Queen's Economics Department Working Paper No. 1413

# How Cluster-Robust Inference Is Changing Applied Econometrics

James G. MacKinnon
Queen's University

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

11-2020 (references updated; minor corrections in 2019 and 2020)

# How Cluster-Robust Inference Is Changing Applied Econometrics*

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

November 14, 2020

**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. It seems more plausible to assume that they are correlated within clusters, such as geographical areas or time periods, but uncorrelated across clusters. 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 grant number 435-2016-0871 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. It was published in the *Canadian Journal of Economics*, 2019, 52(3), 851–881. I am grateful to Arthur Sweetman, Steve Lehrer, Richard Startz, several anonymous referees, and seminar participants at Queen's University, Binghamton University, the University of Toronto, and York University for comments and suggestions. I am particularly grateful to Matt Webb and Morten Nielsen for joint work that made this paper possible, and for their comments and feedback, as well as 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 have been included among the regressors. The idea was that any correlation across observations could be accounted for by the fixed effects.

Beginning in the mid 1990s, the assumption of uncorrelated disturbances became less acceptable in empirical work with cross-section data. After the popular econometrics package Stata offered the option of cluster-robust, or "clustered," standard errors, it 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 for various outcomes might be clustered by village, by province, or by country, depending on the nature of the model and dataset. In experimental economics, the disturbances for choices made by experimental subjects might be clustered by subject. There are examples in many other fields. With panel data, clustering by time periods and/or by cross-sectional units sometimes replaces more traditional approaches, such as random effects models. 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.

Nevertheless, whether and how to account for clustered disturbances is still somewhat controversial; see Section 6. In some cases, it is impossible to include cluster fixed effects, because they would be perfectly collinear with one or more explanatory variables. In such cases, it seems to be essential to allow for clustered disturbances. But residuals often show evidence of clustering even when cluster fixed effects are included, and failing to take this into account can lead to standard errors that are seriously misleading because they are far too small; see the empirical example in Section 5.

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 does not attempt to be comprehensive. Instead, it focuses on a few key concepts and issues, and it discusses 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 problematic. It also deals with bootstrap methods, notably the wild cluster bootstrap, that are 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 how investigators should respond. Section 7 discusses the difficult issues that arise when estimating treatment effects,

where cluster-robust inference can be dangerously unreliable. Section 8 briefly deals with instrumental variables estimation of a linear regression model where at least one regressor is endogenous. Section 9 deals with two important issues where more research is needed, and Section 10 concludes.

## 2 Cluster-Robust Covariance Matrices

For simplicity and concreteness, 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.

In writing the model (1), I am making the (arguably very strong) assumptions that each observation belongs to one and only one cluster, that the investigator knows which cluster it belongs to, and that any correlation across observations occurs only within clusters. It is up to the investigator to decide whether these assumptions are reasonable.

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

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

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

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

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

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 $N/(N-K)$ used in the well-known heteroskedasticity-consistent covariance matrix estimator $\text{HC}_1$ (MacKinnon and White 1985) that is robust only to heteroskedasticity of unknown form. Note that $\text{CV}_1$ reduces to $\text{HC}_1$ when each cluster contains just one observation, so that $G = N$.

Covariance matrix estimators like (4) are often referred to as "sandwich estimators." There are two identical pieces of "bread" on the outside and a "filling" in the middle. The filling in the sandwich in (4) is supposed to estimate the corresponding matrix in (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) typically have rank $K$ unless $N_g < K$. 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).

Since each of the matrices in the filling of (4) has rank 1, $\text{CV}_1$ can have rank at most $G$. In some cases, it will have rank $G-1$. This makes it impossible to test hypotheses involving more than $G$, or perhaps $G-1$, restrictions using Wald tests based on (4). Moreover, for hypotheses that involve numbers of restrictions not much smaller than $G$, the finite-sample properties of Wald tests based on (4) are likely to be very poor; see Section 4.

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 (2018). 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 $\boldsymbol{M}_{\boldsymbol{X}} \equiv \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, the $N \times N$ projection matrix that yields OLS residuals. 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; see MacKinnon, Nielsen and Webb (2020b).

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, and 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, because observations that belong to both $\boldsymbol{X}_g$ and $\boldsymbol{X}_h$ contribute to both of the first two terms in (6).

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. 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 et al. (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). It is $O(1/N)$, as we would expect.[1] 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).

Expression (9) is the true variance of $\bar{y}$. Under the usual assumption of uncorrelated disturbances, we would use $s^2/N$ to estimate it, where $s^2$ is the conventional estimate of $\sigma^2$. Unless $N$ is quite small, the quantity $s^2/N$ generally estimates the first term in (9) very well, but it completely ignores the second term. Unfortunately, because the first term is $O(1/N)$ and the second term is $O(1)$, the ratio of the latter to the former increases without bound as $N$ increases. Thus $s^2/N$ massively underestimates $\text{Var}(\bar{y})$ when $N$ is large.

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 at the same rate as $N$, while not letting the $N_g$ change systematically as $N$ increases. 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 (2019).

[1] Here we have used the "same-order" or "big O" notation, which is a convenient way to indicate how a quantity changes with the sample size $N$. The argument of $O(\cdot)$ is $N$ raised to some power. If something does not change with the sample size, then we write that it is $O(1) = O(N^0)$.

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 $N$. This implies that the amount of information about the parameters of interest contained in extremely large samples (such as the ones increasingly encountered in empirical microeconomics) 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, both the way in which observations are distributed across clusters and the characteristics of the $\boldsymbol{X}_g$ matrices can greatly affect the reliability of finite-sample inference. Simple rules about how many clusters are needed for reliable inference have been proposed; for example, Angrist and Pischke (2008) suggests (partly in jest) that 42 clusters is generally sufficient. Unfortunately, this type of rule of thumb can be extremely misleading.

Suppose we are interested in one element of $\boldsymbol{\beta}$, say $\beta_j$. Then 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 generally assumed to follow the Student's $t$ distribution with $G-1$ degrees of freedom. This approximation makes sense, because the filling of the sandwich in (4) is the sum of $G$ matrices that are not independent since the residuals (normally) sum to zero. Thus we are using $G-1$ pieces of (approximately) independent information to estimate the variance of $\hat{\beta}_j$. Bester, Conley and Hansen (2011) proves that the distribution of $t_j$ actually tends to $t(G-1)$ as $N$ becomes large when $G$ is fixed and the limiting matrices for both $\boldsymbol{X}_g'\boldsymbol{X}_g/N$ and $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g/N$ are the same for all $g$.

As the result of Bester et al. (2011) suggests, inference based on (11) and the $t(G-1)$ distribution can sometimes work very well. It generally does so when there are at least 50 clusters and they are reasonably homogeneous, that is, similar in size and with reasonably similar $\boldsymbol{X}_g'\boldsymbol{X}_g$ and $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ matrices. Note that investigators can observe cluster sizes and the $\boldsymbol{X}_g'\boldsymbol{X}_g$ matrices, so two of these conditions can be checked. Unfortunately, severe over-rejection can occur when these conditions are not satisfied; see, among others, MacKinnon and Webb (2017b) and Djogbenou et al. (2019). The latter paper 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 when there are over 200 clusters. This is true even when the largest

cluster is becoming a smaller fraction of the sample as $G$ increases at a rate fast enough for asymptotic theory to be valid.

