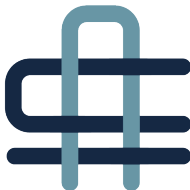


# When Can We Trust Cluster-Robust Inference?

James G. MacKinnon

Queen's University &  
Aarhus Centre for Econometrics (ACE)

Canadian Stata Conference, Oct. 3, 2025



# Introduction

Cluster-robust standard errors are widely used, but  $t$ -statistics and confidence intervals based on them can be unreliable.

For detailed discussions of all but the latest methods, see [MacKinnon, Nielsen, and Webb \(JoE 2023a, “The Guide”\)](#).

- Most asymptotic theories only tell us that inference is reliable when  $G$ , the number of clusters, is very large.
- We cannot know when  $G$  is large enough, because several other features of the model/sample can matter greatly.
- The most commonly used standard errors are the worst choice. Better ones are readily available. We can also use bootstrap methods or coefficient-specific critical values.
- The best methods can yield surprisingly reliable inferences even when  $G$  is quite small.

How can we decide whether various forms of cluster-robust inference can be relied upon in any given case?

# Linear Regression with Clustering

There are  $G$  clusters, indexed by  $g$ . The  $g^{\text{th}}$  cluster has  $N_g$  observations, so the sample size is  $N = \sum_{g=1}^G N_g$ . The model can be written as

$$\mathbf{y}_g = \mathbf{X}_g \boldsymbol{\beta} + \mathbf{u}_g, \quad g = 1, \dots, G, \quad (1)$$

where  $\mathbf{X}_g$  is an  $N_g \times k$  matrix of regressors,  $\boldsymbol{\beta}$  is a  $k$ -vector of coefficients, and  $\mathbf{y}_g$  and  $\mathbf{u}_g$  are  $N_g$ -vectors.

Stacking the  $\mathbf{X}_g$  yields the  $N \times k$  matrix  $\mathbf{X}$ , and stacking the  $\mathbf{y}_g$  and  $\mathbf{u}_g$  yields the  $N$ -vectors  $\mathbf{y}$  and  $\mathbf{u}$ .

The OLS estimator of  $\boldsymbol{\beta}$  is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}. \quad (2)$$

This assumes that the data are actually generated by (1) with  $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ .

If  $\mathbf{s}_g = \mathbf{X}_g^\top \mathbf{u}_g$  is the **score vector** for the  $g^{\text{th}}$  cluster, and (1) is correct,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\mathbf{X}^\top \mathbf{X})^{-1} \sum_{g=1}^G \mathbf{X}_g^\top \mathbf{u}_g = \left( \sum_{g=1}^G \mathbf{X}_g^\top \mathbf{X}_g \right)^{-1} \sum_{g=1}^G \mathbf{s}_g. \quad (3)$$

The random variation in  $\hat{\boldsymbol{\beta}}$  around  $\boldsymbol{\beta}_0$  evidently arises from the randomness in the  $\mathbf{s}_g$ .

### Key assumptions:

$$E(\mathbf{s}_g \mathbf{s}_g^\top) = \boldsymbol{\Sigma}_g \quad \text{and} \quad E(\mathbf{s}_g \mathbf{s}_{g'}^\top) = \mathbf{0}, \quad g, g' = 1, \dots, G, \quad g' \neq g, \quad (4)$$

where  $\boldsymbol{\Sigma}_g$  is the symmetric, positive semidefinite variance matrix of the scores for the  $g^{\text{th}}$  cluster. There are two assumptions here:

- 1 Within each cluster, there may be very general patterns of heteroskedasticity and/or intra-cluster correlation.
- 2 The scores for every cluster are uncorrelated with the scores for every other cluster. **This assumption is crucial!**

The true variance matrix of  $\hat{\beta}$  is the sandwich matrix

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \left( \sum_{g=1}^G \boldsymbol{\Sigma}_g \right) (\mathbf{X}^\top \mathbf{X})^{-1}. \quad (5)$$

If we knew  $\text{Var}(\hat{\beta})$ , and assuming that a central limit theory can be applied, inference could be based on the fact that

$$\hat{\beta} \stackrel{a}{\sim} \text{N}(\beta_0, \text{Var}(\hat{\beta})); \quad (6)$$

see [Djogbenou, MacKinnon, and Nielsen \(2019\)](#). This should work well, except perhaps when  $G$  and/or  $N$  are very small, or the score vectors are very heterogeneous. But  $\text{Var}(\hat{\beta})$  is unknown!

