# How Cluster-Robust Inference Is Changing Applied Econometrics*

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

August 17, 2018

**Abstract**

In many fields of economics, and also in other disciplines, it is hard to justify the assumption that the random error terms in regression models are uncorrelated. Assuming that they are correlated within clusters, such as geographical areas or time periods, but uncorrelated across clusters, seems more plausible. It has therefore become very popular to use "clustered" standard errors, which are robust against arbitrary patterns of within-cluster variation and covariation. Conventional methods for inference using clustered standard errors work very well when the model is correct and the data satisfy certain conditions, but they can produce very misleading results in other cases. This paper discusses some of the issues that users of these methods need to be aware of.

**Keywords:** CRVE, grouped data, clustered data, panel data, wild cluster bootstrap, difference-in-differences, treatment model, fixed effects

*This research was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada. The paper was first presented, under a slightly different title, at the 2018 Joint Statistical Meetings in Vancouver. I am grateful to Matt Webb and Morten Nielsen for comments and for joint work that made this paper possible, and to Stas Kolenikov for comments and for inviting me to write this paper and present it at the JSM.

# 1 Introduction

The assumption that the disturbances (random error terms) in regression models are uncorrelated across observations is a very strong one. Econometricians have long been aware of the potential for serial correlation when using time-series data, and methods for dealing with it have been a major focus of econometric research. But for data at the individual level, it was traditionally assumed that the disturbances are uncorrelated, perhaps after time and/or group fixed effects were included among the regressors. The idea was that any correlation across observations could be accounted for by the fixed effects.

This assumption changed quite rapidly, beginning in the mid 1990s, after the popular econometrics package `Stata` offered the option of cluster-robust, or "clustered," standard errors, which are discussed in Section 2. It soon became common to allow for arbitrary patterns of within-cluster correlation for clusters defined in various ways. In the education literature, for example, the disturbances for models of student performance might be clustered by classroom, by teacher, by school, or perhaps by school district. In the health literature, the disturbances for models of health outcomes might be clustered by doctor, by hospital, or by hospital chain. In the development literature, the disturbances might be clustered by village, by province, by country, or even by regional groups of countries, depending on the nature of the model and dataset. Whenever the observations can plausibly be grouped into a set of clusters, it has become customary, indeed often mandatory, in many areas of applied econometrics to use clustered standard errors.

Cameron and Miller (2015) provides a comprehensive survey of cluster-robust inference in econometrics, but there have been a number of developments since it was written. This paper will not attempt to be comprehensive. Instead, it will focus on a few key concepts and issues, and it will discuss some recent developments. Section 2 briefly reviews the literature on cluster-robust covariance matrices. Section 3 discusses the consequences of clustered disturbances for statistical inference. Section 4 discusses some of the issues that can make finite-sample cluster-robust inference problematical, along with methods such as the wild cluster bootstrap designed to make it more reliable. Section 5 presents an empirical example which illustrates how, in a large sample, inferences can be very sensitive to assumptions about how the disturbances are clustered. Section 6 discusses some of the reasons why residuals may display intra-cluster correlation, and Section 7 concludes.

# 2 Cluster-Robust Covariance Matrices

For simplicity, consider the linear regression model

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

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $N \times 1$ vectors of observations and disturbances, $\boldsymbol{X}$ is an $N \times K$ matrix of exogenous covariates, and $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector. With one-way clustering, which is currently the most common case, there are $G$ clusters, indexed by $g$, where the $g^{\text{th}}$ cluster has $N_g$ observations. The $N \times N$ covariance matrix $\boldsymbol{\Omega}$ is block-diagonal, with $G$ diagonal

blocks that correspond to the $G$ clusters:

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

Here $\boldsymbol{\Omega}_g$ is the $N_g \times N_g$ covariance matrix for the observations belonging to the $g^{\text{th}}$ cluster, which is assumed to be positive definite but unknown. For notational convenience, the observations here are ordered by cluster, although this is not necessary in practice. What is essential is that every observation be known to belong to one and only one cluster.

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

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

where the $N_g \times k$ matrix $\boldsymbol{X}_g$ contains the rows of $\boldsymbol{X}$ that belong to the $g^{\text{th}}$ cluster. The fact that $\text{Var}(\hat{\boldsymbol{\beta}})$ has this form has important consequences for inference; see Section 3.

In order to estimate (3), we replace the $K \times K$ matrices $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ by their sample analogs, using the outer product of the residual vector $\hat{\boldsymbol{u}}_g$ with itself to estimate $\boldsymbol{\Omega}_g$. This yields a cluster-robust variance estimator, or CRVE, of which the most widely-used version is

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

The first factor here is asymptotically negligible, but it makes $\text{CV}_1$ larger when $G$ and $N$ are finite. It is analogous to the factor $1/(N-K)$ used in the well-known $\text{HC}_1$ covariance matrix (MacKinnon and White 1985) that is robust only to heteroskedasticity of unknown form. Note that $\text{CV}_1$ reduces to the latter when each cluster contains just one observation, so that $G = N$.

Covariance matrix estimators like (4) are often referred to as "sandwich estimators" because there are two identical pieces of "bread" on the outside and a "filling" in the middle. The filling in the sandwich is supposed to estimate the corresponding matrix in equation (3). In both cases, the filling involves a sum of $G$ matrices. In the case of (4) and other CRVEs, these matrices have rank 1, even though they are of dimension $K \times K$. In contrast, the matrices $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ in (3) usually have rank $K$. This makes it clear that the individual components of the filling in (4) cannot possibly provide consistent estimators of the corresponding components of the filling in (3).

Because the matrices in the filling of (4) are of rank 1, the CRVE itself can have rank at most $G$. This makes it impossible to test hypotheses involving more than $G$ restrictions using Wald tests. Moreover, for hypotheses that involve numbers of restrictions not much smaller than $G$, the finite-sample properties of Wald tests based on (4) and other CRVEs are likely to be poor. Indeed, as will be discussed in Section 4, any sort of Wald test based on a CRVE, including $t$ tests, can have very poor finite-sample properties in some cases.

Although $\text{CV}_1$ is by far the most commonly employed CRVE, it is not the only one. A more complicated estimator, which is the analog of the $\text{HC}_2$ estimator studied in MacKinnon and White (1985), was proposed in Bell and McCaffrey (2002) and has recently been advocated by Imbens and Kolesár (2016); see also Pustejovsky and Tipton (2017). This estimator is

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

where $\boldsymbol{M}_{gg}^{-1/2}$ is the inverse symmetric square root of the matrix $\boldsymbol{M}_{gg} \equiv \mathbf{I}_{N_g} - \boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g'$, which is the $g^{\text{th}}$ diagonal block of $\boldsymbol{M}_{\boldsymbol{X}} \equiv \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, an $N \times N$ projection matrix. Thus $\text{CV}_2$ omits the scalar factor in $\text{CV}_1$ and replaces the residual subvectors $\hat{\boldsymbol{u}}_g$ by rescaled subvectors $\boldsymbol{M}_{gg}^{-1/2}\hat{\boldsymbol{u}}_g$.