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 quantity 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 $t(G^*)$ distribution, if desired, although this does not seem to be as reliable as inference based on several other approaches.

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).

Yet another approach is to run a regression that uses cluster averages instead of individual data, as suggested in Donald and Lang (2007). Thus there would be $G$ observations instead of $N$. This can work well when the regressors of interest do not vary within clusters, but it can result in serious loss of power in other cases; see MacKinnon and Webb (2019). Using cluster averages would not make sense for the example of Section 5, where the key regressors (measures of educational attainment) vary mainly at the individual level.

When there are two or more restrictions to be tested, we need to use a Wald test instead of a $t$ test. Suppose we wish to test the hypothesis that $\boldsymbol{R\beta} = \boldsymbol{r}$, where $\boldsymbol{R}$ is an $r \times k$ matrix and $\boldsymbol{r}$ is an $r \times 1$ vector. We can test these $r$ restrictions jointly using the cluster-robust Wald statistic

$$W(\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}\hat{\boldsymbol{V}}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \tag{12}$$

where $\hat{\boldsymbol{V}}$ would usually be $\text{CV}_1$ given in (4). We require that $r \leq G$, because the rank of $\hat{\boldsymbol{V}}$ is at most $G$. Asymptotically, as $G$ gets large, $W(\hat{\boldsymbol{\beta}})$ would be distributed as $\chi^2(r)$. However, this is likely to provide a very poor approximation in finite samples, especially when $r$ is not much smaller than $G$.

In most cases, the best way to perform tests of restrictions on the linear regression model (1) seems to be to use a particular version of the wild bootstrap. We first compute either a $t$ statistic like $t_j$ or a Wald statistic like $W(\hat{\boldsymbol{\beta}})$, then compute a large number of bootstrap test statistics, and finally calculate a bootstrap $P$ value that measures how extreme the actual test statistic is relative to the distribution of the bootstrap test statistics. For example, if 26 out of 999 bootstrap statistics were more extreme than the actual test statistic, the bootstrap $P$ value would be $27/999 = 0.027$.

The key issue for bootstrap testing is how to generate the bootstrap samples. In the case of (1), there are several plausible ways to do so. The best approach usually seems to be to

use the restricted wild cluster bootstrap (WCR) proposed in Cameron, Gelbach and Miller (2008), which I now discuss. MacKinnon and Webb (2017b) studies this method in detail, and Djogbenou et al. (2019) proves that it is asymptotically valid and can provide what is called an "asymptotic refinement" in some circumstances.[2] 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 data generating process (DGP) are based on estimates that satisfy the null hypothesis.

Suppose the objective is to test the restriction $\boldsymbol{a}'\boldsymbol{\beta} = \mathbf{0}$, 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 $\boldsymbol{a}'\boldsymbol{\beta} = \mathbf{0}$ 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

$$\hat{\boldsymbol{V}}_b^* = \frac{G(N-1)}{(G-1)(N-K)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\hat{\boldsymbol{u}}_g^{*b}(\hat{\boldsymbol{u}}_g^{*b})'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}; \tag{13}$$

   (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 $\boldsymbol{a}'\boldsymbol{\beta} \neq \mathbf{0}$ and there is no reason to expect the test statistic to have a non-zero mean under the null hypothesis, compute the symmetric bootstrap $P$ value

$$\hat{P}_{\mathrm{S}}^* = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\big(|t_a^{*b}| > |t_a|\big), \tag{14}$$

   where $\mathbb{I}(\cdot)$ denotes the indicator function. If the test statistic may have a non-zero mean, as in the case of a regression estimated by instrumental variables, it would be better to compute an equal-tail $P$ value, which is twice the minimum of the upper-tail and lower-tail bootstrap $P$ values; see Section 8.

[2] When a bootstrap test provides an asymptotic refinement, its performance improves faster as $N$ (or in this case $G$) increases than does the asymptotic test on which it is based.

For the Wald test based on $W(\hat{\boldsymbol{\beta}})$, we would compute the bootstrap statistic $W(\hat{\boldsymbol{\beta}}^{*b})$ instead of $t_a^{*b}$ in step 3(d), and we would compute an upper-tail $P$ value in step 4.

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 this case and many others, bootstrap tests perform better when the bootstrap samples impose the null hypothesis; see Davidson and MacKinnon (1999). Notice that, unlike the familiar pairs bootstrap, the wild cluster bootstrap does not involve resampling the data. The $\boldsymbol{X}$ matrix is identical for every bootstrap sample; only the $\boldsymbol{y}^{*b}$ vectors vary. The pairs bootstrap cannot be used for models like (1) because resampling by observation would destroy the clustered structure of the data. However, the pairs cluster bootstrap, in which entire clusters are resampled, can be used; see the end of this section and Section 9.

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 $v^* = -1$, each with probability 0.5; see MacKinnon (2015). Why the Rademacher distribution is particularly attractive is discussed in Djogbenou et al. (2019). It is not a good idea to use a 2-point distribution like the Rademacher when $G$ is very small, however, because the number of distinct bootstrap samples is just $2^G$; see Webb (2014), which suggests a 6-point distribution for use in such cases.

Provided the number of clusters is not too large, it is possible to generate a large number of wild cluster bootstrap test statistics very efficiently. This is what the Stata routine `boottest` does; see Roodman, MacKinnon, Nielsen and Webb (2019). 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 (13). All of the computations that are $O(N)$ are done just once, rather than for every bootstrap sample. Therefore, for large $B$, the computational cost of computing a WCR bootstrap $P$ value is approximately $O(G^2B)$ instead of $O(NB)$. In consequence, when $G \ll N$, as is the case for the empirical example of Section 5, using the WCR bootstrap can be remarkably inexpensive. Even though this example involves well over a million observations and 57 coefficients, it takes less than 30 seconds to calculate a bootstrap $P$ value based on 99,999 bootstraps for either 36 or 51 clusters.

Because using `boottest` is so inexpensive in most cases, it probably makes sense to employ the restricted wild cluster bootstrap by default whenever making cluster-robust inferences using Stata. If the WCR bootstrap $P$ value is similar to the one based on the $t(G-1)$ distribution, then we can be fairly confident that it is reliable. If they differ substantially, however, it is very likely that the former is more accurate than the latter, but it is by no means certain that it is sufficiently accurate. Unfortunately, the WCR bootstrap can be seriously unreliable in certain cases; see Section 7.

Confidence intervals can easily be obtained by inverting a bootstrap test, and `boottest` does this by default. A 95% bootstrap confidence interval is simply the set of parameter values for which the bootstrap $P$ value is greater than 0.05. Finding such an interval requires iteration, with the bootstrap $P$ value computed for a number of potential upper and lower

limits. MacKinnon (2015) provides simulation evidence which suggests that confidence intervals based on inverting WCR bootstrap tests work better than several alternatives.

Even though the disturbances for bootstrap samples generated by the wild cluster bootstrap are clustered in only one dimension, it can also be used in conjunction with two-way clustered standard errors. MacKinnon et al. (2020b) shows that the WCR bootstrap often works well in this case, and `boottest` makes it easy to do this.

