in fact been used in `Stata` as the default for CRVE-based inference for many years. For small values of G, using it can lead to noticeably more conservative inferences than using the $t(N-k)$ or normal distributions. However, as we discuss in Section 4.2 and Section 5, inferences based on $t(G-1)$ are often not nearly conservative enough.

4.1.2 Large Number of Clusters

The asymptotic approximations discussed so far depend on rather extreme assumptions, namely, that G is either proportional to N or G is fixed with identically sized homogeneous clusters. The former assumption may be relaxed by allowing G to be only approximately proportional to N, so that G/N is roughly constant as $N \to \infty$. This implies that all the clusters must be small. In this case, the quality of the asymptotic approximations is not likely to be harmed much by moderate variation in cluster sizes. If a sample has, say, 500 clusters that vary in size from 10 to 50 observations, we would expect asymptotic inference to perform very well unless there is some other reason (unrelated to cluster sizes) for it to fail.

Djogbenou, MacKinnon and Nielsen (2019) and Hansen and Lee (2019) take a more flexible approach, with primitive conditions that restrict the variation in the N_g relative to the sample size. These conditions allow some clusters to be "small" and others to be "large" in the sense that some but not all $N_g \to \infty$ as $N \to \infty$. Although a key assumption is that $G \to \infty$ (i.e., is "large"), the appropriate normalization factor for $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is usually not $\sqrt{G}$. Instead, this factor depends in a complicated way on the regressors, the relative cluster sizes, the intra-cluster correlation structure, and interactions among these; some examples of different normalizing factors are given in the papers cited above. For this reason, the key result that the t-statistic defined in (15) is asymptotically distributed as standard normal is derived assuming that the rate at which $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ tends to zero is unknown. Of course, this result also justifies using the $t(G-1)$ distribution, which is more conservative.

The application of a CLT to $\sum_{g=1}^{G} \boldsymbol{s}_g$, appropriately normalized, requires a restriction on

the amount of heterogeneity that is allowed. Otherwise, just a few clusters might dominate the entire sample in the limit, thus violating the Lindeberg or Lyapunov conditions. The necessary restrictions on the heterogeneity of clusters may be expressed in terms of two key parameters. The first of these parameters is the number of moments that is assumed to exist for the distributions of $\boldsymbol{s}_{gi}$ (uniformly in g and i). We denote this parameter by $\gamma > 2$. When more moments exist, the distributions of $\boldsymbol{s}_{gi}$ are closer to the normal distribution, and hence the sample will feature fewer outliers or other highly leveraged observations or clusters.

In the clustered regression model, the variance of the scores, which is often referred to as the Fisher information matrix, is given by $\mathcal{J}_N = \sum_{g=1}^{G} \text{Var}(\boldsymbol{s}_g)$. When appropriately normalized, $\mathcal{J}_N$ converges to a nonzero and finite matrix $\mathcal{J}$. The rate of convergence η_N is defined implicitly by $\eta_N^{-1}\mathcal{J}_N \to \mathcal{J}$. This rate is the second key parameter. One interpretation of η_N can be found in (4), from which the stochastic order of magnitude of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is seen to be $O_P(\eta_N^{1/2}/N)$. In general, $\eta_N \geq N$, with the equality holding whenever there is no intra-cluster correlation. The larger the value of η_N, the more slowly does $\hat{\boldsymbol{\beta}}$ converge to $\boldsymbol{\beta}_0$.

The conditions required on the heterogeneity of clusters to apply a CLT can be stated in terms of the parameters γ and η_N. Specifically, when expressed in our notation, Assumption 3 of Djogbenou et al. (2019) states the following condition:

$$\left(\frac{\eta_N^{1/2}}{N}\right)^{\frac{-2\gamma}{2\gamma-2}} \frac{\sup_g N_g}{N} \longrightarrow 0. \tag{18}$$

Because $\eta_N = o(N^2)$ for consistency of $\hat{\boldsymbol{\beta}}$, the condition in (18) makes it clear that we cannot allow a single cluster to dominate the sample, in the sense that its size is proportional to N. More generally, (18) shows that there is a tradeoff between information accumulation and variation in cluster sizes, as measured by the largest cluster size. To interpret the condition in (18), we will consider three different comparative statics.

First, when γ increases, condition (18) becomes less strong. In particular, when the scores are nearly normally distributed, in the sense that all their moments exist, then $\gamma = \infty$, which implies that $-2\gamma/(2\gamma - 2) = -1$. In this case, (18) reduces to $\eta_N^{-1/2} \sup_g N_g \to 0$, so that the size of the largest cluster must increase more slowly than the square root of the rate at which the Fisher information matrix converges. When $\gamma < \infty$, so that there are fewer moments, then the rate at which $\sup_g N_g$ is allowed to increase becomes smaller.

Second, suppose that the scores are uncorrelated, or more generally that a CLT applies to $N_g^{-1/2}\boldsymbol{s}_g$ as assumed in Bester et al. (2011) and Ibragimov and Müller (2010, 2016). In this case, $\text{Var}(\boldsymbol{s}_g) = O(N_g)$, so that $\eta_N = N$. Condition (18) is then $N^{-(\gamma-2)/(2\gamma-2)} \sup_g N_g \to 0$. If the scores are nearly normal, then it reduces further to $N^{-1/2} \sup_g N_g \to 0$, so that the size of the largest cluster must increase no faster than the square root of the sample size.

Once again, the fewer moments there are, the more slowly $\sup_g N_g$ is allowed to increase.

Third, suppose that the scores are generated by the factor model in (8), or by the simpler random-effects model. Then $\text{Var}(\boldsymbol{s}_g) = O(N_g^2)$. If, in addition, $\inf_g N_g$ and $\sup_g N_g$ are of the same order of magnitude, then $\eta_N = N \sup_g N_g$ and the condition (18) collapses to $N^{-1} \sup_g N_g \to 0$, regardless of the number of finite moments.

One, possibly surprising, implication of the above considerations is that, when there is more intra-cluster correlation, so that η_N is relatively large, then greater heterogeneity of cluster sizes is allowed. That is, a higher degree of intra-cluster correlation implies a faster rate of convergence, η_N, of the Fisher information matrix, which in turn allows a larger $\sup_g N_g$ in (18). The intuition is that, with a high degree of intra-cluster correlation, the effective cluster size, as measured by the amount of independent information, is relatively small. In the extreme case in which all observations in a cluster are perfectly correlated, the cluster size is effectively one and not N_g. Note, however, that large clusters are implicitly weighted more heavily than small clusters even in this extreme case.

4.2 When Asymptotic Inference Can Fail

Whenever we rely on asymptotic theory, we need to be careful. What is true for infinitely large samples may or may not provide a good approximation for any actual sample. In general, the number of clusters G and the extent to which the distribution of the scores varies across clusters will determine the quality of the asymptotic approximation.

Unless the very strong assumptions discussed in Section 4.1.1 are satisfied, we cannot expect to obtain reliable inferences when G is small. Unfortunately, there is no magic number for G above which asymptotic inference can be relied upon. In very favorable cases, inference based on the $t(G-1)$ distribution can be fairly reliable when $G = 20$, but in unfavorable ones it can be seriously unreliable even when $G = 200$.

4.2.1 Heterogeneity of Clusters

What determines whether a case is favorable or unfavorable is mostly the heterogeneity of the scores. The discussion in Section 4.1 focused on cluster sizes, which are often particularly important, but any form of heterogeneity can have serious consequences. This includes both heteroskedasticity of the disturbances at the cluster level and systematic variation across clusters in the distribution of the regressors. The more heterogeneity there is across clusters, the worse the asymptotic approximation will likely be.

In principle, a poor asymptotic approximation could lead t-tests based on the $t(G-1)$ distribution either to under-reject or over-reject. However, we have never observed these

tests to under-reject in any simulation experiments. In the very best cases, they may reject between 5% and 6% of the time for tests at the 5% level. More commonly, unless G is fairly large, they tend to over-reject much more than that. It is not difficult to find cases in which t-tests reject more than 20% of the time when $G = 20$. Wald tests of several restrictions typically perform even worse (Pustejovsky and Tipton 2018).

As we discussed in Section 4.1.2, the condition (18) imposes a restriction on the size of the largest cluster relative to the sample size. Thus, the quality of the asymptotic approximation will surely diminish as the size of the largest cluster increases relative to the average cluster size, and over-rejection will consequently increase. This conjecture is supported by simulation evidence in MacKinnon and Webb (2017a) and Djogbenou et al. (2019), as well as by analytic results based on Edgeworth expansions in the latter paper.

One way to quantify the heterogeneity of cluster sizes and regressors is to calculate G^*, the "effective number of clusters," as proposed in Carter, Schnepel and Steigerwald (2017). This number, which is always less than G, can provide a useful warning when G^* is much smaller than G. It is sensitive not only to the variation in cluster sizes but also to other features of the $\boldsymbol{X}_g$ matrices, although it is not sensitive to heteroskedasticity or other features of the disturbances. The value of G^* depends on an unknown parameter ρ, the intra-cluster correlation of the disturbances in the equi-correlated case. Carter et al. (2017) suggest setting $\rho = 1$, as a sort of worst case, but, in our view, it is often more realistic to set $\rho = 0$. These two choices yield what we refer to as G_1^* and G_0^*, respectively; see Section 8.

There are two situations in which cluster-robust t-tests and Wald tests are at risk of over-rejecting to an extreme extent, namely, very large clusters and few treated clusters. In both of these cases, one cluster, or just a few of them, have high leverage or are very influential, in the sense that omitting one of these clusters has the potential to change the OLS estimates substantially (Belsley, Kuh and Welsch 1980). Since both of these situations can occur even when G is not small, all users of cluster-robust inference need to be on guard for them. A computationally efficient way to identify high-leverage and influential clusters is discussed in MacKinnon, Nielsen and Webb (2021a); see also Section 4.4.

The first case in which conventional inference fails is when one or two clusters are very much larger than any of the others. This implies that the distributions of the score vectors for those clusters are much more spread out than the ones for the rest of the clusters. A particularly extreme example is studied in Djogbenou et al. (2019, Figure 3). When half the sample is in one large cluster and all the other clusters are small, rejection rates at the 5% level actually increase as G increases, approaching 50% for $G = 201$. Unfortunately, this extreme case is empirically relevant. Because roughly half of all incorporations in the United States are in Delaware, empirical studies of state laws and corporate governance encounter

precisely this situation whenever they cluster at the state level (Spamann 2019).

Not all forms of heterogeneity are harmful. In particular, having some extremely small clusters in a sample generally does not cause any problems, so long as there is not too much heterogeneity in the remainder of the sample. For example, suppose that a sample consists of, say, 25 large clusters, each with roughly 200 observations, and 15 tiny clusters, each with just one or a handful of observations. Except in very unusual cases, the coefficient estimates and their t-statistics would hardly change if we were to drop the tiny clusters, so this sample effectively has just 25 clusters. The asymptotic approximations would perform just about the same whether or not the tiny clusters were included.

Of course, if we changed the above example so that there were 5 large clusters and 15 tiny ones, then inference would surely be very problematic, because there would effectively be just 5 clusters. This would be apparent if we were to calculate G^*.