Ordinary least squares shrinks the disturbance vector $\boldsymbol{u}$ differentially when it creates the residual vector $\hat{\boldsymbol{u}}$. Because the rescaling in (5) tends to undo the shrinkage, $\text{CV}_2$ typically yields larger and more accurate standard errors than $\text{CV}_1$. However, $\text{CV}_2$ is considerably more expensive to compute than $\text{CV}_1$ when the clusters are large, because it requires finding the inverse symmetric square root of the $N_g \times N_g$ matrix $\boldsymbol{M}_{gg}$ for each cluster. In fact, it seems to be numerically difficult to compute $\text{CV}_2$ once any of the $N_g$ exceeds 5000 or so; see MacKinnon and Webb (2018). Nevertheless, $\text{CV}_2$ should certainly be considered for samples of moderate size.

Using a different CRVE is not the only way to obtain inferences that are more accurate than the ones from Wald tests based on $\text{CV}_1$. A large number of methods is available, some of which, notably ones based on the wild cluster bootstrap, will be discussed in Section 4.

The true covariance matrix (3) and its estimators (4) and (5) allow for one-way clustering. However, there are models and datasets for which it is plausible that there may be multi-way clustering. For example, with individual data gathered at different times in different places, there may be clustering by both time period and location. This led Cameron, Gelbach and Miller (2011) and Thompson (2011) to propose CRVEs that allow for clustering in two or more dimensions.

In the two-dimensional case, the filling in the true covariance matrix (3) becomes

$$\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g + \sum_{h=1}^{H}\boldsymbol{X}_h'\boldsymbol{\Omega}_h\boldsymbol{X}_h - \sum_{g=1}^{G}\sum_{h=1}^{H}\boldsymbol{X}_{gh}'\boldsymbol{\Omega}_{gh}\boldsymbol{X}_{gh}. \tag{6}$$

Here there are $G$ clusters in the first dimension and $H$ in the second, $\boldsymbol{X}_g$ contains the rows of $\boldsymbol{X}$ that belong to cluster $g$ in the first dimension, and $\boldsymbol{X}_h$ contains the rows of $\boldsymbol{X}$ that belong to cluster $h$ in the second dimension. Similarly, $\boldsymbol{\Omega}_g$ is the covariance matrix for cluster $g$ in the first dimension, and $\boldsymbol{\Omega}_h$ is the covariance matrix for cluster $h$ in the second dimension. The matrix $\boldsymbol{X}_{gh}$ contains the rows of $\boldsymbol{X}$ that belong both to cluster $g$ in the first dimension and to cluster $h$ in the second. Similarly, the matrix $\boldsymbol{\Omega}_{gh}$ is the covariance matrix for observations that belong to both clusters $g$ and $h$. Notice the minus sign in (6). Without it, there would be double counting.

The filling in the two-way CRVE analogous to (4) that corresponds to (6) is

$$\sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g' \boldsymbol{X}_g + \sum_{h=1}^{H} \boldsymbol{X}_h' \hat{\boldsymbol{u}}_h \hat{\boldsymbol{u}}_h' \boldsymbol{X}_h - \sum_{g=1}^{G} \sum_{h=1}^{H} \boldsymbol{X}_{gh}' \hat{\boldsymbol{u}}_{gh} \hat{\boldsymbol{u}}_{gh}' \boldsymbol{X}_{gh}, \tag{7}$$

where the notation should be obvious. Notice that, because the last term is subtracted, this matrix may not be positive definite in finite samples. Also, the number of terms in the double summation may be less than $GH$, perhaps much less, because there may be no observations associated with some $gh$ pairs.

The two-way CRVE based on (7) can be extended to multi-way clustering in three or even more dimensions, although the algebra rapidly gets complicated; see Cameron, Gelbach and Miller (2011). In practice, it is often not at all obvious whether to use one-way clustering or two-way clustering, and the choice can be important for inference, as the empirical example in Section 5 illustrates.

# 3 Consequences of Clustered Disturbances

Allowing the disturbances to be correlated fundamentally changes the nature of statistical inference, especially for large samples. This is most easily seen in the context of estimating a population mean. Suppose we have a sample of $N$ uncorrelated observations, $y_i$, each with variance $\text{Var}(y_i)$ that is bounded from below and above. Then the usual formula for the variance of the sample mean $\bar{y}$ is

$$\text{Var}(\bar{y}) = \frac{1}{N^2} \sum_{i=1}^{N} \text{Var}(y_i) = \frac{1}{N} \sigma^2, \tag{8}$$

where $\sigma^2$ is the limiting value of the average of the $\text{Var}(y_i)$. The result (8) is obvious when the disturbances are homoskedastic, since $\text{Var}(y_i) = \sigma^2$ for all $i$. But it also holds under heteroskedasticity of unknown form, provided the limiting value $\sigma^2$ exists and is finite. The sandwich has disappeared in this case, because the only regressor is a constant term, and the product of the two $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrices is just $1/N^2$.

From (8) it is easy to see that $\text{Var}(\bar{y}) \to 0$ as $N \to \infty$. But this result depends crucially on the assumption that the $y_i$ are uncorrelated. Without such an assumption, the variance of the sample mean would be

$$\text{Var}(\bar{y}) = \frac{1}{N^2} \sum_{i=1}^{N} \text{Var}(y_i) + \frac{2}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \text{Cov}(y_i, y_j). \tag{9}$$

The first term on the right-hand side is the middle expression in (8) and is $O(1/N)$, as we would expect. But the second term is $O(1)$, because it is $2/N^2$ times a double summation involving $O(N^2)$ elements. Thus, even if the $\text{Cov}(y_i, y_j)$ are very small, the variance of $\bar{y}$ will never converge to zero as $N \to \infty$. Instead, it will ultimately converge to whatever the second term converges to. Therefore, $\bar{y}$ cannot estimate the population mean consistently. For a more detailed discussion of this type of inconsistency, see Andrews (2005).