Other bootstrap methods can also be used. MacKinnon and Webb (2018) argues that the ordinary wild bootstrap can sometimes work better than the wild cluster bootstrap when interest focuses on a treatment dummy; see Section 7.[3] 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 et al. (2008) and MacKinnon and Webb (2017a). However, it has the advantage that it can be used for nonlinear models like the probit model. The pairs cluster bootstrap will be discussed briefly in that context in Section 9.

# 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 is the logarithm of weekly earnings for men aged 25 to 65, conditional on earnings being greater than \$20 (not adjusted for inflation). 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

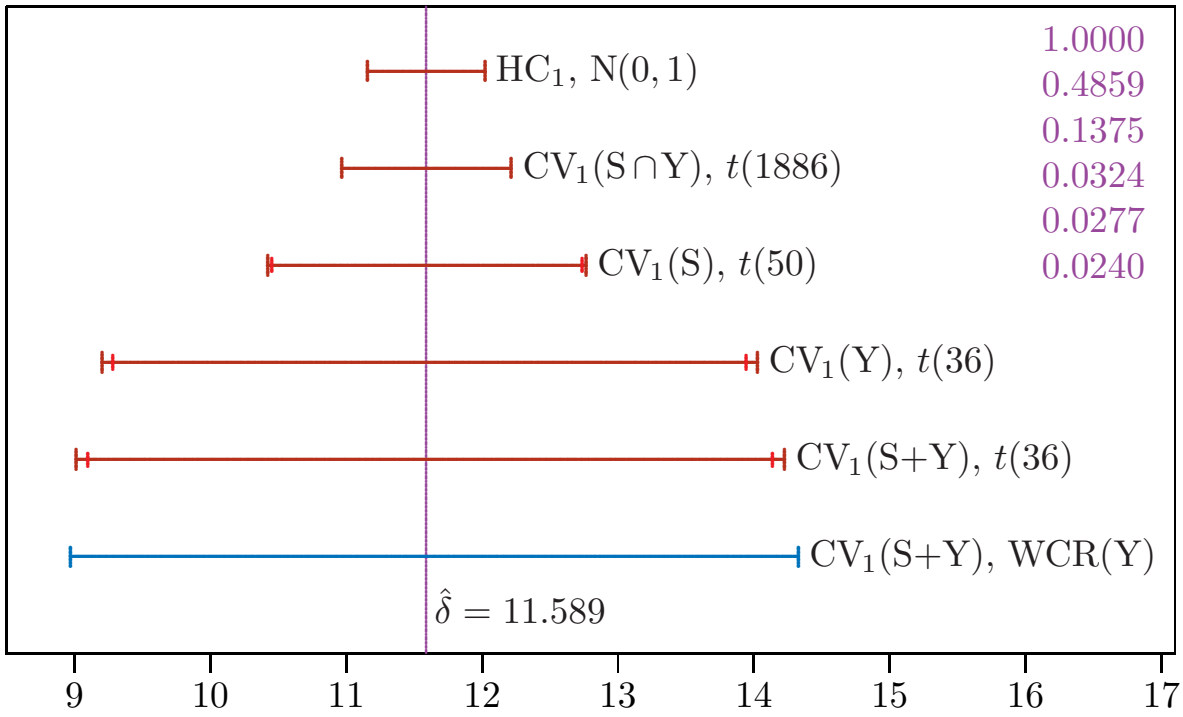
$$y_{gti} = \beta_1 + \sum_{j=2}^{5} \beta_j \texttt{ED}j_{gti} + \beta_6 \texttt{Age}_{gti} + \beta_7 \texttt{Age}^2_{gti} + \sum_{s=1}^{36} \gamma_s \texttt{Year}^s_t + \sum_{k=1}^{50} \eta_k \texttt{State}^k_g + u_{gti}, \quad (15)$$

where $g$ denotes the state, $t$ denotes the year, and $i$ denotes the individual. In equation (15), $\texttt{Year}^s_t$ is a dummy that equals 1 when $s = t$, and $\texttt{State}^k_g$ is a dummy that equals 1 when $g = k$. One year dummy and one state dummy are omitted to avoid perfect collinearity, and the year dummy variables are "absorbed" to save computer time, leaving 57 coefficients to estimate. This equation could be used to answer various economic questions. For concreteness, I focus on the value of obtaining a postgraduate degree.[4]

[3] Unfortunately, the tricks that `boottest` uses to save computer time are not effective for the ordinary wild bootstrap.

[4] 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 1: Confidence Intervals for $\hat{\delta}$




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

$$\hat{\delta} = 100\big(\exp(\hat{\beta}_5 - \hat{\beta}_4) - 1\big) = 100\big(\exp(0.78727 - 0.67762) - 1\big) = 11.589\%. \qquad (16)$$

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 (16), we need a standard error. 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 HC$_1$ standard error is 0.2215, which suggests that we have estimated $\delta$ 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 (15) 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 at a lower level, or perhaps some more complicated pattern of correlated disturbances, the fixed effects would not eliminate all within-cluster correlation. Thus it seems plausible that there may be within-cluster correlations among the disturbances.

Figure 1 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 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 given above and the critical value 1.96, which is the 0.975 quantile of the standard normal distribution. This is probably the interval that most investigators would have used until about the year 2000.

The second interval shown in Figure 1 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 0.975 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 below.

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.[5]

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. 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. An exception is the finance literature (Thompson 2011), where two-way clustering by firm (or asset) and time period is not uncommon.

The belief that clustering by state is safer than clustering by the intersection of state and year appears to have originated with simulation results in Bertrand et al. (2004). This paper was the first to employ "placebo-law" experiments, in which every replication uses the same data for the regressand and all but one of the regressors. The only thing that differs across replications is the regressor of interest, a treatment dummy that affects certain states in certain years. Since the treatment dummies are generated randomly, we would expect valid statistical procedures to reject the null hypothesis about as often as the level of the test. Instead, when the standard errors are either heteroskedasticity-robust or robust only to state-year clustering, the tests over-reject very severely.

[5]Sharp-eyed readers may notice that the intervals in Figure 1 appear to be slightly asymmetric. That is because they were computed by forming symmetric intervals for $\beta_5 - \beta_4$ and then converting those into asymmetric intervals for $\delta$ by applying the function $100\big(\exp(\cdot) - 1\big)$ to the upper and lower limits.

MacKinnon (2016) performs placebo-law experiments using the dataset of this paper, based on equation (15) augmented by a random treatment dummy. These experiments confirm the findings of Bertrand et al. (2004) that there is very severe over-rejection when standard errors account only for heteroskedasticity or only for clustering at the state-year level. They also suggest that the extent of over-rejection depends strongly on the sample size. The larger the sample, the more the tests over-reject. Unfortunately, the experiments in MacKinnon (2016) do not consider clustering by year or two-way clustering.

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. Clustering the bootstrap samples by state yields extremely similar results that are not reported.[6] 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 1 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 et al. (2019) and the discussion in Section 4.

It may seem that we have to choose among the six intervals in Figure 1 somewhat arbitrarily. However, as I discuss below, 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. Considering the sample size, these seem remarkably wide, a point that will be discussed at the end of this section.

Ideally, we could perform a statistical test to determine which level of clustering, and therefore which confidence interval, is appropriate. How to perform such a test is an area of active research. Ibragimov and Müller (2016) proposes a test between one-way clustering at a low level (or no clustering at all) against one-way clustering at a higher level, 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 (2020a).