4.2.2 Treatment and Few Treated Clusters

The second case in which conventional inference fails is when the regressor of interest is a treatment dummy, and treatment occurs only for observations in a small number of clusters. In such cases, the empirical score vectors $\hat{\boldsymbol{s}}_g$ for the treated clusters provide very poor estimates of the actual score vectors $\boldsymbol{s}_g$.

Suppose that d_{gi} is the value of the treatment dummy for observation i in cluster g, and let s_g^d denote the element of $\boldsymbol{s}_g$ corresponding to the dummy. Consider first the extreme case in which only observations in the first cluster are treated. Then $s_g^d = \sum_{i=1}^{N_g} d_{gi} u_{gi}$ is equal to $\sum_{i=1}^{N_1} d_{1i} u_{1i}$ for $g = 1$ and to 0 for all $g \neq 1$. Thus the scores corresponding to the treatment dummy equal zero for the control clusters. Moreover, because the treatment regressor must be orthogonal to the residuals, the empirical score $\hat{s}_1^d = 0$. Since the actual score $s_1^d \neq 0$, this implies that (5) provides a dreadful estimate of (4), at least for the elements corresponding to the coefficient on the treatment dummy. In consequence, the CV_1 standard error of this coefficient can easily be too small by a factor of five or more. When more than one cluster is treated, the problem is not as severe, because the $\hat{s}_g^d$ now sum to zero over the observations in all the treated clusters. This causes them to be too small, but not to the same extent as when just one cluster is treated; see MacKinnon and Webb (2017a, 2018).

The sizes of the treated and control clusters, the values of other regressors, and the number of treated observations within the treated clusters all affect how well the empirical scores mimic the actual scores. Thus they affect the accuracy of cluster-robust standard errors and the extent to which t-statistics based on them over-reject. As the number of treated clusters, say G_1, increases, the problem often goes away fairly rapidly. But increasing G when G_1 is small and fixed does not help and may well cause over-rejection to increase.

For models where all observations in each cluster are either treated or not, having very few control clusters is just as bad as having very few treated clusters. The situation is more complicated for DiD models, however; see MacKinnon and Webb (2017b).

4.2.3 Testing Several Restrictions

Most of the literature on cluster-robust inference has focused on t-tests. Of course, the cluster-robust Wald tests defined in (16) also tend to over-reject in finite samples. In fact, they tend to do so more severely as r, the number of restrictions, increases; see Pustejovsky and Tipton (2018). As is well known, this phenomenon occurs for Wald tests of all kinds. The problem may well be unusually severe in this case, however, because the CV_1 variance matrix (5) has rank at most G (in many cases, only $G-1$). It therefore seems very likely that the inverse of $\boldsymbol{R}^\top \hat{\boldsymbol{V}} \boldsymbol{R}$ will provide a worse approximation to the inverse of $\boldsymbol{R}^\top \text{Var}(\hat{\boldsymbol{\beta}}) \boldsymbol{R}$ as r becomes closer to G, which in turn must cause the distribution of W/r to be less well approximated by the $F(r, G-1)$ distribution.

4.3 Cluster-Robust Inference in Nonlinear Models

Although cluster-robust inference is most commonly used with the linear regression model (1), it can actually be employed for a wide variety of models estimated by maximum likelihood or the generalized method of moments (GMM); see Hansen and Lee (2019).

Consider a model characterized by the log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \ell_{gi}(\boldsymbol{\theta}), \tag{19}$$

where $\boldsymbol{\theta}$ is the $k \times 1$ parameter vector to be estimated, and $\ell_{gi}(\boldsymbol{\theta})$ denotes the contribution to the log-likelihood made by the i^{th} observation within the g^{th} cluster. Let $\hat{\boldsymbol{\theta}}$ denote the vector that maximizes (19), $\boldsymbol{s}_{gi}(\boldsymbol{\theta})$ the $k \times 1$ vector of the first derivatives of $\ell_{gi}(\boldsymbol{\theta})$ (that is, the score vector), and $\boldsymbol{H}_{gi}(\boldsymbol{\theta})$ the $k \times k$ Hessian matrix of the second derivatives. Further, let $\hat{\boldsymbol{s}}_g = \sum_{i=1}^{N_g} \boldsymbol{s}_{gi}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{H}} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \boldsymbol{H}_{gi}(\hat{\boldsymbol{\theta}})$. Then Hansen and Lee (2019, Theorem 10) shows (using somewhat different notation) that the cluster-robust variance estimator for the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \hat{\boldsymbol{H}}^{-1} \left(\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top \right) \hat{\boldsymbol{H}}^{-1}. \tag{20}$$

The resemblance between (20) and the CV_1 variance matrix in (5) is striking. Indeed, since the Hessian is proportional to $\boldsymbol{X}^\top \boldsymbol{X}$ for the linear regression model, CV_1 without the leading

scalar factor is really just a special case of (20).

The variance matrix estimator (20) can be used for a wide variety of models estimated by maximum likelihood. In fact, `Stata` has been using it for various models, including probit and logit, for some years. Hansen and Lee (2019, Theorem 12) provides a similar result for GMM estimation, which is also very widely applicable. Unfortunately, at the present time, very little seems to be known about the finite-sample properties of tests and confidence intervals based on (20) or its GMM analog. They are probably even less reliable than ones based on (15) or (16). It seems quite plausible that bootstrapping would help, but the properties of applicable bootstrap procedures are largely unknown at the present time.

4.4 What Should Investigators Report?

Many empirical studies that use cluster-robust inference fail to report enough information to convince readers that the results should be believed. Investigators often simply use the $t(G-1)$ or $F(r, G-1)$ distributions, instead of the alternative methods discussed in Sections 5 and 6, and they rarely provide any evidence that doing so yields reliable results. This can often be done by reporting some key information about the sample at hand, and we strongly recommend that investigators do so as a matter of routine.

The fundamental unit of inference when the observations are clustered is the cluster, not the observation; this is evident from (2) and (4). The asymptotic theory discussed in Section 4.1.2 therefore depends on G, not N. With the exception of certain very special cases discussed in Section 4.1.1, asymptotic approximations tend to work poorly when there are few clusters. It is therefore absolutely essential to report the number of clusters, G, whenever inference is based on a CRVE. This is even more important than reporting N.

Moreover, because the approximations work best when the scores are homogeneous across clusters, and the most important source of heterogeneity is often variation in cluster sizes, it is extremely important to report measures of this variation. At a minimum, we believe that investigators should always report the median cluster size and the maximum cluster size, in addition to N and G. In addition, it is generally a good idea to report both versions of G^*, the effective number of clusters (Section 4.2.1). When G is small, or when the distribution of the N_g is unusual, it would be good to report the entire distribution of cluster sizes in the form of either a histogram or a table.

Clusters can be heterogeneous in many ways beyond their sizes. Classic measures of observation-level heterogeneity are leverage and influence (Belsley, Kuh and Welsch 1980; Chatterjee and Hadi 1986). These were generalized to cluster-level measures in MacKinnon et al. (2021a). One possible consequence of heterogeneity is that the estimates may change a

lot when certain clusters are deleted. When this is the case, a cluster is said to be influential. In order to identify individually influential clusters, we can construct the matrices $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ and the vectors $\boldsymbol{X}_g^\top \boldsymbol{y}_g$, for $g = 1, \ldots, G$. Then

$$\hat{\boldsymbol{\beta}}^{(g)} = (\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}_g^\top \boldsymbol{X}_g)^{-1}(\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}_g^\top \boldsymbol{y}_g) \tag{21}$$

is the vector of least squares estimates when cluster g is deleted. It should not be expensive to compute $\hat{\boldsymbol{\beta}}^{(g)}$ for every cluster using (21).

When there is a parameter of particular interest, say β, then it will often be a good idea to report the $\hat{\beta}^{(g)}$ for $g = 1, \ldots, G$ in either a histogram or a table. If $\hat{\beta}^{(h)}$ differs a lot from $\hat{\beta}$ for some cluster h, then cluster h is evidently influential. In a few extreme cases, there may be a cluster h for which it is impossible to compute $\hat{\beta}^{(h)}$. If so, then the original estimates should probably not be believed. This will happen, for example, when cluster h is the only treated one, and we saw in Section 4.2.2 that inference is extremely unreliable in that case.

As pointed out in Belsley et al. (1980) and Chatterjee and Hadi (1986), it is often valuable to identify high-leverage observations as well as influential ones. It is perhaps even more valuable to identify high-leverage clusters (MacKinnon et al. 2021a). Loosely speaking, a high-leverage cluster is one whose regressors contain a lot of information. At the observation level, high-leverage observations are associated with a high value of h_i, the i^{th} diagonal element of $\boldsymbol{H} = \boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. The analog of h_i in the cluster case is the $N_g \times N_g$ matrix $\boldsymbol{H}_g = \boldsymbol{X}_g(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}_g^\top$.

Since it is not feasible to report the $\boldsymbol{H}_g$, we suggest that investigators report the values of the trace of $\boldsymbol{H}_g$. Computing this scalar for all g is quite easy, because we have already calculated $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and the $\boldsymbol{X}_g^\top \boldsymbol{X}_g$. For any cluster that contains just one observation, $\text{Tr}(\boldsymbol{H}_g)$ reduces to the usual measure of leverage at the observation level. High-leverage clusters can be identified by comparing the $\text{Tr}(\boldsymbol{H}_g)$ to their own average, which is k/G. If, for some h, $\text{Tr}(\boldsymbol{H}_h)$ is substantially larger than k/G, then cluster h has high leverage. This can happen either because N_h is much larger than G/N or because the matrix $\boldsymbol{X}_h$ is somehow extreme relative to the other $\boldsymbol{X}_g$ matrices, or both.

Regression models often include cluster fixed effects. It is computationally attractive to partial them out before estimation begins, using for example the `areg` procedure in `Stata`. It is essential to do this if we are to compute the $\hat{\beta}^{(g)}$ and the $\text{Tr}(\boldsymbol{H}_g)$. The problem is that, when one of the regressors is a fixed-effect dummy for cluster g, the matrices $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ and $\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}_g^\top \boldsymbol{X}_g$ are singular. However, the problem solves itself if we partial out the fixed-effect dummies and replace $\boldsymbol{X}$ by $\tilde{\boldsymbol{X}}$ and $\boldsymbol{y}$ by $\tilde{\boldsymbol{y}}$, the matrix and vector of deviations from cluster means. For example, the gj^{th} element of $\tilde{\boldsymbol{y}}$ is $y_{gj} - N_g^{-1}\sum_{i=1}^{N_g} y_{gi}$.

Our suggestion that investigators should routinely compute the $\hat{\beta}^{(g)}$ and the $\text{Tr}(\boldsymbol{H}_g)$ and

report them, at least when they provide information beyond that in the distribution of the cluster sizes, seems to be new. At the moment, we do not have much experience with doing this; see Section 8.1. However, we are confident that reporting these measures of influence and leverage would help to identify cases in which inference may be unreliable, as well as sometimes turning up interesting, or possibly disturbing, features of the data.

5 Bootstrap Inference

Instead of basing inference on an asymptotic approximation to the distribution of a statistic of interest, it is often more reliable to base it on a bootstrap approximation. In the next subsection, we briefly review some key concepts of bootstrap testing and bootstrap confidence intervals. Then, in the following two subsections, we discuss bootstrap methods for regression models with clustered data. Bootstrap methods are particularly attractive in this context. They often lead to dramatically more reliable inferences than asymptotic procedures, and, in many cases, they are astonishingly inexpensive to compute.