The variance given in (9) is the variance of the sample mean when there is just one cluster, since every observation may be correlated with every other observation. When there is one-way clustering, the variance is instead

$$\text{Var}(\bar{y}) = \frac{1}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\text{Var}(y_{gi}) + \frac{2}{N^2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{j=i+1}^{N_g}\text{Cov}(y_{gi}, y_{gj}), \tag{10}$$

where $y_{gi}$ is the $i^{\text{th}}$ observation in cluster $g$. The second term here now involves a triple summation, the number of elements in which is of order $G(\max N_g)^2$. For $\bar{y}$ to be consistent, $G(\max N_g)^2/N^2$ must tend to 0 as $N \to \infty$. The easiest way to ensure that this happens is to let $G$ increase with $N$, while bounding all the $N_g$. In that case, $\bar{y}$ will converge to the population mean at rate $G^{-1/2}$, which is proportional to $N^{-1/2}$. However, it is also possible for $G$ to increase more slowly than $N$ and the $N_g$ to increase without bound, provided they do not do so too fast. When that happens, $\bar{y}$ will converge at a rate slower than $G^{-1/2}$. For a detailed discussion of the conditions that must be imposed on the number of clusters and their sizes for $\hat{\boldsymbol{\beta}}$ to be consistent in the regression case, see Djogbenou, MacKinnon and Nielsen (2018).

Equation (10) makes it clear that inference with clustered disturbances can be very different from inference with uncorrelated ones. When $G$ is fixed, or increases less rapidly than $N$, the information contained in a sample grows more slowly than the sample size. As the sample gets larger, the first term in (10) shrinks at rate $N^{-1}$, while the second term either stays roughly constant (when $G$ is fixed) or shrinks at a rate slower than $N^{-1}$ (when $G$ increases more slowly than $N$). Thus, for large samples, the second term must dominate the first term unless $G$ is proportional to $N$, and the rate at which information accumulates is then determined by the latter. This implies that the amount of information about the parameters of interest contained in extremely large samples may be very much less than intuition would suggest. We will encounter an example of this in Section 5.

# 4 Inference in Finite Samples

Much of the work on cluster-robust inference in recent years has focused on inference in finite samples. The meaning of "finite" is not the usual one, however. What matters for reliable inference is not the number of observations, $N$, but the number of clusters, $G$. In addition, the way in which observations are distributed across clusters and the way in which the $\boldsymbol{X}_g$ matrices vary across clusters can greatly affect the reliability of finite-sample inference. Simple rules about how many clusters are needed for reliable inference have been suggested, but they can be misleading. For example, Angrist and Pischke (2008) suggested that 42 clusters is generally sufficient, a conclusion strongly disputed in MacKinnon and Webb (2017b).

Suppose we are interested in one element of $\boldsymbol{\beta}$, say $\beta_j$. Cluster-robust inference is typically based on the $t$ statistic

$$t_j = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}_{jj}}}, \tag{11}$$

where $\beta_{j0}$ is the value under the null hypothesis, and $\hat{V}_{jj}$ is the $j^{\text{th}}$ diagonal element of the $\text{CV}_1$ matrix (4). The statistic $t_j$ is usually assumed to follow the $t(G-1)$ distribution, which can be justified by a result in Bester, Conley and Hansen (2011).

If there are at least 50 clusters and they are reasonably homogeneous, that is, similar in size and similar in their $\boldsymbol{X}_g'\boldsymbol{X}_g$ and $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ matrices, then inference based on (11) and the $t(G-1)$ distribution typically works very well. However, there can be severe over-rejection when these conditions are not satisfied; see, among others, MacKinnon and Webb (2017b) and Djogbenou, MacKinnon and Nielsen (2018). The latter considers a case in which one cluster is much bigger than any of the others and finds that the test based on (11) over-rejects severely even with over 200 clusters. This is true even when the largest cluster is becoming a smaller fraction of the sample at a rate fast enough for asymptotic theory to be valid as $G$ increases.

One way to check whether inference is likely to be reliable is to compute the "effective number of clusters," $G^*$, as defined in Carter, Schnepel and Steigerwald (2017). This depends on $G$, the $N_g$, and the entire $\boldsymbol{X}$ matrix, and it requires assumptions about the extent of intra-cluster correlation. When $G^*$ is substantially less than $G$, and especially when it is small (say, less than 20), tests based on the $t(G-1)$ distribution are almost certain to over-reject. Computing $G^*$ using the entire sample can be costly or even infeasible when $N$ is large, but it is often possible to compute a very good estimate using a subsample. Inference can be based on the $t(G^*-1)$ distribution, if desired, although this does not seem to be the best approach.

As was discussed in Section 2, the test statistic (11) can be modified by using the $\text{CV}_2$ covariance matrix given in (5) instead of $\text{CV}_1$. Bell and McCaffrey (2002) and Imbens and Kolesár (2016) further suggest ways of computing a degrees-of-freedom parameter to be used instead of $G-1$. Young (2016) proposes a different way of accomplishing essentially the same thing. His method first corrects the bias of the $\text{CV}_1$ standard error, thus avoiding the computational difficulties of calculating $\text{CV}_2$, and then calculates a degrees-of-freedom parameter. These methods are discussed and compared in MacKinnon and Webb (2018).

Instead of comparing $t_j$ to the $t$ distribution with a computed number of degrees of freedom, we can compare it to a bootstrap distribution. There are several ways in which the bootstrap samples can be generated. In most cases, the best approach seems to be to use the restricted wild cluster bootstrap (WCR) proposed in Cameron, Gelbach and Miller (2008), which we now discuss. MacKinnon and Webb (2017b) studies this method in detail, and Djogbenou, MacKinnon and Nielsen (2018) proves that it is asymptotically valid. The basic idea is to generate the vector of bootstrap disturbances for each cluster using the vector of residuals for that cluster, so as to retain the intra-cluster covariances of the latter. The method is called "restricted" because the parameters and disturbances of the bootstrap DGP are based on estimates that satisfy the null hypothesis.

Suppose the objective is to test the restriction $\boldsymbol{a}'\boldsymbol{\beta} = \boldsymbol{0}$, where $\boldsymbol{a}$ is a known vector of length $K$. Then the WCR bootstrap works as follows:

1. Obtain OLS estimates $\hat{\boldsymbol{\beta}}$ and the CRVE $\hat{\boldsymbol{V}}$ using (1) and (4). Also, re-estimate (1) subject to the restriction $\boldsymbol{a}'\boldsymbol{\beta} = \boldsymbol{0}$ to obtain restricted estimates $\tilde{\boldsymbol{\beta}}$ and residuals $\tilde{\boldsymbol{u}}$.

2. Calculate the cluster-robust $t$-statistic, $t_a = \boldsymbol{a}'\hat{\boldsymbol{\beta}}/\sqrt{\boldsymbol{a}'\hat{\boldsymbol{V}}\boldsymbol{a}}$.