The top interval in Figure 1, the one based on $\text{HC}_1$, would be appropriate if there were actually no intra-cluster correlation. In that case, the matrix $\boldsymbol{\Omega}$ would be diagonal, and all the intervals would be valid. In the special case of the sample mean, which was discussed in Section 3, $\text{HC}_1$ would estimate only the first term in equation (9), setting the entire second term to zero. The various cluster-robust estimators would instead estimate both the first term and some of the covariances in the second term, setting others to zero. The two-way

[6] Although the bootstrap samples exhibit one-way clustering by either year or state, the CRVE that is computed for both the actual and bootstrap samples has two-way clustering.

CRVE would estimate the largest number of covariances, but even it would set most of them to zero.

If $\mathbf{\Omega}$ were actually diagonal, we would expect the estimated covariances to be, on average, zero. But it is clear from Figure 1 that they are actually positive, on average, because the cluster-robust standard errors are getting larger and larger as we estimate more and more covariances. For clustering by state, by year, and by both state and year, there are a great many of these covariances. It seems highly unlikely that the true covariances could be zero, on average, when the estimates of their average are so large.

As noted above, the confidence intervals that seem most plausible are surprisingly wide. This is because the sample actually contains much less information than it initially appears to. The numbers in the upper right of Figure 1 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.

To investigate this issue further, I reduced the sample size from 1,156,597 to 72,288 by randomly throwing away 15 out of every 16 observations without attempting to preserve the relative sample sizes for each state-year combination. The $\text{HC}_1$ standard error increased by 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 by a factor of just 1.81. Even more surprisingly, the standard error based on two-way clustering increased by a factor of only 1.098. Thus throwing away 15/16 of the sample increased the standard error by just under 10%.

This implies that, for equation (15) 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. The reasons probably vary across models and datasets. In many cases, it seems reasonable to believe that there are unobserved quantities which affect some or all observations within each cluster. 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).

For many years, applied econometricians believed that including cluster fixed effects would eliminate the source of intra-cluster correlation. However, as the example of Section 5 illustrates, cluster fixed effects often explain only part of any intra-cluster correlation. There are many ways in which this could happen. For example, perhaps each cluster contains many unobserved subclusters, each with its own unobserved effect. The cluster fixed effect would account for the average of the subcluster effects, but not for variation around that average. The remaining correlation within subclusters would show up as within-cluster correlation. For related discussion, see Cameron and Miller (2015).

More generally, all sorts of model misspecification might 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, especially when the clusters are large, 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 (15) cannot possibly explain 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. These assumptions seem 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. The assumptions also seem reasonable for villages or hospitals. Conditional on the chosen clusters, the sample may contain all the observations for each cluster, or just some subset of them.

However, the empirical example of Section 5 does not really satisfy the assumptions discussed in the previous paragraph. 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 I 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 aggregate time-series and panel data.

Abadie, Athey, Imbens and Wooldridge (2017) has recently argued that many economic datasets do not satisfy the assumption that the sample is 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 I have speculated that much of the within-cluster correlation that we observe arises 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 necessarily be avoided. This is clearly an area where more work is needed.

# 7 Inference about Treatment Effects

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 (DiD) regressions when all the treated observations belong to just a few clusters. It has been known for some time that using cluster-robust $t$ statistics leads to serious over-rejection in such cases; see, among others, Abadie, Diamond and Hainmueller (2010) and Conley and Taber (2011).

Precisely why this happens is explained in MacKinnon and Webb (2017b, Section 6). Here I provide a somewhat simpler argument.

The reason why cluster-robust inference about treatment effects can sometimes fail may be seen in the context of the simple regression model

$$y_{gi} = \beta + \delta d_{gi} + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{17}$$

where $d_{gi}$ is a dummy variable that equals 1 for treated observations and 0 otherwise. Suppose initially that every observation belongs to its own cluster, so that $N = G$. Then if $d_{ii} = 0$ for $G_0$ observations and $d_{ii} = 1$ for $G_1 = G - G_0$ observations, and neither $G_0$ nor $G_1$ is small, there is no problem. We can use the OLS estimate $\hat{\delta}$ and its heteroskedasticity-robust standard error to make inferences.

Suppose, however, that $G_1 = 1$, where (without loss of generality) the treated observation is number 1. Then $\hat{u}_{11} = 0$, which implies that the heteroskedasticity-robust covariance matrix is singular. In fact, because of the singularity, it can be shown that $\widehat{\text{Var}}(\hat{\delta}) = \widehat{\text{Var}}(\hat{\beta})$. It is clear that this estimator of $\text{Var}(\hat{\delta})$ must typically be much too small, because $\widehat{\text{Var}}(\hat{\beta})$ is consistent, and $\beta$ will usually be estimated with much greater precision than $\delta$. Thus $t$ tests based on $\widehat{\text{Var}}(\hat{\delta})$ are sure to over-reject very severely. Of course, it should come as no surprise that we cannot test the hypothesis $\delta = 0$ with $G_1 = 1$ when we allow for heteroskedasticity of unknown form. There is simply no way to distinguish between $\delta$ being non-zero and $u_{11}$ being large.

Now suppose that the data are divided into clusters, with $N_1$ observations in the first one. It is no longer the case that $\hat{u}_{11} = 0$. Instead, $\sum_{j=1}^{N_1} \hat{u}_{1j} = 0$, because the dummy variable must be orthogonal to the residuals. For heteroskedasticity-robust inference, this would not be much of a problem unless $N_1$ were very small. But for cluster-robust inference, it is. The CRVE is singular, and $\widehat{\text{Var}}(\hat{\delta})$ is once again equal to $\widehat{\text{Var}}(\hat{\beta})$. In the usual case in which $N_1$ is small relative to $N$, this implies that $\widehat{\text{Var}}(\hat{\delta})$ is much too small. Thus, as before, $t$ tests based on $\widehat{\text{Var}}(\hat{\delta})$ are sure to over-reject severely.

It is natural to hope that the bootstrap would solve the problem. Unfortunately, as MacKinnon and Webb (2017b) shows, it does not do so. In fact, the WCR bootstrap tends to under-reject, usually very severely, whenever there are few treated clusters. Just how much it under-rejects depends on the sizes of the treated and untreated clusters. The WCU bootstrap, in which unrestricted instead of restricted residuals are used in the bootstrap DGP, has the opposite problem. For just one treated cluster, it over-rejects about as severely as comparing the cluster-robust $t$ statistic to the $t(G-1)$ distribution.

MacKinnon and Webb (2018) discusses various methods that can work better than the WCR bootstrap. In particular, that paper suggests employing the ordinary wild bootstrap. Djogbenou et al. (2019) proves that combining the ordinary wild bootstrap with cluster-robust standard errors is asymptotically valid. There are cases in which doing so can yield quite reliable inferences, even with just one or two treated clusters. However, there are also cases in which inferences are not very reliable at all.