5.1 General Principles of the Bootstrap

Suppose we are interested in a test statistic τ, which might be a t-statistic or a Wald statistic. Instead of using P values or critical values taken from an asymptotic distribution, we can use ones from the empirical distribution function (EDF) of a large number of bootstrap test statistics. This EDF often provides a good approximation to the unknown distribution of τ. In order to obtain the EDF, we need to generate B bootstrap samples and use each of them to compute a bootstrap test statistic, say τ_b^*, for $b = 1, \ldots, B$.

Precisely how the bootstrap samples are generated is critical, and we will discuss some methods for doing so in the next two subsections. The choice of B also matters. Ideally, it should be reasonably large (Davidson and MacKinnon 2000) and satisfy the condition that $\alpha(B + 1)$ is an integer for any α (the level of the test) of interest (Racine and MacKinnon 2007). In general, the computational cost of generating the bootstrap test statistics is proportional to B times N, so that bootstrapping can be expensive. However, as we discuss in Section 5.3, surprisingly inexpensive methods are available for linear regression models with clustered disturbances. Unless computational cost is an issue, $B = 9{,}999$ and even $B = 99{,}999$ seem to be good choices.

The EDF of the τ_b^* often provides a better approximation to $F(\tau)$, the distribution of τ, than its asymptotic distribution does. This can sometimes be shown formally, but generally only under strong assumptions and at the cost of a great deal of algebra (Djogbenou et al.

2019, Section 5). For the model (1), however, the intuition is quite simple. In many cases, the poor finite-sample properties of cluster-robust test statistics arise because $\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top$ provides a poor approximation to $\sum_{g=1}^{G} \boldsymbol{\Sigma}_g$. We may hope that the bootstrap analog of the former provides a similarly poor approximation to the bootstrap analog of the latter. If so, then it is plausible that the empirical distribution of the τ_b^* will differ from their asymptotic distribution in roughly the same way as the distribution of τ differs from its asymptotic distribution. In that case, the EDF of the bootstrap test statistics should provide a reasonably good approximation to $F(\tau)$.

The EDF of the τ_b^* may be obtained by sorting the τ_b^* from smallest to largest. Number $\alpha(B+1)$ then provides an estimate of the α quantile. There is more than one type of bootstrap P value. If a test rejects in the upper tail, then it is appropriate to use

$$\hat{P}^*(\tau) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \tau). \tag{22}$$

Here τ could be either the Wald statistic (16) or the absolute value of the t-statistic (15). Setting $\tau = |t_a|$ imposes symmetry on the bootstrap distribution of t_a itself. In many cases, it makes sense to do this, because cluster-robust t-statistics for linear regression models with exogenous regressors are often symmetrically distributed around the origin, at least to a good approximation. When $\tau = |t_a|$, then $\hat{P}^*(\tau)$ defined in (22) is often called a symmetric bootstrap P value.

In dynamic models, nonlinear models, and models estimated by instrumental variables, it is common for key coefficients to be biased. This causes t-statistics to have non-zero means in finite samples. In such cases, it makes sense to use the equal-tail bootstrap P value,

$$\hat{P}^*_{\text{et}}(\tau) = \frac{2}{B} \min\left(\sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \tau), \sum_{b=1}^{B} \mathbb{I}(\tau_b^* \leq \tau) \right). \tag{23}$$

Here we compute upper-tail and lower-tail P values, take the minimum of them, and then multiply by 2 to ensure that the nominal level of the test is correct.

There are many ways to construct a bootstrap confidence interval for a regression coefficient β of which we have an estimate $\hat{\beta}$. A method that is conceptually (but not always computationally) simple is to invert a bootstrap test. This means finding two values of the coefficient, say β_l and β_u, with $\beta_u > \beta_l$ and normally on opposite sides of $\hat{\beta}$, such that

$$\hat{P}^*_{\text{et}}\Big(t(\beta = \beta_l)\Big) = \alpha \quad \text{and} \quad \hat{P}^*_{\text{et}}\Big(t(\beta = \beta_u)\Big) = \alpha. \tag{24}$$

Here $t(\beta = \beta_a)$, for $a = l$ and $a = u$, is a cluster-robust t-statistic for the hypothesis that $\beta = \beta_a$. The desired $1 - \alpha$ confidence interval is then $[\beta_l, \beta_u]$. When the bootstrap DGP

imposes the null hypothesis, the distribution of the bootstrap samples depends on the value of β_a. Solving the two equations in (24) therefore requires iteration; see Hansen (1999) and MacKinnon (2015). In general, these methods tend to be expensive, but that is often not the case for the bootstrap method to be discussed in Section 5.3.

For bootstrap confidence intervals, it is common to use bootstrap DGPs that do not impose the null hypothesis, because no iteration is then required. The simplest method is just to calculate the standard deviation of the $\hat{\beta}_b^*$ and use this number, say $s^*(\hat{\beta})$, as an estimate of the standard error of $\hat{\beta}$. The confidence interval is then

$$\Big[\hat{\beta} - c_{1-\alpha/2}\, s^*(\hat{\beta}),\ \hat{\beta} + c_{1-\alpha/2}\, s^*(\hat{\beta})\Big], \tag{25}$$

where $c_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of (in this case) the $t(G-1)$ distribution. A better approach, at least in theory, is to use the studentized bootstrap, or percentile-t, confidence interval advocated in Hall (1992), which is

$$\Big[\hat{\beta} - s_\beta c^*_{1-\alpha/2},\ \hat{\beta} - s_\beta c^*_{\alpha/2}\Big], \tag{26}$$

where s_β is the standard error of $\hat{\beta}$ from the CRVE, and c_z^* denotes the z quantile of the bootstrap t-statistics τ_b^*. Although the higher-order theory in Djogbenou et al. (2019) does not explicitly deal with confidence intervals, it strongly suggests that the intervals (25) and (26) should not perform as well as inverting a bootstrap test based on a bootstrap DGP that imposes the null hypothesis. Simulation results in MacKinnon (2015) are consistent with these predictions. However, the intervals (25) and (26) have the advantage that they are easy to compute. No iteration is required, and a single set of bootstrap samples can be used to compute confidence intervals for all the parameters of interest.

5.2 Pairs Cluster Bootstrap

The most important aspect of any bootstrap procedure is how the bootstrap samples are generated. The only procedure applicable to every model that uses clustered data is the pairs cluster bootstrap, which is also sometimes referred to as the cluster bootstrap, the block bootstrap, or resampling by cluster. The pairs cluster bootstrap works by grouping the data for every cluster into a $[\boldsymbol{y}_g, \boldsymbol{X}_g]$ pair and then resampling from the G pairs. Every bootstrap sample is constructed by choosing G pairs at random with equal probability $1/G$.

Although this procedure ensures that every bootstrap sample contains G clusters, the number of observations inevitably varies across the bootstrap samples. When the cluster sizes vary considerably, the size of the bootstrap samples can vary greatly, because the largest clusters may be over-represented in some bootstrap samples and under-represented

in others. This limits the ability of the bootstrap samples to mimic the actual sample. So does the fact that the $\boldsymbol{X}^\top \boldsymbol{X}$ matrix is different for every bootstrap sample.

Because the pairs cluster bootstrap does not impose the null hypothesis, care must be taken when calculating the bootstrap test statistics. If the null hypothesis is that $\beta = \beta_0$, the actual t-statistic will have numerator $\hat{\beta} - \beta_0$, but the bootstrap t-statistic must have numerator $\hat{\beta}_b^* - \hat{\beta}$, where $\hat{\beta}_b^*$ is the estimate of β for the b^{th} bootstrap sample. In this case, $\hat{\beta}$ is the parameter value associated with the bootstrap DGP. Because the bootstrap DGP does not impose the null hypothesis, the pairs cluster bootstrap cannot be used to construct the confidence interval (24), but it can be used to construct the intervals (25) and (26).

Cameron and Miller (2015, Section VI) discussed several problems that can arise with the pairs cluster bootstrap and sensibly suggested that investigators should examine the empirical distributions of the bootstrap coefficient estimates and test statistics. For example, if the bootstrap distribution has more than one mode, then it probably does not provide a good approximation to the actual distribution. This can happen when one or two clusters are very different from all the others.

Ferman and Pinto (2019) proposed several bootstrap procedures that can be thought of as variants of the pairs (not pairs cluster) bootstrap. The first step is to run regressions at either the individual level or the group $\times$ time-period level, then aggregate the residuals so that there is just one residual per cluster, and finally run bootstrap regressions on the resampled residuals. These procedures include parametric methods to correct for the heteroskedasticity generated by variation in the number of observations per group. Remarkably, they can work well even with just one treated cluster. However, this is possible only because, unlike CRVE-based methods, they do not allow for unrestricted heteroskedasticity.

In general, the pairs cluster bootstrap is expensive to compute. However, a computational shortcut for linear regression models can make it feasible even when N and B are both large; see MacKinnon (2021). Nevertheless, we do not recommend this method for the linear regression model (1), because, as we discuss in the next subsection, a much better method is available. With nonlinear models such as the probit model, the pairs cluster bootstrap may be attractive even though it can be expensive. However, we cannot recommend it without reservation, because it appears that very little is known about its finite-sample properties. At least in the case of treatment and DiD models, it can either over-reject or under-reject quite severely (MacKinnon and Webb 2017b).

5.3 Wild Cluster Bootstrap

For the linear regression model (1), the best method at present seems to be the restricted wild cluster, or WCR, bootstrap, which was first suggested in Cameron et al. (2008). Suppose that $\tilde{\boldsymbol{\beta}}$ denotes the OLS estimate of $\boldsymbol{\beta}$ subject to the restriction $\boldsymbol{a}^\top \boldsymbol{\beta} = \boldsymbol{a}^\top \boldsymbol{\beta}_0$, which is to be tested, and $\tilde{\boldsymbol{u}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g \tilde{\boldsymbol{\beta}}$ denotes the vector of restricted residuals for the g^{th} cluster. Then the bootstrap DGP is

$$\boldsymbol{y}_g^{*b} = \boldsymbol{X}_g \tilde{\boldsymbol{\beta}} + \boldsymbol{u}_g^{*b}, \quad \boldsymbol{u}_g^{*b} = v_g^{*b} \tilde{\boldsymbol{u}}_g, \quad g = 1, \ldots, G, \tag{27}$$

where the v_g^{*b} are independent realizations of an auxiliary random variable v^* with zero mean and unit variance. In practice, the best choice for v^* is usually the Rademacher distribution, in which case v^* equals 1 or -1 with equal probabilities (Davidson and Flachaire 2008; Djogbenou et al. 2019). This imposes symmetry on the bootstrap disturbances.

Because the $\boldsymbol{X}_g$ are the same for every bootstrap sample, the WCR bootstrap can also be computed directly using the restricted scores. The bootstrap DGP (27) is replaced by

$$\boldsymbol{s}_g^{*b} = v_g^{*b} \tilde{\boldsymbol{s}}_g, \quad g = 1, \ldots, G, \tag{28}$$

where $\tilde{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \tilde{\boldsymbol{u}}_g$ is the score vector for the g^{th} cluster evaluated at the restricted estimates. Plugging the $\boldsymbol{s}_g^{*b}$ into (2) then yields the bootstrap estimates $\hat{\boldsymbol{\beta}}_b^*$. This method is computationally inexpensive (MacKinnon 2021), and it provides an intuitive justification for the WCR bootstrap because, as we stressed in Section 4, the finite-sample properties of cluster-robust tests depend mainly on the properties of the scores.