To estimate  $\text{Var}(\hat{\beta})$ , we have to estimate the  $\boldsymbol{\Sigma}_g$  consistently, and there is more than one way to do so.

Unfortunately, replacing the  $\boldsymbol{\Sigma}_g$  in (6) by consistent estimates  $\hat{\boldsymbol{\Sigma}}_g$  may lead to seriously unreliable inferences.

## Three Cluster-Robust Variance Estimators

There are several ways to estimate the middle factor in (5). Each yields a different **cluster-robust variance estimator**, or **CRVE**.

The simplest way is to replace  $\Sigma_g$  by  $\hat{s}_g \hat{s}_g^\top$ , where  $\hat{s}_g = \mathbf{X}_g^\top \hat{u}_g$  is the empirical score vector for the  $g^{\text{th}}$  cluster. Multiplying by a d-o-f correction, we obtain

$$\text{CV}_1: \quad \hat{V}_1(\hat{\beta}) = \frac{G(N-1)}{(G-1)(N-k)} (\mathbf{X}^\top \mathbf{X})^{-1} \left( \sum_{g=1}^G \hat{s}_g \hat{s}_g^\top \right) (\mathbf{X}^\top \mathbf{X})^{-1}. \quad (7)$$

The leading scalar is chosen so that, when  $G = N$ ,  $\hat{V}_1(\hat{\beta})$  reduces to the familiar  $\text{HC}_1$  estimator of **Mackinnon and White (1985)**.

The  $\hat{s}_g$  do not always provide good estimates of the  $s_g$ , because the  $\hat{u}_g$  do not always provide good estimates of the  $u_g$ .

In general,  $\hat{u} = M_X u$ , where  $M_X = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  projects  $u$  off  $\mathbf{X}$ . This projection means that  $\hat{u}$  and  $u$  can have very different properties.

Depending on the  $X$  matrix and the  $u$  vector, the  $\hat{s}_g$  can sometimes differ greatly from the  $s_g$ , causing the middle factor in (7) to provide a poor estimate of  $\sum_{g=1}^G \Sigma_g$ . Thus  $CV_1$  can perform very badly.

There are CRVE analogs of the two other popular variance estimators for heteroskedasticity discussed in [M & W \(1985\)](#).

- For  $HC_2$ , the  $\hat{u}_i$  are replaced by  $\hat{u}_i/M_{ii}^{1/2}$ , where  $M_{ii}$  is the  $i^{\text{th}}$  diagonal element of  $M_X$ .  $CV_2$  is analogous; it involves the inverse symmetric square roots of the  $M_{gg}$  (diagonal blocks).
- For  $HC_3$ , the  $\hat{u}_i$  are replaced by  $\hat{u}_i/M_{ii}$ .  $CV_3$  is analogous; it involves the inverses of the  $M_{gg}$ .
- $CV_2$  and  $CV_3$  were first proposed in [Bell and McCaffrey \(2002\)](#), using computational procedures like those for  $HC_2$  and  $HC_3$ .
- Better computational procedures (unless all the  $N_g$  are very small) are discussed in [MacKinnon, Nielsen, and Webb \(JAE 2023b\)](#).
- $CV_3$  is really a **cluster-jackknife** estimator.

The OLS estimates of  $\beta$  when cluster  $g$  is omitted are

$$\hat{\beta}^{(g)} = (\mathbf{X}^\top \mathbf{X} - \mathbf{X}_g^\top \mathbf{X}_g)^{-1} (\mathbf{X}^\top \mathbf{y} - \mathbf{X}_g^\top \mathbf{y}_g), \quad g = 1, \dots, G. \quad (8)$$

To obtain the  $\hat{\beta}^{(g)}$  efficiently, start by calculating

$$\mathbf{X}_g^\top \mathbf{X}_g \quad \text{and} \quad \mathbf{X}_g^\top \mathbf{y}_g, \quad g = 1, \dots, G. \quad (9)$$

Unless  $G$  is very large, this involves very little net cost, because the quantities in (9) can be used to construct  $\mathbf{X}^\top \mathbf{X}$  and  $\mathbf{X}^\top \mathbf{y}$ .

The main cost, after  $\hat{\beta}$  and its ingredients have been computed, is calculating the (generalized) inverse of a  $k \times k$  matrix for each  $\hat{\beta}^{(g)}$ .

The simplest version of the cluster-jackknife variance matrix is

$$\text{CV}_3: \quad \hat{V}_3(\hat{\beta}) = \frac{G-1}{G} \sum_{g=1}^G (\hat{\beta}^{(g)} - \hat{\beta})(\hat{\beta}^{(g)} - \hat{\beta})^\top. \quad (10)$$

This matrix is numerically identical to the original  $\text{CV}_3$  matrix. However, it is usually very much cheaper to compute.