3. For each of $B$ bootstrap replications, indexed by $b$,

   (a) generate a set of bootstrap disturbances $\boldsymbol{u}^{*b}$, where the subvector corresponding to cluster $g$ is equal to $\boldsymbol{u}_g^{*b} = v_g^{*b}\tilde{\boldsymbol{u}}_g$, and the $v_g^{*b}$ are independent realizations of an auxiliary random variable $v^*$ with zero mean and unit variance;

   (b) generate the bootstrap dependent variables according to $\boldsymbol{y}^{*b} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{u}^{*b}$;

   (c) obtain the bootstrap estimate $\hat{\boldsymbol{\beta}}^{*b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}^{*b}$, the bootstrap residuals $\hat{\boldsymbol{u}}^{*b}$, and the bootstrap covariance matrix

$$\hat{\boldsymbol{V}}_b^* = \frac{G(N-1)}{(G-1)(N-K)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\hat{\boldsymbol{u}}_g^{*b}(\hat{\boldsymbol{u}}_g^{*b})'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}; \tag{12}$$

   (d) calculate the bootstrap $t$-statistic

$$t_a^{*b} = \frac{\boldsymbol{a}'\hat{\boldsymbol{\beta}}^{*b}}{\sqrt{\boldsymbol{a}'\hat{\boldsymbol{V}}_b^*\boldsymbol{a}}}.$$

4. If the alternative hypothesis is $\boldsymbol{a}'\boldsymbol{\beta} \neq \boldsymbol{0}$ and there is no reason to expect the test statistic to have a nonzero mean, compute the symmetric bootstrap $P$ value

$$\hat{P}_{\mathrm{S}}^* = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\big(|t_a^{*b}| > |t_a|\big),$$

   where $\mathbb{I}(\cdot)$ denotes the indicator function. Alternatively, we can compute an upper-tail, lower-tail, or equal-tail $P$ value, as appropriate.

The WCR bootstrap has two key features. The first is that the same realization of the auxiliary random variable, $v_g^{*b}$, multiplies every residual within cluster $g$ for bootstrap sample $b$. This ensures that the bootstrap DGP retains the intra-cluster covariances of the residuals, which, on average, should look like the intra-cluster covariances of the disturbances. The second is that the bootstrap DGP imposes the null hypothesis. In principle, the $v^*$ could follow any distribution with mean 0 and variance 1. However, in most cases, it seems to be best to employ the Rademacher distribution, for which $v^* = 1$ or $-1$, each with probability 0.5. This is not a good idea when $G$ is very small, however, because the number of distinct bootstrap samples is just $2^G$; see Webb (2014).

Provided the number of clusters is not too large, it is possible to generate a large number of wild cluster bootstrap statistics very efficiently. This is what the Stata routine `boottest` does; see Roodman, MacKinnon, Nielsen and Webb (2018). The algorithm it uses actually computes the $t_a^{*b}$ without explicitly calculating either the bootstrap residuals $\hat{\boldsymbol{u}}^{*b}$ or the bootstrap CRVE (12). All of the computations that are $O(N)$ are just done once, rather than for every bootstrap sample.

Because `boottest` is remarkably efficient in most cases, it probably makes sense to use the restricted wild cluster bootstrap most of the time. Confidence intervals can easily be obtained by inverting the bootstrap test, and `boottest` does this by default. Note that, even though the disturbances for bootstrap samples generated by the wild cluster bootstrap are clustered in only one dimension, this bootstrap method can be used in conjunction with multi-way clustered standard errors. MacKinnon, Nielsen and Webb (2017) shows that this often works well, and `boottest` makes it easy to do.

Other bootstrap methods can also be used. MacKinnon and Webb (2018) argues that the ordinary wild bootstrap can work better than the wild cluster bootstrap in certain cases (notably, when interest focuses on a treatment dummy, very few clusters are treated, and clusters are otherwise homogeneous). Djogbenou, MacKinnon and Nielsen (2018) proves that using it is asymptotically valid, even though it cannot mimic the intra-cluster correlations of the disturbances.[1] Bertrand, Duflo and Mullainathan (2004) suggests using the pairs cluster bootstrap, in which the data are resampled by cluster. This typically does not work as well as the wild cluster bootstrap; see Cameron, Gelbach and Miller (2008), MacKinnon and Webb (2017a), and below. However, it has the advantage that it can be used for models which are not regression models.

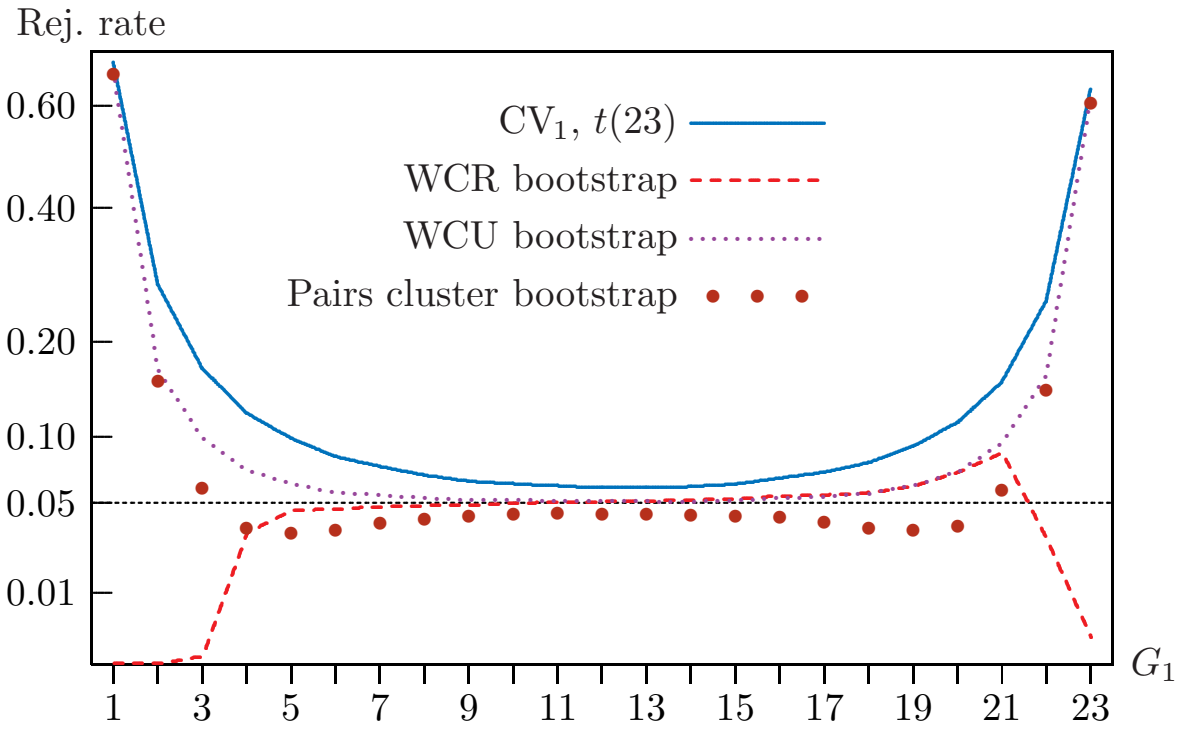
Reliable inference is particularly challenging when the parameter of interest is the coefficient on a treatment dummy variable, treatment is assigned at the cluster level, and there are very few treated clusters. This includes the case of difference-in-differences regressions when all the treated observations belong to just a few clusters. It has been known for some time that inference based on cluster-robust $t$ statistics greatly over-rejects in such cases; see Abadie, Diamond and Hainmueller (2010) and Conley and Taber (2011). Precisely why this happens is explained in MacKinnon and Webb (2017b). Unfortunately, as MacKinnon and Webb (2017b, 2018) show, the wild cluster bootstrap does not solve the problem. In fact, the wild cluster bootstrap either greatly under-rejects or greatly over-rejects, depending on whether or not the null hypothesis is imposed on the bootstrap DGP.