Another approach, pioneered in Conley and Taber (2011) and studied in more detail in MacKinnon and Webb (n.d., 2019), is to use randomization inference, or RI. The idea is to

compare the actual test statistic (or actual parameter estimate) with the distribution of a (preferably large) number of test statistics (or parameter estimates) computed by randomly assigning treatment to clusters that were not actually treated. These procedures can work very well, especially when there are at least two treated clusters and the clusters are not too dissimilar in size. Even when they do not perform particularly well, RI procedures seem not to over-reject as severely as comparing cluster-robust $t$ statistics to the $t(G-1)$ distribution or under-reject as severely as using the WCR bootstrap.

The papers by Matthew Webb and myself cited above contain Monte Carlo simulations and/or placebo-law experiments that show just how badly standard methods can perform when there are very few treated clusters. Additional simulation results may be found in MacKinnon and Webb (2017a), which provides a reasonably accessible discussion of why $t$ tests and bootstrap tests fail in this case. It also contains simulation results for the pairs cluster bootstrap mentioned in Sections 4 and 9. I therefore do not present any simulation results in this section. Instead, I briefly discuss an empirical example which illustrates how inference about treatment effects can be extremely sensitive to assumptions about the appropriate level of clustering.

The example uses a model estimated in Decarolis (2014). The treatment is a change from average-bid auctions (ABA) to first-price auctions (FPA) for public works contracts in Italian cities. Only one municipality made this change during the sample period, namely, Turin. There are 1262 observations, but only 15 municipalities. Turin used ABA in the years 2000–2002 and FPA in 2003–2006. Other municipalities used ABA in all periods. Thus the FPA treatment dummy is equal to 1 only for Turin in the years 2003–2006. The dependent variable is the "discount" over the reserve price. A positive coefficient on the treatment dummy indicates lower bids.

Decarolis (2014) reports a coefficient of 6.136 on the FPA dummy. Whether or not this is significant depends on how the disturbances are assumed to be clustered. The paper clusters them by the intersection of city and year, so that there are 105 clusters, of which 4 are treated. This leads to a standard error of 1.305 and a $t$ statistic of 4.703. If instead we allow for heteroskedasticity but do not cluster at all, the standard error is 2.277, and the $t$ statistic is 2.695. On the other hand, if we cluster at the city level, so that there are 15 clusters of which just one is treated, the standard error is 0.785, and the $t$ statistic is 7.817.

These results and $P$ values computed in various ways are reported in Table 1. They are quite worrying. The theory of Section 3 suggests that standard errors should become larger when we cluster more coarsely, unless the additional correlations that are estimated with coarser clustering are actually zero. The empirical example of Section 5 displays precisely this pattern. But in Table 1, the standard error is smaller for clustering at the city-year level than for no clustering, and smaller still for city-level clustering. This is not a surprise for the city-level case, because one treated cluster out of 15 is an extreme example of the few-treated problem. However, the fact that the standard error drops from 2.277 to 1.305 when we move from no clustering to city-year clustering provides strong evidence that 4 treated clusters out of 105 is too few for making reliable inferences.

What can we conclude? Unless we believe that there is actually no clustering, which seems very unlikely, we evidently cannot rely on the first set of results. The assumptions

Table 1: $P$ values for FPA coefficient

| Clustering | Std. Error | $t$ Statistic | $P$ Value | Method |
|---|---|---|---|---|
| None | 2.277 | 2.695 | 0.0070 | $N(0,1)$ |
| | | | 0.0089 | WR |
| City/Year | 1.305 | 4.703 | 0.0000 | $t(104)$ |
| | | | 0.1765 | WCR |
| | | | 0.0000 | WCU |
| | | | 0.0206 | WR |
| City | 0.785 | 7.817 | 0.0000 | $t(14)$ |
| | | | 0.1305 | WCR |
| | | | 0.0000 | WCU |
| | | | 0.0574 | WR |

**Notes:** WCR is the restricted wild cluster bootstrap discussed in Section 4. WCU is the unrestricted wild cluster bootstrap, which uses unrestricted residuals in step 3(a) of the algorithm given there. WR is the ordinary wild restricted bootstrap, in which there is one realization of the auxiliary random variable per observation instead of one per cluster. Bootstrap $P$ values are based on 99,999 bootstrap samples.

behind the second and third sets seem more plausible, but, based on the theory in MacKinnon and Webb (2017b, 2018), we clearly cannot rely on either the $t(G-1)$ distribution or the WCU bootstrap, both of which tell us to reject the null hypothesis. The same theory implies that we also cannot rely on the WCR bootstrap, which tells us not to reject it. The only results that might be reliable are the ones from the WR bootstrap, which are somewhat ambiguous. I tentatively conclude that there is modest, but far from compelling, evidence that the switch to FPA increased bidding discounts. Results from some other methods that use city-level clustering, presented in MacKinnon and Webb (2019), suggest that we can reject the null hypothesis at the 10% level, but not at the 5% level.

# 8 Simultaneous Equations

Most of the work on cluster-robust inference has assumed that all regressors are exogenous. When, in addition, some of the regressors are endogenous, investigators have to deal with two issues instead of one. Instrumental variables (IV) estimation, and other methods that allow simultaneous equations models to be estimated consistently, can yield seriously biased estimates and extremely unreliable confidence intervals when the instruments are weak.

There is an enormous literature on inference with weak instruments, which began with Nelson and Startz (1990a, b). Particularly influential papers include Staiger and Stock (1997), Kleibergen (2002), Moreira (2003), and Andrews, Moreira and Stock (2006). However, these papers and most others in this literature do not even allow for heteroskedasticity, let alone clustering. One bootstrap procedure that allows for heteroskedasticity of unknown form is a variant of the wild bootstrap proposed in Davidson and MacKinnon (2010). This procedure was extended to allow for clustering in Finlay and Magnusson (2019).

For concreteness, consider the two-equation model

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{Z}_1 \boldsymbol{\gamma} + \boldsymbol{u}_1 \tag{18}$$

$$\boldsymbol{y}_2 = \boldsymbol{Z}_1 \boldsymbol{\pi}_1 + \boldsymbol{Z}_2 \boldsymbol{\pi}_2 + \boldsymbol{u}_2 = \boldsymbol{Z} \boldsymbol{\pi} + \boldsymbol{u}_2, \tag{19}$$

in which there are two endogenous variables, the observations on which are contained in the $N$-vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Here (18) is the structural equation in which we are interested, and (19) is an unrestricted reduced form equation for the other endogenous variable. There are $N$ observations, and the matrices $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ have $K-1$ and $L-K+1$ columns, respectively. Each column contains the data for one exogenous variable. Thus the total number of exogenous variables in the matrix $\boldsymbol{Z} = [\boldsymbol{Z}_1 \;\; \boldsymbol{Z}_2]$ is $L \geq K$. When $L > K$, the model is overidentified.

The OLS estimate of $\beta$ is inconsistent whenever the disturbances $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are correlated. Equations like (18) are therefore typically estimated by IV, or perhaps by some asymptotically equivalent method, such as limited-information maximum likelihood (LIML). This generally works fine when the instruments are strong, which means that dropping $\boldsymbol{Z}_2$ from equation (19) would substantially reduce the explanatory power of that regression. When the instruments are weak, however, conventional asymptotic theory typically provides a poor approximation to the finite-sample properties of the IV estimate $\hat{\beta}$ and its standard error.