Djogbenou et al. (2019) established the asymptotic validity of the WCR bootstrap and also studied the unrestricted wild cluster, or WCU, bootstrap. The difference between the WCR and WCU bootstraps is that, for the latter, $\hat{\boldsymbol{\beta}}$ is used in (27) instead of $\tilde{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{u}}_g$ is used instead of $\tilde{\boldsymbol{u}}_g$. In general, the WCU bootstrap does not perform as well in finite samples as the WCR one. The paper contains both theoretical results (based on Edgeworth expansions) and simulation results to support this assertion. However, the WCU bootstrap has the advantage that the bootstrap DGP does not depend on the restrictions to be tested. This means that the same set of bootstrap samples can be used to perform tests on any restriction or set of restrictions and/or to construct confidence intervals based on either (25) or (26) for any coefficient of interest.

Djogbenou et al. (2019) also proved the asymptotic validity of two variants of the ordinary wild bootstrap, restricted (WR) and unrestricted (WU), for the model (1). The ordinary wild bootstrap uses N realizations of v^*, one for each observation, instead of just G. This means that the disturbances for the bootstrap samples are uncorrelated within clusters.

Although this implies that the distribution of the $\hat{\boldsymbol{\beta}}_b^*$ cannot possibly match that of $\hat{\boldsymbol{\beta}}$, it often does not prevent the distribution of the τ_b^* from providing a good approximation to the distribution of τ. With one important exception (see below), the WR bootstrap seems to work less well than the WCR bootstrap, as asymptotic theory predicts. It can also be much more expensive to compute when N is large.

In many cases (exceptions will be discussed below), the WCR bootstrap yields very accurate inferences, which are generally more accurate than those for other bootstrap methods. In addition to Djogbenou et al. (2019), see Cameron et al. (2008) and MacKinnon and Webb (2017a, 2018). An important feature of all versions of the wild cluster bootstrap is that it is able to replicate what is often the main source of heterogeneity, namely, variation in cluster sizes, in every bootstrap sample. This is not the case for the pairs cluster bootstrap, and it surely contributes to the superior accuracy of inferences based on the WCR bootstrap.

Both versions of the wild cluster bootstrap can be surprisingly inexpensive to compute. Computations based on (28) use only the matrices $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ and the vectors $\boldsymbol{X}_g^\top \boldsymbol{y}_g$ for all the bootstrap computations; see MacKinnon (2021). The calculations can be made even faster by rewriting the bootstrap test statistic so that it depends on all the sample data in the same way for every bootstrap sample; see Roodman, MacKinnon, Nielsen and Webb (2019). The only thing that varies across the bootstrap samples is the G-vector $\boldsymbol{v}^{*b}$ of realizations of the auxiliary random variable. There are some initial computations that may be expensive when N is large, but they only have to be done once. After that, the $\boldsymbol{v}^{*b}$ and the results of the initial computations are used to compute all the bootstrap test statistics.

This fast procedure is implemented in the `Stata` package `boottest`; for details, see Roodman et al. (2019). In Section 8, there is an illustration of how fast it can be; see the notes to Table 3. Importantly, `boottest` not only computes WCR bootstrap P values for both t-tests and Wald tests; it also computes WCR bootstrap confidence intervals based on (24). The package has many other capabilities as well.

Although the WCR bootstrap generally works better than other bootstrap methods, and often works very well indeed, it does not work perfectly. Not surprisingly, its performance tends to deteriorate as G becomes smaller and the clusters become more heterogeneous. In particular, it can sometimes perform very badly when the number of treated clusters G_1 is very small (MacKinnon and Webb 2017a, b, 2018). This is true both for pure treatment models, where all the observations in each cluster are either treated or not, and for DiD models, where only some observations in the treated clusters are treated.

Unlike other methods, which generally over-reject severely when there are few treated clusters (Section 4.2.2), the WCR bootstrap usually under-rejects. This happens because the distribution of the bootstrap statistics τ_b^* depends on the value of the actual test statistic τ.

The larger is τ, the more dispersed are the τ_b^*. This makes $\hat{P}^*(\tau)$ in (22) larger than it should be. In the most extreme case of just one treated cluster, rejection frequencies may be essentially zero. In this case, the bootstrap distribution is often bimodal (MacKinnon and Webb 2017a, Figure 4), so that plotting it can provide a useful diagnostic. When there are few treated clusters and the WCR bootstrap P value seems suspiciously large, it may be worth trying the ordinary wild restricted (WR) bootstrap, which can sometimes work surprisingly well in this context (MacKinnon and Webb 2018).

Because it usually seems to work well and is often remarkably inexpensive to compute, we recommend using the WCR bootstrap (preferably with $B \geq 9{,}999$) almost all the time. When G is not too small and the clusters are not too heterogeneous, WCR bootstrap P values and confidence intervals may be quite similar to ones based on the $t(G-1)$ distribution. In that case, it is likely that finite-sample issues are not severe, and there is probably no need to do anything else. When there is a large discrepancy between bootstrap and asymptotic results, however, it makes sense to try other methods as well. These might include alternative bootstrap methods or some of the methods discussed in Section 6. The WCU bootstrap tends to reject more often than the WCR bootstrap, especially in the few-treated case. When the two methods give similar results, it is probably safe to rely on the WCR bootstrap ones.

The WCR bootstrap can sometimes work remarkably well even when G is very small. In fact, Canay et al. (2021) showed that it can yield exact inferences in certain cases where N is large and G is small. These results were obtained by exploiting the relationship between the WCR bootstrap with Rademacher auxiliary random variables and randomization inference (Section 6.2). However, they require rather strong homogeneity conditions on the distribution of the covariates across clusters, as well as limits on the amount of dependence within each cluster similar to those in Bester et al. (2011).

At this point, a word of warning is in order. Almost all of the simulation results for the WCR bootstrap and other bootstrap methods that we have referred to are based on models with one or just a few regressors, and these regressors are typically generated in a fairly simple way. Moreover, almost all existing simulations focus on t-statistics rather than Wald statistics. There is evidence that rejection frequencies for all methods increase as the number of regressors increases (Djogbenou et al. 2019, Section C.2) and that Wald tests are less reliable than t-tests (Pustejovsky and Tipton 2018). Thus it is possible that the WCR bootstrap may perform less well in empirical applications with large numbers of regressors than it has typically done in simulations, especially when there is more than one restriction.

Many regression models include fixed effects at the cluster level. The estimates of these fixed effects are generally not consistent. Even when $N_g \to \infty$, many types of intra-cluster correlation imply that the amount of information about the fixed effects does not increase

without limit. Fixed effects are discussed in Djogbenou et al. (2019, Section 2.2) and Pustejovsky and Tipton (2018). The key idea is to replace the scores $\boldsymbol{s}_g$ in (3) and the empirical scores $\hat{\boldsymbol{s}}_g$ in (5) by ones where the fixed effects have been partialed out. This is very often desirable for computational reasons anyway, especially when the number of clusters is large. Conveniently, `boottest` works with the `areg`, `xtreg`, and `reghdfe` estimators in `Stata`, which provide efficient methods for estimating regression models with fixed effects.

When G is small, the WCR bootstrap encounters an important practical problem. For the Rademacher distribution, or any other two-point distribution, the number of possible bootstrap samples is just 2^G. Webb (2014) proposed a six-point distribution which largely solves this problem, because 6^G is much larger than 2^G. This distribution seems to work almost as well as Rademacher for most values of G, and sometimes much better when G is very small. Whenever either 2^G (for Rademacher) or 6^G (for six-point) is smaller than the chosen value of B, we can enumerate all possible bootstrap samples instead of drawing them at random. For example, when $G = 16$, there are just $65,536$ Rademacher bootstrap samples to enumerate (one of which is identical to the actual sample). This eliminates simulation randomness from the bootstrap procedure. Note that `boottest` uses enumeration by default whenever B is greater than the number of possible bootstrap samples.

6 Other Inferential Procedures

Bootstrap methods are not the only way to obtain inferences more accurate than the ones given by tests and confidence intervals based on cluster-robust t-statistics and cluster-robust Wald statistics that use CV_1. Numerous other procedures have been proposed, which broadly fall into two categories. We discuss several of these in the following two subsections.

6.1 Alternative CRVEs and Critical Values

Because standard errors based on CV_1 tend to be too small, as do critical values based on the $t(G-1)$ distribution, it is natural to seek better ways of computing standard errors and/or critical values. Several related approaches have been proposed.

The residual vectors $\hat{\boldsymbol{u}}_g$ are not always good estimators of the disturbance vectors $\boldsymbol{u}_g$. CV_1 attempts to compensate for this by including a degrees-of-freedom factor. An alternative CRVE, proposed by Bell and McCaffrey (2002), incorporates a more sophisticated way to do so. It omits the scalar factor in CV_1 and replaces the empirical score vectors $\hat{\boldsymbol{s}}_g$ by modified

score vectors $\ddot{\boldsymbol{s}}_g$ that use transformed residuals. The alternative CRVE is

$$\text{CV}_2\text{:} \qquad (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\Big(\sum_{g=1}^{G} \ddot{\boldsymbol{s}}_g \ddot{\boldsymbol{s}}_g^\top\Big)(\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{29}$$

In the middle factor here,

$$\ddot{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \boldsymbol{M}_g^{-1/2} \hat{\boldsymbol{u}}_g, \quad \text{where} \quad \boldsymbol{M}_g = \mathbf{I}_{N_g} - \boldsymbol{X}_g(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}_g^\top. \tag{30}$$

Thus $\boldsymbol{M}_g$ is the diagonal block corresponding to the g^{th} cluster of the projection matrix $\boldsymbol{M}_{\boldsymbol{X}}$, which satisfies $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, and $\boldsymbol{M}_g^{-1/2}$ is its inverse symmetric square root. The $\ddot{\boldsymbol{s}}_g$ should generally provide better approximations to the $\boldsymbol{s}_g$ than do the $\hat{\boldsymbol{s}}_g$, because the $\ddot{\boldsymbol{s}}_g$ compensate for the tendency of the $\hat{\boldsymbol{s}}_g$ to be too small for clusters with high leverage; see Section 4.4, and notice that $\boldsymbol{M}_g = \mathbf{I}_{N_g} - \boldsymbol{H}_g$.

From a theoretical perspective, the CV_2 estimator (29) is quite attractive. It generalizes the HC_2 estimator discussed in MacKinnon and White (1985) and reduces to the latter when all the N_g are equal to 1. If the variance matrix of $\boldsymbol{u}_g$ were proportional to an identity matrix, CV_2 would actually be unbiased (Pustejovsky and Tipton 2018). This suggests that CV_2 will very often be a better estimator than CV_1 in finite samples. For completeness, we note that there is also a CRVE called CV_3, which differs from CV_2 only by using $\boldsymbol{M}_g^{-1}$ instead of $\boldsymbol{M}_g^{-1/2}$ in (30). This closely approximates the jackknife variance matrix that would be obtained by omitting one cluster at a time. When CV_2 is unbiased, CV_3 is biased upwards.