Figure 1, which is taken from MacKinnon and Webb (2017a), illustrates what can happen in a model where the regressor of interest is a treatment dummy that equals 0 or 1 for every observation in each cluster. In this case, there are 24 clusters, which vary in size from 32 to 235, for a total of 2400 observations. The figure shows rejection frequencies at the .05 level for four tests, based on 400,000 replications, as a function of $G_1$, the number of treated clusters, when clusters are treated from smallest to largest. The usual test, based on the $\text{CV}_1$ covariance matrix (4), over-rejects very severely when $G_1$ is small or large. When there is just one treated or non-treated cluster, it rejects over 60% of the time. The over-rejection drops quite sharply as $\min(G_1, G - G_1)$ increases, however. For $8 \leq G_1 \leq 16$, the usual test always rejects less than 7% of the time.

The two bootstrap tests that do not impose the null hypothesis, WCU and pairs cluster, also over-reject very severely when there is just one treated or non-treated cluster. However, their performance improves very rapidly as $\min(G_1, G - G_1)$ increases. The pairs cluster bootstrap actually under-rejects noticeably for intermediate values of $G_1$, most severely for

[1] Unfortunately, the tricks that `boottest` uses to save computer time are not very effective for the ordinary wild bootstrap, and the program can encounter memory limitations with large samples.

Figure 1: Rejection Frequencies, Treatment Model, $G = 24$, $N = 2400$




$G_1 = 5$ and $G_1 = 19$. The WCU bootstrap never under-rejects, and it works almost perfectly for $9 \leq G_1 \leq 15$.

The WCR bootstrap performs very differently from the other two bootstrap methods. It never rejects (in 400,000 replications) for $G_1 = 1$ and $G_1 = 2$, and it under-rejects severely for $G_1 \leq 4$ and $G_1 \geq 22$. However, it over-rejects noticeably for $19 \leq G_1 \leq 21$. For most intermediate values of $G_1$, the WCR and WCU bootstraps perform very similarly here, something that is generally not true for more complicated models.

The asymmetry that is evident in Figure 1 arises from the facts that cluster sizes vary and that clusters are treated from smallest to largest. Thus the treated clusters on the left-hand side of the figure are relatively small, and the non-treated ones on the right-hand side are relatively large. The theoretical analysis of treatment models in MacKinnon and Webb (2017b, 2018) explains how the numbers and sizes of treated and non-treated clusters affect rejection frequencies. In practice, it would rarely be the case that only the smallest or largest clusters are treated, so the figure is in some respect unrealistically extreme.

## 5 An Empirical Example

The impact of alternative assumptions about how the disturbances are clustered can be striking. In this section, I illustrate this in the context of a simple earnings equation. The dependent variable $y_{ti}$ is the logarithm of weekly earnings for men aged 25 to 65, conditional on earnings being greater than \$20. The key regressors are age, age squared,

and four education dummies. `Ed2` is a dummy for completing high school, `Ed3` is a dummy for completing two years of college or university, `Ed4` is a dummy for obtaining a university degree, and `Ed5` is a dummy for obtaining a postgraduate degree. The data come from the U.S. Current Population Survey (CPS) for the years 1979 through 2015 (37 years). There are 1,156,597 observations from 51 states (including the District of Columbia). On average, there are about 31,259 observations per year. The largest state (California) has 87,427 observations, and the smallest (Hawaii) has only 4,068.

The equation that I estimate, using ordinary least squares, is

$$y_{gti} = \beta_1 + \sum_{j=2}^{5} \beta_j \texttt{ED}j_{gti} + \beta_6 \texttt{Age}_{gti} + \beta_7 \texttt{Age}_{gti}^2 + \sum_{s=1}^{36} \gamma_s \texttt{Year}_t^s + \sum_{k=1}^{50} \eta_k \texttt{State}_g^k + u_{gti}, \quad (13)$$

where $g$ denotes the state, $t$ denotes the year, and $i$ denotes the individual. In equation (13), $\texttt{Year}_t^s$ is a dummy that equals 1 when $s = t$, and $\texttt{State}_g^k$ is a dummy that equals 1 when $g = k$. One year dummy and one state dummy have been omitted to avoid perfect collinearity. This equation could be used to answer various economic questions. For concreteness, I focus on the value of obtaining a postgraduate degree.[2]

The coefficients on `Ed4` and `Ed5` are $\hat{\beta}_4 = 0.67762$ and $\hat{\beta}_5 = 0.78727$. Thus the estimated percentage increase in earnings associated with having obtained the higher degree is

$$\hat{\delta} = 100\big(\exp(\hat{\beta}_5 - \hat{\beta}_4) - 1\big) = 100\big(\exp(0.78727 - 0.67762) - 1\big) = 11.589\%. \quad (14)$$

Of course, since people make choices about how much education to obtain, we cannot naively interpret this number as an estimate of how much more someone who chose to obtain only an undergraduate degree would earn if they had chosen to obtain a postgraduate degree as well. At best, it is simply an empirical regularity.