To obtain IV estimates, we replace $\boldsymbol{y}_2$ in equation (18) by $\boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y}_2$, where $\boldsymbol{P}_{\boldsymbol{Z}}$ is the projection matrix $\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, so that $\boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y}_2$ is the vector of fitted values from regressing $\boldsymbol{y}_2$ on $\boldsymbol{Z}$. Thus we use OLS to estimate the equation

$$\boldsymbol{y}_1 = \beta \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y}_2 + \boldsymbol{Z}_1 \boldsymbol{\gamma} + \boldsymbol{u}_1. \tag{20}$$

Equation (20) could be written in the same form as equation (1) if we defined $\boldsymbol{X}$ to be the $N \times K$ matrix $[\boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y}_2 \;\; \boldsymbol{Z}_1]$. If we interpret $\boldsymbol{X}$ in this way, the top left element of the CRVE (4) gives us the square of the cluster-robust standard error of $\hat{\beta}$.[7] The cluster-robust IV $t$ statistic $t_\beta$ is simply $\hat{\beta} - \beta_0$ divided by this standard error.

The bootstrap method proposed in Davidson and MacKinnon (2010) is called the WRE bootstrap (for "wild, restricted, efficient"). When modified to allow for clustered disturbances, it becomes the WCRE bootstrap, which works very much like the WCR bootstrap described in Section 4. There are two key differences. The first is that the bootstrap DGP needs estimates of the parameters of both equations. For the structural equation (18), it uses the null value $\beta_0$, restricted estimates $\tilde{\boldsymbol{\gamma}}$ obtained by an OLS regression of $\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2$ on $\boldsymbol{Z}_1$, and restricted residuals $\tilde{\boldsymbol{y}}_1$ from that same regression. For the reduced-form equation (19), it uses efficient estimates $\tilde{\boldsymbol{\pi}} = [\tilde{\boldsymbol{\pi}}_1' \;\; \tilde{\boldsymbol{\pi}}_2']'$ obtained by regressing $\boldsymbol{y}_2$ on $\boldsymbol{Z}$ and $\tilde{\boldsymbol{u}}_1$, and residuals $\tilde{\boldsymbol{u}}_2 = \boldsymbol{y}_2 - \boldsymbol{Z}\tilde{\boldsymbol{\pi}}$. Including the residuals $\tilde{\boldsymbol{u}}_1$ in the equation to estimate $\tilde{\boldsymbol{\pi}}$ yields more efficient estimates than estimating (19) directly when $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are correlated, which they must be whenever OLS estimation of (18) is inconsistent. It is also possible to obtain

[7] Strictly speaking, the degrees-of-freedom correction should be different, but this will not matter if we use the bootstrap and apply the same correction to the bootstrap test statistics as to the actual one.

efficient estimates that are asymptotically equivalent to $\tilde{\boldsymbol{\pi}}$ by estimating equations (18) and (19) jointly by maximum likelihood, and this is what `boottest` does.

The second way in which the WCRE bootstrap differs from the WCR bootstrap is that each realization of the auxiliary random variable $v_g^*$ multiplies the residuals from both equations for cluster $g$. This ensures that the bootstrap DGP retains both the correlations between the structural and reduced-form equations and the intra-cluster correlations. For cluster $g$, steps 3(a) and 3(b) in the WCR bootstrap algorithm are replaced by

$$\boldsymbol{y}_{2gi}^* = \boldsymbol{Z}_{gi}\tilde{\boldsymbol{\pi}} + v_g^*\tilde{\boldsymbol{u}}_{2gi}, \tag{21}$$

$$\boldsymbol{y}_{1gi}^* = \beta_0\boldsymbol{y}_{2gi}^* + \boldsymbol{Z}_{1gi}\tilde{\boldsymbol{\gamma}} + v_g^*\tilde{\boldsymbol{u}}_{1gi}. \tag{22}$$

The order of the two equations has been reversed here to emphasize the fact that we have to generate $\boldsymbol{y}_{2gi}^*$ before we can generate $\boldsymbol{y}_{1gi}^*$.

We can test the hypothesis $\beta = \beta_0$ by generating one IV $t$ statistic $t_\beta$ and $B$ bootstrap $t$ statistics, say $t_\beta^{*b}$, using bootstrap samples generated from equations (21) and (22). These are used to calculate the equal-tail $P$ value

$$\hat{P}_{\text{ET}}^* = \min\left(\frac{2}{B}\sum_{b=1}^{B}\mathbb{I}\big(t_\beta^{*b} > t_\beta\big), \frac{2}{B}\sum_{b=1}^{B}\mathbb{I}\big(t_\beta^{*b} \leq t_\beta\big)\right).$$

Because $t_\beta$ does not have mean zero in finite samples under the null hypothesis, this test will have better properties than one based on the symmetric $P$ value (14). In addition, confidence intervals obtained by inverting it will be asymmetric and should be more accurate.

The WCRE bootstrap can be used with other test statistics. In particular, Finlay and Magnusson (2019) investigates its use with a cluster-robust version of the AR statistic proposed in Anderson and Rubin (1949). The `boottest` package computes WCRE bootstrap tests based on both AR statistics and IV $t$ statistics, for the model (18) and (19) and also for similar models with two or more right-hand-side endogenous variables.[8] The computations for the AR statistic can take advantage of the tricks that make the WCR bootstrap so fast (see Section 4), but the ones for $t_\beta$ cannot do so; see Roodman et al. (2019). Thus, in large samples, it can be much faster to bootstrap the cluster-robust AR statistic than the cluster-robust IV $t$ statistic. The AR statistic is also unaffected by weak instruments. Indeed, when the disturbances are homoskedastic and normally distributed, the original AR statistic follows the $F(L-K+1, N-L)$ distribution in finite samples.

The previous paragraph appears to suggest that it is preferable to use bootstrap AR tests instead of bootstrap IV $t$ tests. However, AR tests must be used with care. They are implicitly testing two hypotheses, one that $\beta = \beta_0$ and the other that $\boldsymbol{Z}_2$ does not appear in equation (18). The latter hypothesis is that the overidentifying restriction(s) hold. This implies that an AR test may reject the null hypothesis for the wrong reason. Moreover, inverting an AR test, whether bootstrapped or not, can yield confidence intervals that are extremely misleading. They may be too long, too short, or even empty, depending on how much evidence there is against the overidentifying restrictions. For a detailed discussion,

[8] `boottest` can also compute WCRE bootstrap tests based on several other estimators, including LIML.

see Davidson and MacKinnon (2014b). In contrast, confidence intervals based on inverting WCRE bootstrap $t$ tests seem to work well unless the instruments are extremely weak; see Davidson and MacKinnon (2014a).

# 9 Some Open Issues

Despite the large amount of work on cluster-robust inference in recent years, there are still a number of issues where results are lacking. In this section, I briefly discuss two of them.

Most surveys report sample weights, and it is very common to run weighted regressions that employ them. If observation $i$ has weight $w_i$, then it represents $w_i$ observations, perhaps because it was selected with probability proportional to $1/w_i$. Weighted least squares means replacing $y_i$ and $\boldsymbol{X}_i$ by $y_i^\circ = \sqrt{w_i} y_i$ and $\boldsymbol{X}_i^\circ = \sqrt{w_i}\boldsymbol{X}_i$, respectively. Thus the weighted least squares estimates are

$$\hat{\boldsymbol{\beta}}_{\text{w}} = (\boldsymbol{X}^{\circ\prime}\boldsymbol{X}^\circ)^{-1}\boldsymbol{X}^{\circ\prime}\boldsymbol{y}^\circ = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}, \tag{23}$$

where $\boldsymbol{W}$ is an $N \times N$ diagonal matrix with $w_i$ as the $i^{\text{th}}$ diagonal element.