Although it has some attractive features, CV_2 seems to be used only rarely in practice. The problem is the need to compute the $\boldsymbol{M}_g$ matrices and their inverse symmetric square roots. When the N_g are large, simply storing an $N_g \times N_g$ matrix for each cluster may be challenging. In such cases, inverting these matrices and taking their symmetric square roots (which involves finding their eigenvalues and eigenvectors) can be extremely time-consuming, and perhaps even numerically unstable (MacKinnon and Webb 2018). Calculating CV_2 in the usual way is impossible whenever the regression includes cluster-level fixed effects, because the $\boldsymbol{M}_g$ matrices are singular in that case. For such models, it is necessary to partial out the fixed effects, as discussed in Section 4.4 and in Pustejovsky and Tipton (2018).

In addition to proposing the use of CV_2 instead of CV_1, Bell and McCaffrey (2002) suggested a method based on what is called a "Satterthwaite approximation" for calculating the degrees of freedom for t-tests. This is done under the assumption that the variance matrix of $\boldsymbol{u}$ is proportional to an identity matrix. Note that the degrees of freedom are different for every hypothesis to be tested.

Imbens and Kolesár (2016) proposed a similar procedure under the alternative assumption that the variance matrix of $\boldsymbol{u}$ corresponds to a random-effects model; see Section 3.3. Such

a model implies that the disturbances within each cluster are equi-correlated, with intra-cluster correlation ρ that must be estimated from the residuals. When fixed effects are partialed out, doing so absorbs any random effects, making this approach inapplicable.

Young (2016) proposed a related method that uses CV_1 instead of CV_2. It involves two steps. In the first step, the CV_1 standard error for the coefficient of interest is multiplied by a factor greater than one. In the second step, a degrees-of-freedom (d-o-f) parameter is calculated. The standard error and the d-o-f parameter can then be used to compute either a t-statistic, and its P value, or a confidence interval. When some or all of the N_g are large, Young's procedure can be much less computationally demanding than the ones that use CV_2.

The three procedures just discussed are described in some detail in MacKinnon and Webb (2018, Appendix B). Limited simulation evidence suggests that the performance of Young's method is similar to those of the two much more costly methods based on CV_2. However, this evidence is by no means definitive, because the simulations focused on only a narrow set of treatment models.

Using Hotelling's T^2 distribution with estimated degrees of freedom, Pustejovsky and Tipton (2018) generalized the procedure of Bell and McCaffrey (2002) to the case of Wald tests based on (16). Their simulation evidence suggests that the resulting tests always reject less often than standard Wald tests. They rarely over-reject but often under-reject, and they sometimes do so quite severely. The `clubSandwich` package for both `R` and `Stata` implements the procedures discussed in Pustejovsky and Tipton (2018).

Although the procedures discussed in this subsection have some theoretical appeal and seem to work well in many cases, we are not aware of any evidence that they outperform the WCR bootstrap over a wide range of models and DGPs. One limitation is that the Student's t distribution is not very flexible. Even when the d-o-f parameter is estimated very well, the best we can hope for is that a test based on this distribution will be accurate for some level of interest. It may well over-reject at some levels and under-reject at others.

Our recommendation is to use the methods dealt with in this subsection primarily to confirm (or perhaps cast doubt on) the results of the WCR bootstrap when there is concern about the reliability of the latter. Cases of particular concern are ones with few but balanced clusters (say, $G \leq 12$), ones with balanced but few treated clusters (say $G_1 \leq 7$), ones with seriously unbalanced cluster sizes, and ones with any other sort of heterogeneity that causes a few clusters to be influential or to have high leverage; see Section 4.4.

6.2 Randomization Inference

Randomization inference (RI) was proposed by Fisher (1935) as a distribution-free way to perform tests in the context of experiments. Lehmann and Romano (2005, Chapter 15) gives a formal introduction, and Imbens and Rubin (2015, Chapter 15) provides a more accessible discussion. The key idea of RI is to compare an outcome that is actually observed with a set of outcomes that might have been observed if treatment had been assigned differently. The outcome, say τ, could be a sample average, a coefficient estimate, or some other statistic.

Specifically, consider a clustered regression model with treatment at the cluster level, which could be a DiD model where only some observations within the treated clusters receive treatment. Then τ might be the average treatment effect on some measure of outcome. Suppose there are G clusters, G_1 of which received treatment. The number of ways in which treatment could have been assigned to G_1 out of the G clusters is

$$ {}_GC_{G_1} = \frac{G!}{G_1!(G-G_1)!}. \tag{31} $$

One of these corresponds to the actual assignment, and the remaining $S = {}_GC_{G_1} - 1$ are called re-randomizations. Each re-randomization involves pretending that a particular set of G_1 clusters was treated, with the remaining $G - G_1$ serving as controls. The values of the dependent variable do not change across re-randomizations, only the values of the treatment dummy. For every re-randomization, indexed by j, we could calculate a test statistic τ_j^*. If the observable characteristics of the clusters were all the same, it would make sense to compare τ with the empirical distribution of the τ_j^*. For example, we could calculate the P value for an upper-tail test as either

$$ P_1^*(\tau) = \frac{1}{S}\sum_{j=1}^{S} \mathbb{I}(\tau_j^* \geq \tau) \quad \text{or} \quad P_2^*(\tau) = \frac{1}{S+1}\left(1 + \sum_{j=1}^{S} \mathbb{I}(\tau_j^* \geq \tau)\right). \tag{32} $$

Here P_2^* implicitly includes the actual assignment to treatment, and P_1^* omits it.

When αS is an integer for a test at level α, both P values in (32) yield the same result, and the test is exact if the distributions of the τ_j^* are the same as that of τ. However, P_1^* and P_2^* can differ noticeably when S is small. The latter is more conservative and seems to be more popular. When S is not too large, it is often easy to calculate all of the τ_j^*, but this is infeasible even when G is not particularly large. In such cases, we choose a number of re-randomizations, say $B = 99{,}999$, at random. In principle, these should be drawn without replacement, but that is not important if B is small relative to S.

It seems to be widely believed that RI tests are always exact. This is not true. When treatment is not assigned at random, or when the observed characteristics of the treated

clusters differ systematically from those of the controls, we cannot expect the distributions of τ and the τ_j^* to coincide.

For DiD models, Conley and Taber (2011) proposed a method that is very similar to RI based on the least squares estimate of the coefficient on a treatment dummy. It also described a way to obtain confidence intervals by inverting the RI P values. MacKinnon and Webb (2020a) studied the coefficient-based RI procedure for DiD models (called RI-β) and a similar one in which inference is based on cluster-robust t-statistics (called RI-t). Both procedures work very well when the clusters are homogeneous, even when $G_1 = 1$. However, when the clusters differ in size, and the treated clusters are systematically larger or smaller than average, neither RI-β nor RI-t works well, and it appears that G may have to be quite large (much larger than for the WCR bootstrap) before either of them work well.

There is evidently a close relationship between RI and the wild cluster bootstrap. Evaluating all possible bootstrap samples by enumeration is quite similar to evaluating all possible re-randomizations. More importantly, the results of Canay et al. (2021) and MacKinnon and Webb (2020a) strongly suggest that homogeneity across clusters is more important for RI than for the WCR bootstrap. For the former, the regressors change as we compute each of the τ_j^*, but the regressand stays the same. For the latter, the regressand changes as we compute each of the τ_b^*, but the regressors stay the same. Thus, the empirical distribution to which τ is being compared is conditional on the actual regressors (including the clusters actually treated) for the WCR bootstrap, but not for the RI procedures discussed above.

The RI approach discussed above is not the only one. Canay, Romano and Shaikh (2017) proposed RI tests based on the cluster-level estimators of Ibragimov and Müller (2010); see Cai, Canay, Kim and Shaikh (2021) for a guide to their approach. Because cluster-level estimation is required, every cluster must include both treated and untreated observations in models for treatment effects. This can be accomplished by merging clusters, but at the cost of making G smaller. For the test to have reasonable power, it is desirable that $G \geq 8$, so it is not useful for cases with very few treated clusters. Hagemann (2019a, b) developed RI tests that can be used even when G is quite small and there is substantial heterogeneity across clusters. These tests do not require cluster-level estimation, but G_1 and $G - G_1$ should both be no less than 4. Spamann (2019) developed an RI procedure based on RI-t for the case where one cluster is much larger than any of the others.

An alternative way to deal with the problem of one or very few treated clusters, which can involve an RI-like procedure, is the method of synthetic controls surveyed in Abadie (2021).

7 At What Level Should We Cluster?

Choosing the right level at which to cluster is not always easy, and choosing the wrong level can have serious consequences. Suppose there are two possible levels of clustering, coarse and fine. With G coarse clusters, the middle matrix in (4) is $\sum_{g=1}^{G} \boldsymbol{\Sigma}_g$. If we assume that each coarse cluster contains M_g fine clusters indexed by h, then $\boldsymbol{\Sigma}_g$ can be written as

$$\boldsymbol{\Sigma}_g = \sum_{h_1=1}^{M_g} \sum_{h_2=1}^{M_g} \boldsymbol{\Sigma}_{g,h_1h_2}, \tag{33}$$

where $\boldsymbol{\Sigma}_{g,h_1h_2}$ denotes the covariance of the scores for fine clusters h_1 and h_2 within coarse cluster g. Under the assumption of fine clustering, $\boldsymbol{\Sigma}_{g,h_1h_2} = \boldsymbol{\Sigma}_{gh}$ whenever $h_1 = h_2 = h$, and it equals zero whenever $h_1 \neq h_2$. In the latter case, the middle matrix in (4) is $\sum_{g=1}^{G} \sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh}$.

From (33), the difference between the middle matrices for coarse and fine clustering is

$$\sum_{g=1}^{G} \boldsymbol{\Sigma}_g - \sum_{g=1}^{G} \sum_{h=1}^{M_g} \boldsymbol{\Sigma}_{gh} = 2 \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2=h_1+1}^{M_g} \boldsymbol{\Sigma}_{g,h_1h_2}. \tag{34}$$

Under the assumption of fine clustering, all of the terms on the right-hand side of (34) are equal to zero. Under the assumption of coarse clustering, all of these terms must be estimated. If we cluster at the fine level when coarse clustering is appropriate, the CRVE is inconsistent. On the other hand, if we cluster at the coarse level when fine clustering is appropriate, the CRVE has to estimate what is often a very large number of terms that actually equal zero. This makes the CRVE less accurate than it should be, leads to loss of power, and may well lead to poor inferences in finite samples, especially when the number of coarse clusters is small.

The consequences of clustering at an incorrect level were investigated by simulation in MacKinnon and Webb (2020b). Under-clustering (that is, clustering at too fine a level) generally led to serious over-rejection, which became worse as the sample size increased with the numbers of clusters at all levels held constant. This is exactly what we would expect; see the discussion following (7). Over-clustering (that is, clustering at too coarse a level) also led to some over-rejection, which could be fairly serious when using the $t(G-1)$ distribution but much less so when using the WCR bootstrap. Power also decreased when the level used was coarser than necessary. Thus, in the setup of these simulations, the consequences of over-clustering are noticeable but not severe.