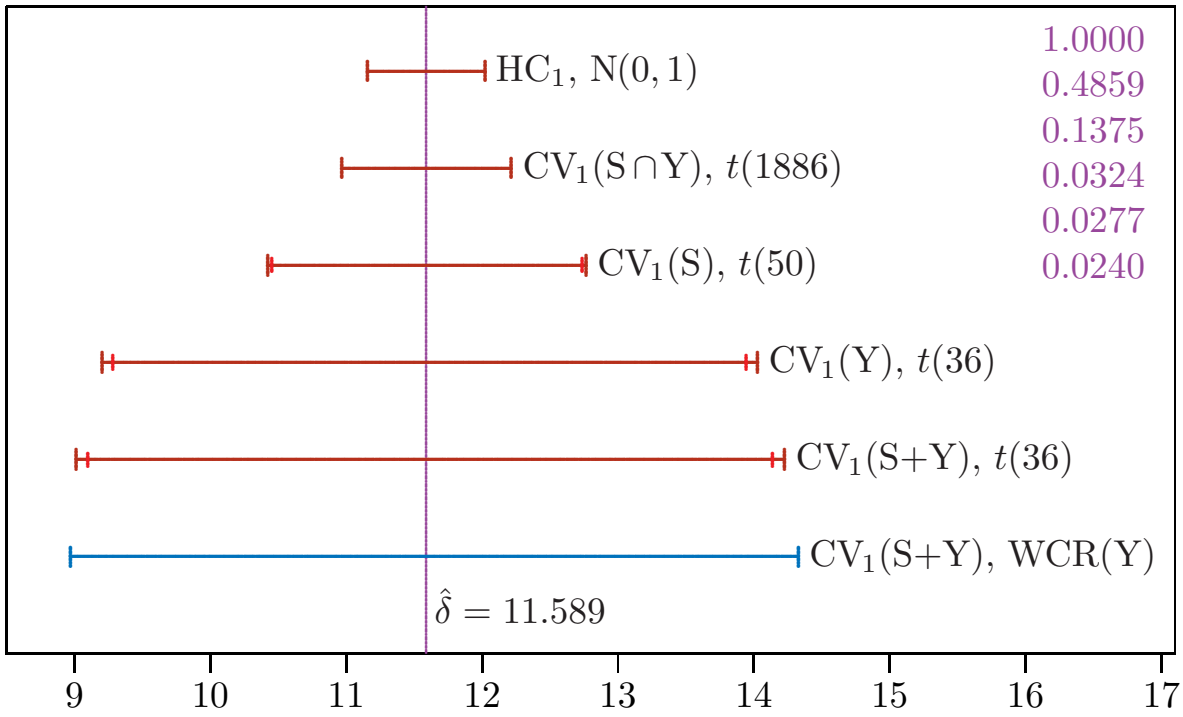
In order to compute a confidence interval for $\delta$, the population equivalent of $\hat{\delta}$ defined in (14), we need a standard error for the quantity $\hat{\beta}_5 - \hat{\beta}_4 = 0.10965$. The traditional approach would be to argue that, since the fixed effects account for any within-cluster correlation, we can just use a conventional heteroskedasticity-robust standard error. The $\text{HC}_1$ standard error is 0.001985, which suggests that we have estimated the difference between $\beta_5$ and $\beta_4$ with great accuracy.

However, including state and year fixed effects does not in fact eliminate all within-cluster correlation. It would only do so if the $u_{gti}$ in (13) followed a random-effects model, where $u_{gti}$ is the sum of a random state effect, a random year effect, and an individual effect. If there were instead random effects at the state-year level, or perhaps some more complicated pattern, the fixed effects could not eliminate all within-cluster correlation. Thus it seems plausible that there may be within-cluster correlations among the disturbances.

Figure 2 shows six different 95% confidence intervals for $\delta$. These are based on five different assumptions about how the disturbances are clustered. It is evident that, for this

[2] This equation was previously estimated using the same dataset in MacKinnon (2016), which contains a number of results not reported here, but no results for clustering by year or two-way clustering.

Figure 2: Confidence Intervals for $\hat{\delta}$




model and dataset, the assumptions we make about clustering have an enormous impact on the intervals we obtain.

The topmost interval in the figure, of which the lower and upper limits are 11.156 and 12.024, respectively, is based on the $\mathrm{HC}_1$ standard error 0.001985 given above and the critical value 1.96, which is the .975 quantile of the standard normal distribution. This is probably the interval that most investigators would have used until around the year 2000.

The second interval shown in Figure 2 is based on clustering by the intersection of state and year, which in the figure is denoted $\mathrm{CV}_1(\mathrm{S}\cap\mathrm{Y})$. There are $37 \times 51 = 1887$ clusters, so the .975 quantile of the $t(1886)$ distribution is used to obtain the limits of the interval, which are 10.968 and 12.214. This interval is wider than the first one, but not dramatically so. Some investigators still use intervals like this one, although they have little theoretical or empirical justification; see Bertrand, Duflo and Mullainathan (2004).

The next two intervals also use one-way clustering, but at a much higher level. For the third interval, clustering is by state, and for the fourth, it is by year. Each horizontal line here actually shows two intervals. The narrower one is based on the standard normal distribution, and the wider ones are based on the $t(50)$ and $t(36)$ distributions for clustering by state and year, respectively. The numbers of clusters are now small enough that the differences between the standard normal and Student's $t$ distributions are important.

It is not surprising that clustering by state yields a considerably wider interval than clustering by the intersection of state and year. The number of off-diagonal elements of $\mathbf{\Omega}$ that are allowed to be non-zero is very much greater with 51 big clusters than with 1887

small ones. However, it may be surprising that clustering by year yields a much wider interval than clustering by state. Based on results in Bertrand, Duflo and Mullainathan (2004), it appears to be widely believed that, in the context of data with both a time and a cross-section component, clustering by the latter (in this case states) is the right thing to do, because it allows for general patterns of serial correlation within states. Clustering by time period seems to be much less common.

To allow for both serial correlation within states and contemporaneous correlation across states, the fifth and sixth intervals use two-way clustering by state and year, where the middle matrix in the CRVE is (7). The fifth one is a conventional confidence interval based on the $t(36)$ distribution, while the sixth is a bootstrap interval using the restricted wild cluster bootstrap with bootstrap clustering by year; this is denoted WCR(Y) in the figure. The bootstrap samples were clustered by year because simulation results in MacKinnon, Nielsen and Webb (2017) suggest that, with multi-way clustering, it may be desirable to cluster them by the dimension with the fewest clusters. Clustering the bootstrap samples by state yielded extremely similar results. In this case, the impact of bootstrapping is quite modest, because the numbers of clusters in both dimensions are not all that small.

The bootstrap interval in Figure 2 is based on 99,999 bootstrap samples, so that there is very little simulation error. This may seem like an extraordinarily large number for a sample of over a million observations, but it was not computationally demanding to compute this interval using `boottest`, even though it involved numerically inverting a bootstrap test; see Roodman, MacKinnon, Nielsen and Webb (2018).

It may seem that we have to choose somewhat arbitrarily among the six intervals in Figure 2. However, as we discuss below, this is not the case. There appears to be strong, albeit informal, evidence that the top three intervals are too narrow. Whether we need to use two-way clustering or one-way clustering by year is not so clear, however. Bootstrapping the interval based on the latter makes it slightly wider, as expected, but still somewhat narrower than the bootstrapped two-way interval. These bootstrap intervals are [9.187, 14.088] and [8.971, 14.328], respectively.