From the perspective of asymptotic theory, it is perfectly valid to treat $\boldsymbol{y}^\circ$ and $\boldsymbol{X}^\circ$ as if they were the original data. The weighting will inevitably introduce heteroskedasticity, but every CRVE already allows for heteroskedasticity. Thus, in principle, we can simply deal with weighted regressions as if they were unweighted regressions. This is also true if we use any version of the wild bootstrap, and the `boottest` package in Stata is designed to work with regressions that explicitly use weights.

Although sample weights cause no problem asymptotically, they almost certainly make finite-sample inference less reliable. Unfortunately, I am not aware of any research that has explicitly studied cluster-robust inference for weighted regressions. MacKinnon and Webb (2018) provides some evidence that a particular form of heteroskedasticity can seriously harm the finite-sample properties of the WCR bootstrap in a model of treatment effects. On the other hand, Djogbenou et al. (2019) provides evidence that a different form of heteroskedasticity has only a modest effect on those properties in a model with a continuous regressor. These contradictory results are not very helpful.

What scholars who employ Canadian survey data really need is a study of the type of heteroskedasticity caused by sample weights, for models with either 10 or 13 clusters that vary in size like actual samples containing data for Canadian provinces (or provinces and territories). In the absence of such a study, it would be wise for anyone using regressions with sample weights to perform their own simulation study based on the actual sample they are using, including the actual sample weights. This advice applies in particular to models of treatment effects, including DiD models, where inference may also be problematic in the Canadian context because there are likely to be few treated (or untreated) clusters and because cluster sizes may vary greatly.

A second issue that needs more study is what to do when the dependent variable is binary. One approach is simply to estimate a linear probability model (LPM) using least squares and proceed in the usual way. This allows one to make cluster-robust inferences based on (4) or some other CRVE and to employ the wild cluster bootstrap, if desired. A second

approach, which is perhaps more appealing, is to estimate a binary response model that constrains the fitted values to lie between 0 and 1. But this requires a CRVE designed for the binary response model, along with an appropriate bootstrap method when the number of clusters is not large.

A binary response model with a linear index function can be written as

$$P_i \equiv \mathrm{E}(y_i \,|\, \boldsymbol{X}_i) = F(\boldsymbol{X}_i\boldsymbol{\beta}), \tag{24}$$

where $y_i$ equals either 0 or 1, $\boldsymbol{X}_i$ is a row vector of observations on explanatory variables, and $\boldsymbol{\beta}$ is a vector of coefficients to be estimated. The transformation function $F(\cdot)$ maps from the real line to the [0-1] interval. In most applications, it is either the logistic function or the cumulative standard normal distribution function; these yield the logit model and the probit model, respectively. In practice, both of these models generally produce very similar results, except for the scale of the parameter estimates.

When $\boldsymbol{X}_i$ includes variables that can take on relatively extreme values, the function $F(\cdot)$ in (24) plays a very important role by ensuring that $0 < P_i < 1$. However, in many cases, the (unknown) true values of the $P_i$ are well away from both 0 and 1, and all of the regressors are dummy variables. In such cases, least squares typically yields estimated probabilities that lie in the [0-1] interval and are quite similar to the ones from a binary response model. Thus it is common, and not very harmful, for investigators to estimate an LPM when the nonlinear model (24) would be more appropriate.

If it is appropriate to use an LPM, and the conditions for CRVE-based inference to perform well are satisfied (see the discussion in Section 4), then cluster-robust $t$ statistics should be fairly reliable. If not, one possibility is to use the wild cluster bootstrap. For any form of wild bootstrap that uses the Rademacher distribution, the bootstrap dependent variable can take on only two values, each with probability $\frac{1}{2}$. If $\hat{P}_i$ denotes the fitted value from the LPM, these are

$$y_i^* = \hat{P}_i + (y_i - \hat{P}_i) = y_i \quad \text{and} \quad y_i^* = \hat{P}_i - (y_i - \hat{P}_i) = 2\hat{P}_i - y_i. \tag{25}$$

The first value here is just the actual value of $y_i$, which is 0 or 1. But the second is either $2\hat{P}_i$ or $2\hat{P}_i - 1$. Thus the $y_i^*$ must look very different from the $y_i$ and are sure to lie outside the [0-1] interval in many cases. However, they do at least have the right expectation, since

$$\tfrac{1}{2}\mathrm{E}(y_i) + \tfrac{1}{2}\big(2\hat{P}_i - \mathrm{E}(y_i)\big) = \tfrac{1}{2}(2\hat{P}_i) = \hat{P}_i, \tag{26}$$

where $\mathrm{E}(y_i) = \hat{P}_i$ because the bootstrap DGP is conditional on the sample. At present, unfortunately, it is unknown whether the wild cluster bootstrap yields reasonably reliable inferences in the linear probability model, or even whether it yields less unreliable inferences than comparing cluster-robust $t$ statistics to the $t(G-1)$ distribution.

If we estimate a logit or probit model, a good way to obtain cluster-robust standard errors for the elements of $\hat{\boldsymbol{\beta}}$ is to use the square roots of the diagonals of the matrix

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \big(\boldsymbol{X}'\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\big)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{s}_g(\hat{\boldsymbol{\beta}})\boldsymbol{s}_g'(\hat{\boldsymbol{\beta}})\right)\big(\boldsymbol{X}'\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\big)^{-1}, \tag{27}$$

where $\boldsymbol{s}_g$ is the score vector for cluster $g$, of which a typical element is

$$s_{gj} = \sum_{i=1}^{N_g} \left( \frac{y_{gi}}{F(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})} + \frac{y_{gi} - 1}{1 - F(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})} \right) f(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}}) X_{gij}, \quad j = 1, \ldots, k, \tag{28}$$

and $\boldsymbol{\Upsilon}(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix with typical diagonal element

$$\Upsilon_i(\boldsymbol{\beta}) \equiv \frac{f^2(\boldsymbol{X}_i\boldsymbol{\beta})}{F(\boldsymbol{X}_i\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_i\boldsymbol{\beta})\big)}. \tag{29}$$

If there were no clustering, the covariance matrix would simply be $\big(\boldsymbol{X}'\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\big)^{-1}$. The built-in routines for binary response models in Stata compute an estimated covariance matrix that is asymptotically equivalent to (27). These routines replace $\boldsymbol{X}'\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})\boldsymbol{X}$ by the Hessian of the loglikelihood function. In my view, this is not the best thing to do, because the Hessian is noisier than what it replaces, but using the Hessian is probably not very harmful when the sample size is reasonably large.

When the number of clusters is large, both the matrix (27) and its counterpart in Stata probably yield reasonably accurate inferences. However, they surely cannot do so in all cases. Consider a sample where the dependent variable is continuous and inference based on cluster-robust $t$ statistics is unreliable. Then replacing the continuous dependent variable by a binary one and using, say, a probit model instead of a linear regression model is not going to improve the properties of cluster-robust $t$ statistics. If we need to use the bootstrap in the former case, we almost certainly need to use it in the latter case as well.