At least three rules of thumb are commonly suggested as ways to choose the right level of clustering. The simplest is just to cluster at the coarsest feasible level (Cameron and Miller 2015, Section IV). A second rule of thumb is to cluster at whatever level yields the largest

standard error(s) for the coefficient(s) of interest (Angrist and Pischke 2008, Section 8.2). This rule will often lead to the same outcome as the first one, although it is more robust to the problem of few treated clusters (Section 4.2.2), which can cause standard errors to become smaller as the level of clustering becomes coarser. As the discussion above makes clear, these two rules are conservative and are likely to avoid the severe over-rejection associated with under-clustering, but they can lead to poor finite-sample properties and loss of power (or, equivalently, confidence intervals that are unnecessarily long).

The third widely-recommended rule of thumb is to cluster at the level at which treatment is assigned for models that estimate treatment effects (Bertrand et al. 2004). When treatment is assigned by cluster, treatment dummies vary only at the cluster level (or, in the DiD case, only within some clusters). This implies that, whenever there is any intra-cluster correlation of the disturbances, the scores will be correlated within the treated clusters; see the discussion between (6) and (7). From the perspective of both the model-based and design-based approaches (Section 3.1), it would never make sense to cluster at a level lower than the one at which treatment is assigned. However, from the perspective of the model-based approach, it may well be desirable to cluster at a higher level (for example, at the school level rather than the classroom level even when treatment is assigned by classroom). In this case, knowing how treatment was assigned tells us the finest level at which it makes sense to cluster, but it does not tell us whether it is appropriate to cluster at a coarser level.

7.1 Testing for the Correct Level of Clustering

MacKinnon, Nielsen and Webb (2020) proposed what they call score-variance tests for the correct level of clustering. These are based on comparing the variance of the scores for two different levels of nested clusters, say "fine" and "coarse." The idea is to construct a test statistic based on the empirical analog of (34), which is a function of the empirical scores. Score-variance tests are most easily computed when there is just one coefficient. The model can be put into this form by partialing out all the regressors except the one of interest. The empirical analog of (34), divided by the square root of an estimate of its variance, is then asymptotically distributed as $\mathrm{N}(0,1)$. MacKinnon et al. (2020) also proposed wild (cluster) bootstrap implementations of these tests. Bootstrapping can sometimes greatly improve their finite-sample properties, especially when there are fixed effects at the level of the coarse clusters. When score-variance tests reject the null hypothesis, investigators should cluster at the coarse level. When it is not rejected, they may choose to cluster at the fine level.

Score-variance tests directly test the level at which the scores are clustered. The only other test for the level of clustering of which we are aware is an ingenious but indirect one

proposed in Ibragimov and Müller (2016). Their test requires the model to be estimated separately for every coarse cluster, something that is not possible when the regressor of interest is invariant within some clusters, which is very often the case for treatment models and DiD models. The test statistic compares the observed variation of the estimates across clusters with an estimate of what that variation would be if clustering were actually at a finer level. But it is invalid if the parameter β has different meanings for different clusters. This will happen whenever a model includes fixed effects for categorical variables and not all categories are observed in every cluster.

Choosing the level of clustering based on a test is a form of pre-testing. As is well known, pre-testing and other forms of model selection can lead to estimators with undesirable properties and distributions that are poorly approximated by asymptotic theory (Leeb and Pötscher 2005). In this case, however, the estimator $\hat{\boldsymbol{\beta}}$ remains the same whatever the outcome of a score-variance test. Only the variance matrix estimator that is employed depends on that outcome. Thus the usual objections to pre-testing do not quite apply. Simulations in MacKinnon et al. (2020) suggest that, while it would of course be better to know the actual level of clustering, choosing the level based on a score-variance test can lead to improved inference compared with simply using a rule of thumb.

8 Empirical Example

We illustrate many of our recommendations by revisiting a long-standing empirical question in labor economics, namely, the impact of the minimum wage on young people. In the past two decades, many U.S. states have significantly increased their minimum wages. In fact, from 2000 to 2019, every state increased the nominal minimum wage by at least 27%, and six states doubled it. Moreover, recent proposals to increase the national minimum wage to $15 per hour, including by President Biden, have reinvigorated the debate about the effects of the minimum wage.

Some classic references on the impact of the minimum wage are Mincer (1976) and Card and Krueger (1994). The latter paper was among the very first and most influential applications of the DiD methodology, which continues to be used in this literature. For example, Jardim et al. (2017) used a DiD analysis to study the effects of a large increase in the minimum wage in Seattle. Wolfson and Belman (2019) and Neumark and Shirley (2021) have surveyed many recent studies on the impacts of minimum wages. Both conclude that the majority of studies, but not all, find dis-employment effects that are concentrated among teenagers and those with low levels of education. Manning (2021) explores why the evidence on employment effects is mixed.

Instead of using a DiD approach, we exploit state-level differences in the minimum wage and analyze their impacts on labor-market and education outcomes at the individual level. Although we treat the minimum wage as exogenous in our analysis, we hesitate to call our estimates causal. There is reason to believe that state-level minimum wages may be endogenous, because states may be more likely to increase them during good economic times. However, we ignore this issue, because our principal objective is to illustrate the importance of clustering for statistical inference.

The model we estimate is

$$y_{ist} = \alpha + \beta \mathrm{mw}_{st} + \boldsymbol{Z}_{ist}\boldsymbol{\gamma} + \mathrm{year}_t\,\boldsymbol{\delta}_t + \mathrm{state}_s\,\boldsymbol{\delta}_s + u_{ist}. \tag{35}$$

Here y_{ist} is the outcome of interest for person i in state s in year t. There are three outcome variables. "Hours" records the usual hours worked per week, which is not defined for unemployed individuals. "Employed" is a binary variable equal to 1 if person i is employed and to 0 if they are either unemployed or not in the labor force. "Student" is a binary variable equal to 1 if person i is currently enrolled in school and to 0 otherwise. The parameter of interest is β, which is the coefficient on the minimum wage in state s at time t, denoted mw_{st}. The row vector $\boldsymbol{Z}_{ist}$ collects a large set of individual-level controls, including race, gender, age, and education. There are also year and state fixed effects. Neumark and Wascher (2007) estimate models similar to (35) with individual-level data and cluster at the state level.

Data at the individual level from the American Community Survey (ACS) were obtained from IPUMS (Ruggles et al. 2020) and cover the years 2005–2019. The minimum wage data were provided by Neumark (2019), and we have collapsed them to state-year averages to match the ACS frequency. Following previous literature, we restrict attention to teenagers aged 16–19. We keep only individuals who are "children" of the respondent to the survey and who have never been married. We drop individuals who had completed one year of college by age 16 and those reporting in excess of 60 hours usually worked per week. We also restrict attention to individuals who identify as either black or white.

We consider six different clustering structures that lead to six different estimated standard errors for $\hat{\beta}$. These are no clustering (with HC_1 standard errors), one-way clustering at either the state-year, state, or region level (with CV_1 standard errors), and two-way clustering by state and year or by region and year (also with CV_1 standard errors).[1] Early empirical research on the impacts of the minimum wage would have used either conventional or HC_1

[1]Regions are defined as the U.S. Census Divisions, with the following partitioning of states. New England: CT, MA, MN, NH, RI, VT; Middle Atlantic: NJ, NY, PA; South Atlantic: DC, DE, FL, GA, MD, NC, SC, VA, WV; East South Central: AL, KY, MS, TN; East North Central: IL, IN, MI, OH, WI; West North Central: IA, KS, MN, MO, ND, NE, SD; West South Central: AR, LA, OK, TX; Mountain: AZ, CO, ID, MT, NM, NV, UT, WY; Pacific: AK, CA, HI, OR, WA.

Table 1: Summary Statistics for Cluster Heterogeneity

Clustering	G	G_1^*	G_0^*	$\bar{N}_g$	minimum	1st quartile	median	3rd quartile	maximum
Hours data: $N = 492{,}827$									
State-year	765	37.3	79.4	644	6	176	480	860	3,052
State	51	4.5	16.2	9,663	258	2,495	7,082	13,481	35,995
Year	15	5.6	6.6	32,855	28,262	28,839	30,733	40,224	40,394
Region	9	3.3	7.5	54,759	27,849	37,396	50,489	65,389	96,337
Employment and student data: $N = 1{,}531{,}360$									
State-year	765	25.4	66.0	2,002	42	524	1,413	2,426	10,794
State	51	2.6	13.1	30,027	927	7,363	22,845	37,020	144,914
Year	15	6.0	6.5	102,091	92,701	95,589	102,319	108,858	110,528
Region	9	2.0	7.0	170,151	74,172	104,703	181,767	208,099	291,955

Notes: The values of G^* were calculated using 28 regressors after the state dummies had been partialed out. G_1^* and G_0^* use $\rho = 1$ and $\rho = 0$, respectively. See MacKinnon et al. (2021a).

standard errors, but modern studies would almost always cluster at some level.

The design-based approach (Section 3.1) can be interpreted as saying that state-year clustering is appropriate. Since the minimum wage is invariant within state-year clusters, no finer level should be considered (Abadie et al. 2017, Corollary 2). Furthermore, because treatment effects are probably not heterogeneous across years, and (35) includes state fixed effects, no higher level is appropriate (Abadie et al. 2017, Section 4). However, these arguments may not apply, because the minimum wage is not randomly assigned by state-year.

After a state has increased its minimum wage, it almost always remains at the new level until it is increased again. This implies that minimum wages must be correlated across years within each state. Unless the disturbances happen to be uncorrelated across years within states, the scores will therefore be correlated, which suggests that state-level clustering may be appropriate, as it accounts for both the invariance of the minimum wage at the state-year level and for within-state correlations. We also consider region-level clustering based on the nine census divisions because there may be correlations among nearby states. Finally, largely for completeness, we consider two-way clustering by either state or region and year.

In Table 1, we present a number of summary statistics for cluster-size heterogeneity. Specifically, we report the number of clusters, G, two variants of the effective number of clusters, G_1^* and G_0^* (Section 4.2.1), as well as the average, minimum, first and third quartiles, median, and maximum of the N_g. These statistics suggest that the state-year clusters are extremely unbalanced. Although there are $G = 765$ clusters, the effective numbers G_1^* and G_0^* are much smaller than that, by factors that range from nearly 10 to over 30. The maximum

Table 2: Score-Variance Tests for Level of Clustering

	Hours	Employed	Student	Hours	Employed	Student
	H_{none} vs. $H_{state\text{-}year}$			$H_{state\text{-}year}$ vs. H_{state}		
$\hat{\tau}$ statistic	18.2723	10.1108	2.8558	7.9453	2.2800	3.1105
P value, asymptotic	0.0000	0.0000	0.0021	0.0000	0.0113	0.0009
P value, bootstrap	0.0000	0.0000	0.0026	0.0000	0.0227	0.0085
	H_{none} vs. H_{state}			$H_{state\text{-}year}$ vs. H_{region}		
$\hat{\tau}$ statistic	38.9394	10.9183	4.6608	11.2952	2.5226	1.2701
P value, asymptotic	0.0000	0.0000	0.0000	0.0000	0.0058	0.1020
P value, bootstrap	0.0000	0.0000	0.0007	0.0000	0.0279	0.1388
	H_{none} vs. H_{region}			H_{state} vs. H_{region}		
$\hat{\tau}$ statistic	49.6559	9.4096	2.7201	2.2566	0.6757	−0.3206
P value, asymptotic	0.0000	0.0000	0.0033	0.0120	0.2496	0.6257
P value, bootstrap	0.0120	0.0000	0.0000	0.0410	0.2662	0.5497

Notes: There are 765 state-year clusters, 51 state clusters, and 9 region clusters. The test statistic $\hat{\tau}$ is asymptotically distributed as $\mathrm{N}(0,1)$. The bootstrap P values were calculated with $B = 99,999$.

cluster size is over six times the median, and the third quartile is nearly twice the median.