Ideally, we could test which interval is appropriate, and this is an area of active research. A test between one-way clustering at a low level (or no clustering at all) against one-way clustering at a higher level has been proposed by Ibragimov and Müller (2016), but it is not applicable to two-way clustering and requires that the key parameter(s) be identifiable using the data for each cluster. More widely applicable tests are being developed in MacKinnon, Nielsen and Webb (2018).

The interval based on $\text{HC}_1$ is surely too narrow. Suppose that there were actually no intra-cluster correlations. Then the matrix $\boldsymbol{\Omega}$ would be diagonal, and all the intervals would be valid. However, because the $\text{HC}_1$ interval is the only one that makes use of the assumption that $\boldsymbol{\Omega}$ is diagonal, all the other intervals would be based on standard errors that are estimated inefficiently. Consider the matrix in the middle of the $\text{CV}_1$ covariance matrix given in (4). The element of this matrix that corresponds to the $k^{\text{th}}$ regressor is

$$\sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} X_{gi,k}^2 X_{gj,k}^2 \hat{u}_{gi} \hat{u}_{gj}. \tag{15}$$

When $i = j$ and there is no intra-cluster correlation, this quantity is an estimate of

$$\sum_{g=1}^{G} \sum_{i=1}^{N_g} \sigma_{gi}^2 X_{gi,k}^2. \tag{16}$$

The corresponding element of the middle matrix in the $\mathrm{HC}_1$ covariance matrix is

$$\sum_{g=1}^{G} \sum_{i=1}^{N_g} X_{gi,k}^2 \hat{u}_{gi}^2 = \sum_{i=1}^{N} X_{i,k}^2 \hat{u}_i^2, \tag{17}$$

where the expression on the right-hand side is the one that would normally be used in the absence of clustering. The difference between (15) and (17) is

$$\sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{j \neq i} X_{gi,k} X_{gj,k} \hat{u}_{gi} \hat{u}_{gj}. \tag{18}$$

When the $N_g$ are large, this expression may involve a great many terms, but they would all have expectation zero if the residuals were replaced by disturbances. To highest order, the residuals and disturbances should have the same properties here. Expression (18) may have either sign. When $G$ is fixed, both (17) and (18) are $O(N)$. In the former case, there are $N$ terms, each of them $O(1)$. In the latter case, there are $O(N^2)$ terms, each of them $O(N^{1/2})$ because they are assumed to have mean zero.

The results in Figure 2 make it clear that, for our empirical example, expression (18) must be very large when we cluster by state or by year, and the corresponding quantity for two-way clustering must be very large when we cluster by both of them. Thus it seems unlikely that the population analogs of these quantities are in fact zero. This conclusion must be tempered by the fact that, when $G$ is fixed, we cannot estimate $1/N$ times expression (18) consistently using a CRVE.

Although the sample is large, it actually contains much less information than it initially appears to because the disturbances are clustered. The purple numbers in the upper right of Figure 2 attempt to quantify this information loss. Each of them corresponds to one of the displayed intervals, in the same order from top to bottom. The number for a given interval is the ratio of the sample size that would be needed to obtain an interval of that length if the disturbances really were uncorrelated to the actual sample size. The smallest number here, 0.0240, tells us that the length of the bootstrap interval with two-way clustering is what we would expect to obtain using a sample of only $0.0240 \times 1{,}156{,}597 = 27{,}758$ observations with uncorrelated disturbances.

The information contained in this sample is thus very much less than it would seem to be if we did not allow for clustered disturbances. To investigate this issue further, I reduced the sample size from 1,156,597 to 72,288 by throwing away 15 out of every 16 observations. The $\mathrm{HC}_1$ standard error increased from 0.001985 to 0.007848, a factor of 3.95. This is just about what we would expect when the sample size is reduced by a factor of 16. However, the various clustered standard errors increased by much less. For example, the standard error

based on state-level clustering increased from 0.005206 to 0.009422, a factor of 1.81. Even more surprisingly, the standard error based on two-way clustering increased from 0.01148 to just 0.01261, a factor of only 1.098. Amazingly, throwing away 15/16 of the sample has increased the standard error by just under 10%.

Thus, for equation (13) with two-way clustering (and also with one-way clustering by year), the extra information we gain when we increase the sample size by a factor of 16 is very modest. But recall expression (10) for the variance of a sample mean when there is clustering. When we increase $N$ and all the $N_g$ by a factor of 16, the first term shrinks by a factor of 16, but the second term remains essentially the same size. This is also true when there is two-way clustering. Thus, if the second term is already fairly large relative to the first term, the net effect of increasing the sample size, even by a large factor, may be only a modest reduction in the sampling variance of an estimator.

# 6 Why Is There Intra-Cluster Correlation?

Precisely why residuals appear to be correlated within clusters in a great many econometric applications is not entirely clear and probably varies across models and datasets. In many cases, it seems reasonable to believe that there are unobserved quantities which affect all observations in at least some clusters. For data on educational outcomes, as an example, there may be unobserved random effects at the teacher and/or school and/or district levels; see Koedel, Parsons, Podgursky and Ehlert (2015).

More generally, all sorts of model misspecification could cause residuals to be correlated within clusters. There is, of course, a risk that misspecification might also cause residuals to be correlated across clusters. However, it seems plausible that the parts of any omitted variables which cannot be explained by cluster fixed effects and other regressors should be more highly correlated within than across clusters.

In the case of the empirical example, the design of the Current Population Survey probably accounts for some of the state-level intra-cluster correlation. The CPS is a complex survey. It uses various sampling techniques such as clustering, stratification, multiple stages of selection, and unequal probabilities of selection, in order to achieve a reasonable balance between the cost and statistical accuracy of the survey. However, the design of the CPS also ensures that the observations are not entirely independent within states. The basic unit of sample selection is the census tract, not the household. Once a tract has been selected, it typically contributes a number of households to the surveys that are done over several adjacent years. Any sort of dependence within census tracts will then lead to residuals that are correlated within states both within and across years.

In principle, it may be possible to take account of the features of the design of a particular survey; see, among others, Binder (1983) and Rao and Wu (1988). Kolenikov (2010) provides an accessible introduction to this literature along with Stata code for bootstrap inference when the survey design is known. When the survey design is very complex, however, it would be extremely difficult to implement this sort of procedure. When the design is unknown to the investigator, it would be impossible. In many cases, the best we can do is to use a CRVE clustered at the appropriate level. A widely recommended rule of thumb is to cluster

at the highest feasible geographic level (for example, by state in the empirical example of Section 5), because survey design issues would typically manifest themselves within but not across large geographic areas.

The other type of intra-cluster correlation observed in the empirical example, namely, correlation of observations for the same year, is almost certainly a consequence of misspecification. Because business cycles at the state or industry levels are not perfectly correlated with the national business cycle, the year fixed effects included in equation (13) cannot possibly account for all the effects of business cycles on earnings. This surely accounts for much of the clustering by year that we observe. The magnitude of the effect, and its consequences for the accuracy of parameter estimates, are strikingly large.