However, it is not entirely clear what sort of bootstrap to use. The wild cluster bootstrap is not available, because binary response models do not have residuals that can be used to generate bootstrap samples. The pairs cluster bootstrap is appealing, but it has some deficiencies. First, it is likely to be expensive, because we need to estimate a probit model for every bootstrap sample. Second, if cluster sizes vary a lot, then so will the sizes of the bootstrap samples. Some bootstrap samples will happen to include a lot of large clusters, and others will happen to include a lot of small clusters. In neither of these cases will the bootstrap samples look much like the actual sample. Finally, if the model concerns treatment effects, the number of treated clusters will vary across bootstrap samples. When there are few treated clusters in the actual sample, some of the bootstrap samples may contain no treated clusters at all and will therefore have to be thrown out.

Another approach is to use the score cluster bootstrap of Kline and Santos (2012). This method, which `boottest` implements, involves applying what is essentially the wild cluster bootstrap to the score vectors that appear in expression (27). It has the advantage of being inexpensive to compute, because the model is not re-estimated for each bootstrap sample. However, this computational advantage is also a theoretical disadvantage, because the bootstrap scores do not vary across the bootstrap samples in the same way that actual scores would vary. The distribution of the bootstrap test statistics is essentially based on a linear approximation. Limited simulation evidence in Kline and Santos (2012, Table 3) suggests that the score cluster bootstrap is less reliable than the pairs cluster bootstrap in all cases, especially when $G = 10$ and $G = 20$. However, because these simulations involve

equal cluster sizes, the pairs cluster bootstrap probably performs better than it would have if the cluster sizes varied a lot.

# 10 Conclusions

It has become extremely common in many areas of applied econometrics to divide the data into $G$ somewhat arbitrarily chosen clusters, compute "clustered" standard errors, and rely on the $t(G-1)$ distribution for inference. When the disturbances are correlated within clusters but uncorrelated across them, this can be a reasonable thing to do. Failing to allow for intra-cluster correlation is often much worse. But this approach can also lead to seriously misleading inferences in many cases, even when the appropriate level of clustering is known.

One often overlooked feature of clustered disturbances is that the relationship between the sample size $N$ and the accuracy of parameter estimates does not have its usual form. When there is intra-cluster correlation, 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 of Section 5 illustrates, "big" datasets may actually contain much less information than we might expect them to.

Standard methods for cluster-robust inference based on the $t(G-1)$ distribution generally work acceptably well when the number of clusters is quite large (at least 50) and the clusters are fairly homogeneous in terms of the numbers of observations and the characteristics of the regressors and disturbances. One situation in which inference based on cluster-robust standard errors can be extremely misleading, however, is when interest focuses on a treatment dummy variable and only a few clusters are treated; see Section 7 and MacKinnon and Webb (2017b, 2018). In this case, of course, the key regressor is very heterogeneous across clusters.

As discussed in Section 4, there is a 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 try two or three of them and see whether they agree. In many cases, but not all, the restricted wild cluster bootstrap works very well. Perhaps surprisingly, it can often be computed remarkably quickly using the Stata routine `boottest`; see Roodman et al. (2019).

As the empirical examples of Sections 5 and 7 illustrate, inference can be very sensitive to how (and whether) 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 Paper 24003

Anderson, Theodore W., and Herman Rubin (1949) 'Estimation of the parameters of a single equation in a complete set of stochastic equations.' *Annals of Mathematical Statistics* 20(1), 46–63

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

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock (2006) 'Optimal two-sided invariant similar tests for instrumental variables regression.' *Econometrica* 74(3), 715–752

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(1), 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. 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.' *Review of Economics and Statistics* 90(3), 414–427

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

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.' *Review of Economics and Statistics* 93(1), 113–125

Davidson, Russell, and James G. MacKinnon (1999) 'The size distortion of bootstrap tests.' *Econometric Theory* 15(3), 361–376

Davidson, Russell, and James G. MacKinnon (2010) 'Wild bootstrap tests for IV regression.' *Journal of Business & Economic Statistics* 28(1), 128–144

Davidson, Russell, and James G. MacKinnon (2014a) 'Bootstrap confidence sets with weak instruments.' *Econometric Reviews* 33(6), 651–675

Davidson, Russell, and James G. MacKinnon (2014b) 'Confidence sets based on inverting Anderson-Rubin tests.' *Econometrics Journal* 17(2), S39–S58

Decarolis, Francesco (2014) 'Awarding price, contract performance, and bids screening: Evidence from procurement auctions.' *American Economic Journal: Applied Economics* 6(1), 108–132

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

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

Finlay, Keith, and Leandro M. Magnusson (2019) 'Two applications of wild bootstrap methods to improve inference in cluster-IV models.' *Journal of Applied Econometrics* 34, 911–933

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

Kleibergen, Frank (2002) 'Pivotal statistics for testing structural parameters in instrumental variables regression.' *Econometrica* 70(5), 1781–1803

Kline, Patrick, and Andres Santos (2012) 'A score based approach to wild bootstrap inference.' *Journal of Econometric Methods* 1(1), 23–41

Koedel, Cory, Eric Parsons, Michael Podgursky, and Mark Ehlert (2015) 'Teacher preparation 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(2), 165–199

MacKinnon, James G. (2015) 'Wild cluster bootstrap confidence intervals.' *L'Actualité économique* 91(1-2), 11–33

MacKinnon, James G. (2016) 'Inference with large clustered datasets.' *L'Actualité économique* 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

MacKinnon, James G., and Matthew D. Webb (2019) 'Wild bootstrap randomization inference for few treated clusters.' In *The Econometrics of Complex Survey Data: Theory and Applications,* ed. Kim P. Huynh, David Tomás Jacho-Chávez, and Gautam Tripathi, vol. 39 of *Advances in Econometrics* (Emerald Group) chapter 3, pp. 61–85

MacKinnon, James G., and Matthew D. Webb 'Randomization inference for difference-in-differences with few treated clusters.' *Journal of Econometrics* 218(2), 435–450

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

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2020b) 'Wild bootstrap and asymptotic inference with multiway clustering.' *Journal of Business & Economic Statistics* 38, to appear

Moreira, Marcelo J. (2003) 'A conditional likelihood ratio test for structural models.' *Econometrica* 71(4), 1027–1048

Nelson, Charles R., and Richard Startz (1990a) 'Some further results on the exact small sample properties of the instrumental variables estimator.' *Econometrica* 58(4), 967–976

Nelson, Charles R., and Richard Startz (1990b) 'The distribution of the instrumental variables estimator and its $t$-ratio when the instrument is a poor one.' *Journal of Business* 63(1), S125–S140

Pustejovsky, James E., and Elizabeth Tipton (2018) 'Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models.' *Journal of Business and Economic Statistics* 36, 672–683

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 (2019) 'Fast and wild: Bootstrap inference in Stata using boottest.' *Stata Journal* 19(1), 4–60

Staiger, Douglas, and James H. Stock (1997) 'Instrumental variables regression with weak instruments.' *Econometrica* 65(3), 557–586

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.' QED Working Paper 1315, Queen's University, Department of Economics

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