From Table 1, we also see that the state clusters are extremely unbalanced, based on their sample sizes, the values of G_1^*, and (to a somewhat lesser extent) the values of G_0^*. The region clusters are fairly balanced in terms of their sample sizes, with the maximum N_g only about four times the minimum. The values of G_1^* suggest that they are seriously unbalanced, but the values of G_0^* are less worrisome. For both states and regions, we find G_0^* more plausible than G_1^*, because the average intra-cluster correlation is evidently much closer to 0 than to 1; see Table 3. The year clusters (which we only use in two-way clustering) are very balanced in terms of their sample sizes but less so in terms of both G^* variants. We would expect asymptotic theory to perform rather poorly in all cases, because there is a lot of cluster heterogeneity for clustering by state-year and state, a small number of clusters for clustering by region, and a small number of effective clusters for clustering by year.

We next apply the score-variance tests described in Section 7.1 to test for the appropriate level of clustering.[2] Table 2 presents results from six tests for each of the three models. We report results from tests of the null of no clustering (independence) against alternatives of state-year, state, and region clustering, as well as state-year vs. state, state-year vs. region, and state vs. region. Following a systematic, sequential testing approach, we would test

[2]We did not calculate the tests proposed in Ibragimov and Müller (2016), because (35) cannot be estimated on a cluster-by-cluster basis for state-level or state-year clusters.

independence vs. state-year, then state-year vs. state, and finally state vs. region, and (apart from the possibility of two-way clustering) we would conclude that the appropriate level of clustering is that of the first non-rejected null hypothesis.

The null hypothesis of no clustering is strongly rejected for all three models. The null of state-year clustering is also rejected against the alternative of state clustering for all three models, and the rejections are very strong except perhaps when the regressand is employment status. This is interesting, because, as we noted earlier in this section, the design-based approach (Section 3.1) seems to suggest state-year clustering.

This brings us to the null hypothesis of state clustering. For the hours model, the bootstrap P value is 0.041, which is quite different from the asymptotic P value of 0.012. For the other two models, the null hypothesis of state clustering never comes close to being rejected by either asymptotic or bootstrap tests, which yield similar results. Thus, for the hours model, the score-variance tests marginally favor region clustering over state clustering, but, for the employed and student models, they clearly favor state clustering.

Table 3 presents the estimates of β from regression (35) for all three regressands, along with t-statistics and P values for each of the clustering levels considered. The coefficients are $\hat{\beta} = -0.1539$ for hours, $\hat{\beta} = -0.00367$ for employed, and $\hat{\beta} = 0.00221$ for student. Under the surely false assumption of no clustering, all of these coefficients are extremely significant, with P values below 0.0000. We do not compute bootstrap P values, because it would be prohibitively costly, and they must all be very close to zero.

When we instead follow the design-based approach and cluster at the state-year level, the t-statistics become smaller, especially for the employed model. Nevertheless, clustering at this level still leads us to conclude that all three coefficients are significant, with bootstrap P values ranging from 0.0005 to 0.0145. However, our conclusions change radically when we cluster at the state or region levels. For all three outcome variables, the t-statistics become smaller, and the P values (especially the bootstrap ones) become larger. The bootstrap P values often differ substantially from the ones based on the $t(G-1)$ distribution, which is expected given that the clusters are either very unbalanced or small in number (Table 1). Clustering by region always yields larger bootstrap P values than clustering by state.

Based on the results in Table 2, it seems appropriate to cluster at the state level for the employed and student models, and at either the state or region level for the hours model. When we do this, we find from Table 3 that an increase in the minimum wage is associated with a decrease in employment, but the coefficient is not significant at any conventional level. It is also associated with an increased probability of being a student, which is significant at the 5% level regardless of clustering level. For hours worked, the coefficient is negative. It is significant for state clustering and almost significant for region clustering.

Table 3: Estimated Impact of the Minimum Wage

Clustering level		Hours	Employed	Student
	$\hat{\beta}$	−0.1539	−0.0037	0.0022
None: HC_1	t-statistic	−5.4469	−5.2801	4.9719
	P value, $N(0,1)$	0.0000	0.0000	0.0000
State-year: CV_1	t-statistic	−3.3823	−2.6492	4.0776
	P value, $t(764)$	0.0008	0.0082	0.0001
	P value, WCR	0.0027	0.0145	0.0005
State: CV_1	t-statistic	−2.4696	−1.3679	2.9780
	P value, $t(50)$	0.0170	0.1775	0.0045
	P value, WCR	0.0362	0.2141	0.0238
Region: CV_1	t-statistic	−2.2478	−1.0230	3.1743
	P value, $t(8)$	0.0548	0.3362	0.0131
	P value, WCR	0.0527	0.3826	0.0430
State & year: two-way CV_1	t-statistic	−2.5197	−1.4776	3.4443
	P value, $t(14)$	0.0245	0.1617	0.0039
	P value, WCR (year)	0.1281	0.2148	0.0034
Region & year: two-way CV_1	t-statistic	−2.2842	−1.0999	3.5766
	P value, $t(8)$	0.0517	0.3034	0.0072
	P value, WCR (region)	0.0736	0.3711	0.0367

Notes: There are 765 state-year clusters, 51 state clusters, and 9 region clusters, with 492,827 observations in the hours dataset and 1,531,360 observations in the employed and student dataset. The bootstrap dimension for two-way clustering is given in parentheses. In most cases, WCR bootstrap P values are calculated with $B = 99,999$ using the Rademacher distribution. When bootstrapping by region, they are calculated using Webb's six-point distribution. When bootstrapping by year, they are calculated by enumeration using the Rademacher distribution, so that $B = 32{,}768$. Obtaining all the results in this table took 6 minutes and 12 seconds using Stata 16 and boottest 2.5.3 on one core of an Intel i9 processor running at 3.6 GHz. The data and programs are available at http://qed.econ.queensu.ca/pub/faculty/mackinnon/guide/

The table also reports asymptotic and bootstrap t-statistics and P values for two-way clustering, either by state and year or by region and year. The bootstrap method we use combines the one-way WCR bootstrap with the two-way variance matrix (14). This can be done in two different ways, corresponding to each of the two clustering dimensions. Based on simulation evidence in MacKinnon et al. (2021b), we bootstrap by the dimension with the smallest number of clusters. However, two-way clustering does not change most of the results very much, so there appears to be little reason to use it in this case.

We tentatively conclude that an increase in the minimum wage is associated with a significant decrease in hours worked and a significant increase in the likelihood of being a

student. Jardim et al. (2017) find a similar reduction in hours following a minimum wage increase in Seattle, and the effects of the minimum wage on school enrollment are discussed in Neumark and Wascher (1995). For employment, we obtain a small negative effect, but it is not close to being statistically significant when we cluster at either the state or region levels. Thus our results are consistent with, and add support for, the conclusions of Manning (2021) about the "elusive" employment effect of the minimum wage.

8.1 Leverage and Influence

As we discussed in Section 4.4, investigators should be suspicious of results that are overly dependent on very few clusters. We therefore calculated measures of influence and leverage for each of the three levels of one-way clustering. With respect to leverage, it appears that, for both data sets, there is no heterogeneity in the clusters other than what is implied by cluster sizes. The $\text{Tr}(\boldsymbol{H}_g)$ are essentially proportional to cluster sizes.

For the student model with any level of clustering, we find no evidence of influential clusters. For the hours and employed models with state-year clustering, there is also not much to suggest that any clusters are influential, with only about a 20% difference between the smallest and largest of the 765 values of $\hat{\beta}^{(g)}$. On the other hand, for the employment model with state clustering, there are two noticeable values of $\hat{\beta}^{(g)}$: North Dakota has $\hat{\beta}^{(g)} = -0.0053$ and Colorado has $\hat{\beta}^{(g)} = -0.0020$, with $\hat{\beta} = -0.0037$ and all the remaining $\hat{\beta}^{(g)}$ in the interval $[-0.0046; -0.0030]$. For comparison, the standard error of $\hat{\beta}$ is 0.0027.

For the hours model, four values of the $\hat{\beta}^{(g)}$ stand out. They are Illinois (-0.1776) and North Dakota (-0.1770) at one end and Ohio (-0.1274) and Colorado (-0.1227) at the other. For comparison, $\hat{\beta} = -0.1539$, and all the other $\hat{\beta}^{(g)}$ lie in the interval $[-0.1713; -0.1406]$. In this case, the standard error of $\hat{\beta}$ is 0.0623.

For region-level clustering, the $\hat{\beta}^{(g)}$ vary somewhat, but, with only nine regions, it is difficult to determine whether any clusters are influential. We tentatively conclude that, no matter what level we cluster at, no clusters are so influential as to be a cause for concern.

8.2 Placebo Regressions for the Empirical Example

As we discussed in Section 3.2, placebo regressions provide a simple way to check the level at which the residuals are clustered, even when the pattern of intra-cluster correlation is unknown and perhaps very complicated. If a placebo regressor is clustered at, say, the state level, then the empirical scores will also be clustered at that level unless the residuals are clustered only at a finer level. When the residuals display intra-cluster correlation at the state level, we would expect placebo regressions with standard errors clustered at that

Table 4: Rejection Percentages for Placebo Regressions

	DiD-type treatment						State-level AR(1) component					
Method	Hours		Employed		Student		Hours		Employed		Student	
HC_1, N(0, 1)	31.1	48.6	48.4	62.6	40.0	43.5	37.8	26.0	54.3	44.4	28.6	21.7
State-year												
CV_1, $t(764)$	23.5	30.0	30.3	34.8	28.1	37.1	16.4	16.3	21.1	25.2	14.6	16.2
CV_1, WCR	19.7	25.5	25.0	29.8	26.5	33.0	14.7	15.5	18.9	23.6	13.4	15.5
State												
CV_1, $t(50)$	6.8	14.7	7.7	14.2	7.1	20.7	7.5	6.6	9.0	8.3	6.5	6.9
CV_1, WCR	3.5	6.0	2.0	5.1	5.1	9.4	5.2	4.9	5.2	5.6	5.1	5.3
Region												
CV_1, $t(8)$	7.5	11.0	7.3	11.6	7.9	10.8	7.6	7.2	8.2	8.0	7.8	7.2
CV_1, WCR	4.4	6.3	2.5	5.7	3.6	6.4	6.3	5.8	6.0	6.2	6.1	5.7

Notes: The numbers are rejection percentages at the nominal 5% level based on 10,000 simulations. For the DiD-type treatment, the first number in each pair is the smallest rejection percentage over all parameter values used to simulate the placebo regressor, and the second is the largest; see text. For the state-level AR(1), the first number is for $\rho = 0.5, \delta = 0.9$, and the second is for $\rho = 0.8, \delta = 0.5$. There are 765 state-year clusters, 51 state clusters, and 9 region clusters. The WCR bootstrap used $B = 999$.

level to reject about as often as they should, and we would expect placebo regressions with standard errors clustered at finer levels to over-reject.