There is one important issue that this paper has not discussed, and will not discuss in any depth. All of the analysis has implicitly assumed that the data are actually generated by the regression model (1), and that the sample is very small relative to the population being studied. Thus the population contains a very large number of clusters, and the sample is obtained by choosing a small proportion of them at random. This seems quite reasonable in the education context, for example, where we are clustering by school, because in a country of any size there will be a great many schools, and most samples will contain only a small fraction of them. Conditional on the chosen clusters, the sample may contain all the observations for each cluster, or just some subset of them.

Formally, the empirical example of Section 5 does not satisfy the assumptions of the previous paragraph, however. If we think of the "population" as all employed men aged 25 to 65 in the United States between 1979 and 2015, then the number of clusters in the population (37 years or 51 states) is the same as the number of clusters in the sample. Implicitly, for the methods we have discussed to make sense, we must be trying to make inferences about a meta-population of states and a meta-population of years, from which actual states and actual years have been drawn at random. Whether or not this is a reasonable thing to do is a matter of opinion. Of course, econometricians do it all the time when they analyze time-series data.

Abadie, Athey, Imbens and Wooldridge (2017) has recently argued that many economic datasets do not satisfy the assumption that the sample is very small relative to the population being studied. In the context of cross-section studies of treatment effects, which may vary across units, they analyze cases in which the sample is large relative to the population and contains a large proportion of the clusters. The sample may contain all the observations in the included clusters, or only some of them. They find that, unless the number of clusters in the sample is very small relative to the number in the population, or there is no heterogeneity in treatment effects, cluster-robust standard errors tend to be too large, perhaps much too large. In some cases, heteroskedasticity-robust standard errors lead to more accurate inferences, even though there is considerable intra-cluster correlation.

Whether the conclusions of Abadie et al. (2017) apply to any given case is not at all clear, however. In the empirical example of Section 5, for example, all clusters are included in the sample, but the observations presumably come from a very small fraction of the neighborhoods within those clusters, and we expect much of the within-cluster correlation to arise from within-neighborhood correlation. Moreover, the example does not concern

treatment effects. Therefore, even if we are interested in the actual 51 states instead of a meta-population of states, the results of Abadie et al. (2017) do not imply that methods for cluster-robust inference should be avoided.

# 7 Conclusions

It has become extremely common in many areas of applied econometrics to "cluster" the standard errors at an often arbitrarily chosen level with $G$ clusters and rely on the $t(G-1)$ distribution for inference. This can be a reasonable thing to do, and failing to allow for clustering is often much worse. But this approach can also lead to seriously misleading inferences in many cases.

Even if the appropriate level of clustering is known, there can be serious problems. In general, standard methods work reasonably well when the number of clusters is reasonably large (at least 50) and the clusters are fairly homogeneous in terms of the numbers of observations and the characteristics of the regressors. One situation in which inference based on cluster-robust standard errors can be extremely misleading is when interest focuses on a treatment dummy variable and only a few clusters are treated; see MacKinnon and Webb (2017b, 2018). In this case, of course, the key regressor is very heterogeneous across clusters.

In Section 4, we discussed the large and rapidly growing literature aimed at improving cluster-robust inference in finite samples. A wide variety of methods is available, and it would often make sense to employ two or three of them. In many cases, but not all, the restricted wild cluster bootstrap works well. It can often be computed remarkably quickly using the Stata routine `boottest`; see Roodman, MacKinnon, Nielsen and Webb (2018).

One often overlooked feature of clustered disturbances is that the relationship between the sample size $N$ and the efficiency of parameter estimates does not have its usual form. When the disturbances are correlated within clusters, we saw in Section 3 that information accumulates at a rate slower than $\sqrt{N}$ unless the number of clusters increases at the same rate as the sample size. Thus, as the empirical example illustrates, "big" datasets may actually be much smaller than we think they are.

As the empirical example of Section 5 illustrates, clustered standard errors can be very sensitive to how the observations are clustered. In practice, investigators would therefore be wise to put a lot of thought into this. When there is more than one natural way to cluster, it generally makes sense to investigate all of them.

# References


Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) 'Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program.' *Journal of the American Statistical Association* 105(490), 493–505

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge (2017) 'When should you adjust standard errors for clustering?' NBER Working Papers 24003, Nov.

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

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

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), 249–275

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011) 'Inference with dependent data using cluster covariance estimators.' *Journal of Econometrics* 165(2), 137–151

Binder, David A. (1983) 'On the variances of asymptotically normal estimators from complex surveys.' *International Statistical Review* 51(3), 279–292

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

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

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

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2017) 'Asymptotic behavior of a $t$ test robust to cluster heterogeneity.' *Review of Economics and Statistics* 99(4), 698–709

Conley, Timothy G., and Christopher R. Taber (2011) 'Inference with "difference in differences" with a small number of policy changes.' *The Review of Economics and Statistics* 93(1), 113–125

Djogbenou, Antoine, James G. MacKinnon, and Morten Ø. Nielsen (2018) 'Asymptotic theory and wild bootstrap inference with clustered errors.' Working Paper 1399, Queen's University, Department of Economics

Ibragimov, Rustam, and Ulrich K. Müller (2016) 'Inference with few heterogeneous clusters.' *Review of Economics & Statistics* 98(1), 83–96

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

Koedel, Cory, Eric Parsons, Michael Podgursky, and Mark Ehlert (2015) 'Teacher prepatation programs and teacher quality: Are there real differences across programs.' *Educational Finance and Policy* 10(4), 508–534

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

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2017) 'Bootstrap and asymptotic inference with multiway clustering.' Technical Report 1386, Queen's University, Department of Economics

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2018) 'Testing for the level of (multiway) clustering.' Technical Report pending, Queen's University, Department of Economics

MacKinnon, James G. (2016) 'Inference with large clustered datasets.' *L'Actualité Economique* 92(4), 649–665

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

MacKinnon, James G., and Matthew D. Webb (2017a) 'Pitfalls when estimating treatment effects using clustered data.' *The Political Methodologist* 24(2), 20–31

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

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

Pustejovsky, James E., and Elizabeth Tipton (2017) 'Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models.' *Journal of Business and Economic Statistics* 35, to appear

Rao, J. N. K., and C. F. J. Wu (1988) 'Resampling inference with complex survey data.' *Journal of the American Statistical Association* 83(401), 231–241

Roodman, David, James G. MacKinnon, Morten Ø. Nielsen, and Matthew D. Webb (2018) 'Fast and wild: Bootstrap inference in Stata using boottest.' Working Paper 1406, Queen's University, Department of Economics

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

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

Young, Alwyn (2016) 'Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections.' Technical Report, London School of Economics