We perform two sets of placebo-regression experiments for each of the three equations estimated in Table 3. In the first set, the placebo regressor is a DiD-style treatment dummy similar to the ones used in Bertrand et al. (2004) and the other papers cited in Section 3.2. Treatment is randomly applied to 5, 10, ..., 45 states. For each state, it begins randomly in any year excluding 2005 (to avoid collinearity with the state fixed effects) and continues through 2019.[3] Rejection percentages are shown in the top panel of Table 4. The first number in each pair is the smallest rejection percentage observed over the nine experiments for each equation, and the second number is the largest one. These often differ substantially. In many cases, the rejection percentages associated with either 5 or 45 treated states are particularly extreme, suggesting that these numbers may be too close to 1 and 51 for reliable inference; see Section 4.2.2.

From the first three columns in Table 4, we see that not clustering, or clustering at the state-year level, always leads to severe over-rejection. There can also be noticeable over-

[3]Of course, this sort of DiD model with two-way fixed effects is somewhat obsolete; see Callaway and Sant'Anna (2021) and other papers cited therein. But we would still expect to find no effects for placebo treatments beyond those attributable to chance.

rejection for clustering at the state and region levels when using the $t(G-1)$ distribution. The results are much better when using the WCR bootstrap, although there can be some under-rejection for the employed model when we cluster at either the state or region levels. This is most severe when 45 states are treated. The WCR also shows some over-rejection for the student model when clustering is at the state level and 40 or 45 states are treated.

For the second set of simulation experiments, the placebo regressor is generated by

$$x_{ist} = \delta v_{st} + (1-\delta)\epsilon_{ist}, \quad v_{st} = \rho v_{s,t-1} + e_{st}, \quad 0 \leq \rho < 1, \quad 0 \leq \delta \leq 1, \tag{36}$$

where the ϵ_{ist} and the e_{st} are independent standard normals. Thus the v_{st} are 51 separate stationary AR(1) processes, and x_{ist} is a weighted average of v_{st} and ϵ_{ist}. When either $\rho = 0$ or $\delta = 0$, the x_{ist} are independent. When both ρ and δ are positive, they are correlated within both state-years and states. They are never correlated across states within regions. Because x_{ist} is assigned at the individual level, the design-based approach (Section 3.1) would suggest that not clustering is appropriate in this context despite any within-state correlation.

The extent to which the x_{ist} are correlated within state-years and states depends on both of the parameters in (36). In order to avoid a figure with several panels, we report rejection percentages for just two cases. In the first case, $\rho = 0.5$ and $\delta = 0.9$, so there is a lot of correlation within each state-year. In the second case, $\rho = 0.8$ and $\delta = 0.5$, so there is less correlation within state-years but more correlation across years within each state.

From the last three columns of Table 4, we see that failing to cluster, or clustering at the state-year level, always leads to serious over-rejection. This over-rejection occurs despite the no-clustering recommendation from the design-based approach. Clustering at the state level yields reliable inferences when the tests are bootstrapped, but there is noticeable over-rejection when they are not. This is to be expected given the unbalanced cluster sizes at the state level. Clustering at the region level always leads to modest over-rejection even when the tests are bootstrapped, which probably reflects the fact that, with just 9 regions, the tests are not entirely reliable.

The placebo-regression results are largely consistent with those of the score-variance tests. They suggest that the results for state-level clustering in Table 3 can probably be relied upon, but that the results for not clustering and for state-year clustering should not be believed.

9 Conclusion: A Summary Guide

We conclude by presenting a brief summary guide. This is essentially a checklist for cluster-robust inference in regression models. The first item should be checked off prior to estimation.

The remaining ones should be dealt with as soon as some estimates are available and kept in mind throughout the process of estimation and inference.

1. List all plausible clustering dimensions and levels for the data at hand. For each of these, report the number of clusters, G, and a summary of the distribution of the cluster sizes (the N_g). We suggest reporting at least the minimum, maximum, mean, and median of the N_g. These could be reported in tabular form, as in Table 1, or graphically, perhaps using box plots.
2. Make an informed decision regarding the clustering structure. This could be based on the design-based approach (Section 3.1), the model-based approach (Section 7), or a combination thereof. The decision can depend on what is to be estimated and why. For the model-based approach, formal tests (Section 7.1) can be helpful in making this decision. In some cases, placebo regressions (Sections 3.2 and 8.2) may also be informative.
3. For the key regression specification(s) considered, report a summary of measures of influential and high-leverage clusters (Sections 4.4 and 8.1). These may be particularly informative for difference-in-differences and other treatment models. Inferences may not be reliable when a few clusters are highly influential or have high leverage. If possible, also report the effective number of clusters, G^* (Section 4.2.1).
4. For models with treatment at the cluster level, where either the treated clusters or the controls are few in number and/or atypical, cluster-robust inference can be quite unreliable (Section 4.2.2). In such cases, it is important to verify that the results are robust, perhaps by using methods based on randomization inference (Section 6.2).
5. Employ the restricted wild cluster (WCR) bootstrap (Section 5.3) as a matter of course for both tests and confidence intervals. When G is large and the clusters are homogeneous, bootstrap and conventional inferences may well coincide. When they do not, the WCR bootstrap will almost always yield more accurate inferences than the $t(G-1)$ distribution. However, even the former may be unreliable in extreme versions of the situations discussed in items 3 and 4.
6. Verify that the empirical results are robust. In standard practice, this means that they are robust with respect to the choice of regressors and fixed effects. In the context of cluster-robust inference, however, it also means that they are robust with respect to changes in the clustering structure (Section 7) and perhaps also to alternative methods of cluster-robust inference (Sections 5 and 6).

References

Abadie, A., 2021. Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of Economic Literature 59, to appear.

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

Andrews, I., Stock, J.H., Sun, L., 2019. Weak instruments in instrumental variables regression: Theory and practice. Annual Review of Economics 11, 727–753.

Angrist, J.D., Pischke, J.S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, New York.

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

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

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics. Wiley, New York.

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

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

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

Cai, Y., Canay, I.A., Kim, D., Shaikh, A.M., 2021. A user's guide to approximate randomization tests with a small number of clusters. ArXiv e-prints 2102.09058.

Callaway, B., Sant'Anna, P.H., 2021. Difference-in-differences with multiple time periods. Journal of Econometrics 223, to appear.

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

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

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

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

Canay, I.A., Santos, A., Shaikh, A., 2021. The wild bootstrap with a 'small' number of 'large' clusters. Review of Economics and Statistics 103, to appear.

Card, D., Krueger, A., 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. American Economic Review 84, 772–793.

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

Chatterjee, S., Hadi, A.S., 1986. Influential observations, high-leverage points, and outliers in linear regression. Statistical Science 1, 379–416.

Chiang, H.D., Kato, K., Ma, Y., Sasaki, Y., 2021. Multiway cluster robust double/debiased machine learning. Journal of Business & Economic Statistics 39, to appear.

Chiang, H.D., Kato, K., Sasaki, Y., 2020. Inference for high-dimensional exchangeable arrays. ArXiv e-prints 2009.05150.

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

Conley, T.G., Taber, C.R., 2011. Inference with "difference in differences" with a small number of policy changes. Review of Economics and Statistics 93, 113–125.

Corrado, L., Fingleton, B., 2012. Where is the economics in spatial econometrics? Journal of Regional Science 52, 210–239.

Davezies, L., D'Haultfœuille, X., Guyonvarch, Y., 2021. Empirical process results for exchangeable arrays. Annals of Statistics 49, to appear.

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

Davidson, R., MacKinnon, J.G., 2000. Bootstrap tests: How many bootstraps? Econometric Reviews 19, 55–68.

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

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

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

Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd, Edinburgh.

Gelfand, A.E., Diggle, P., Fuentes, M., Guttorp, P., 2010. Handbook of Spatial Statistics. Chapman and Hall/CRC Press, Boca Raton.

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

Hagemann, A., 2019b. Permutation inference with a finite number of heterogeneous clusters. ArXiv e-prints 1907.01049.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer-Verlag, New York.

Hansen, B.E., 1999. The grid bootstrap and the autoregressive model. Review of Economics and Statistics 81, 594–607.

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

Hansen, C.B., 2007. Asymptotic properties of a robust variance matrix estimator for panel data when T is large. Journal of Econometrics 141, 597–620.

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

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

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

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, New York.

Jardim, E., Long, M.C., Plotnick, R., van Inwegen, E., Vigdor, J., Wething, H., 2017. Minimum wage increases, wages, and low-wage employment: Evidence from Seattle. Working Paper 23532. National Bureau of Economic Research.

Leeb, H., Pötscher, B.M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21, 21–59.

Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses. Springer, New York.

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, 851–881.

MacKinnon, J.G., 2021. Fast cluster bootstrap methods for linear regression models. QED Working Paper. Queen's University.

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

MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2021a. Influence and leverage in clustered regression models. QED Working Paper. Queen's University.

MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2021b. Wild bootstrap and asymptotic inference with multiway clustering. Journal of Business & Economic Statistics 39, 505–519.

MacKinnon, J.G., Webb, M.D., 2017a. Wild bootstrap inference for wildly different cluster

sizes. Journal of Applied Econometrics 32, 233–254.

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

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

MacKinnon, J.G., Webb, M.D., 2020a. Randomization inference for difference-in-differences with few treated clusters. Journal of Econometrics 218, 435–450.

MacKinnon, J.G., Webb, M.D., 2020b. When and how to deal with clustered errors in regression models. QED Working Paper 1421. Queen's University.

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

Manning, A., 2021. The elusive employment effect of the minimum wage. Journal of Economic Perspectives 35, 3–26.

Menzel, K., 2021. Bootstrap with cluster-dependence in two or more dimensions. Econometrica 89, to appear.

Miglioretti, D.L., Heagerty, P.J., 2006. Marginal modeling of nonnested multilevel data using standard software. American Journal of Epidemiology 165, 453–463.

Mincer, J., 1976. Unemployment effects of minimum wages. Journal of Political Economy 84, S87–S104.

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

Neumark, D., 2019. State minimum wage data set through Sept. 2019. URL: http://www.economics.uci.edu/~dneumark/datasets.html.

Neumark, D., Shirley, P., 2021. Myth or measurement: What does the new minimum wage research say about minimum wages and job loss in the United States? Working Paper 28388. National Bureau of Economic Research.

Neumark, D., Wascher, W., 1995. Minimum wage effects on employment and school enrollment. Journal of Business & Economic Statistics 13, 199–206.

Neumark, D., Wascher, W., 2007. Minimum wages, the earned income tax credit, and employment: Evidence from the post-welfare reform era. IZA Discussion Papers 2610.

Pustejovsky, J.E., Tipton, E., 2018. Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. Journal of Business & Economic Statistics 36, 672–683.

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

Roodman, D., MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2019. Fast and wild: Bootstrap

inference in Stata using boottest. Stata Journal 19, 4–60.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., Sobek, M., 2020. IPUMS USA: Version 10.0 [dataset].

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.

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

White, H., 1984. Asymptotic Theory for Econometricians. Academic Press, San Diego.

Wolfson, P., Belman, D., 2019. 15 years of research on US employment and the minimum wage. Labour 33, 488–506.

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.