Theoretical results in Hansen (2025a) and simulation evidence in MacKinnon, Nielsen, and Webb (JAE 2023b, SJ 2023c) suggest that  $CV_3$  standard errors are always larger than  $CV_1$  ones and that  $t$ -statistics based on the former almost always yield more reliable inferences.

What about  $CV_2$  standard errors?

- With neither heteroskedasticity nor intra-cluster correlation, and some conditions on  $X$ , the diagonal elements of  $CV_2$  are unbiased, like the diagonal elements of  $HC_2$ .
- In contrast, the diagonal elements of  $CV_3$  and  $HC_3$  are generally biased upwards in the special case of i.i.d. disturbances.
- However, because the numerators of cluster-robust  $t$ -statistics are not independent of the denominators, using the square root of an unbiased variance estimator in the denominator does *not* guarantee that inference will be reliable.
- Simulations suggest that  $CV_3$  usually outperforms  $CV_2$ , so I won't say any more about the latter.

# Inference Using CRVEs

Suppose we wish to test the hypothesis that  $\beta_j = \beta_{0j}$  or construct a confidence interval for  $\beta_j$ . We start with the  $t$ -statistic

$$t_j^m = \frac{\hat{\beta}_j - \beta_{0j}}{\text{se}_m(\hat{\beta}_j)}, \quad m = 1, 2, 3, \quad (11)$$

where  $\text{se}_m(\hat{\beta}_j)$  is the square root of the  $j^{\text{th}}$  diagonal element of  $CV_m$ .

By combining (11) with an assumed distribution, we can perform hypothesis tests or construct confidence intervals.

It might seem natural to employ the standard normal distribution, but this a really bad idea.

The asymptotic theory of [Bester, Conley, and Hansen \(2011\)](#) holds  $G$  fixed, with the  $N_g$  increasing and intra-cluster correlation diminishing. For  $CV_1$ , (11) is then asymptotically distributed as  $t(G - 1)$ .

Even when the **BCH** assumptions do not hold,  $t(G - 1)$  seems to work better than  $N(0, 1)$ , but not perfectly.

Nevertheless, inferences based on any CRVE plus  $t(G - 1)$  have two unsatisfactory features:

- They use the  $t(G - 1)$  distribution, which does not depend on  $X$ .
- They do not attempt to correct any of the CRVEs for bias.

Several authors suggest using a calculated degrees-of-freedom parameter, say  $d_i$ , in place of  $G - 1$ . See **Bell and McCaffrey (2002)**, **Imbens and Kolesár (2016)**, **Young (2016)**, and **Hansen (2025a, b)**.

Others suggest rescaling the standard errors based on the bias of the estimated variance, which is different for each coefficient. See **Young (2016)**, **Boot, Niccodemi, and Wansbeek (2023)**, and **Hansen (2025a, b)**.

These two approaches are combined in Hansen's two papers. His Stata package `jregress` rescales each  $CV_3$  standard error differently and computes  $d_i$ , often much less than  $G - 1$ , for each coefficient.

# Bootstrap Inference

Generate a large number, say  $B$ , of bootstrap samples, indexed by  $b$ . Compute  $\hat{\beta}^{*b}$  and (usually) the  $t_j^{*b}$  for each of them.

Conceptually the simplest bootstrap DGP is the **pairs cluster bootstrap** (PCB), or **resampling bootstrap**, which resamples from the pairs

$$[\mathbf{X}_g^\top \mathbf{X}_g, \mathbf{X}_g^\top \mathbf{y}_g], \quad g = 1, \dots, G. \quad (12)$$

The bootstrap standard error  $\text{se}^*(\hat{\beta}_j)$  is the square root of  $\widehat{\text{Var}}(\hat{\beta}_j^{*b})$ . We can compare it with the  $\text{se}_m(\beta_j)$ , for  $m = 1, 2, 3$ , or use it for inference.

For any choice of  $\text{se}(\hat{\beta}_j^{*b})$ , we can compute the bootstrap  $t$ -statistic

$$t_j^{*b} = \frac{\hat{\beta}_j^{*b} - \hat{\beta}_j}{\text{se}(\hat{\beta}_j^{*b})}, \quad b = 1, \dots, B. \quad (13)$$

This is testing the hypothesis that  $\beta_j = \hat{\beta}_j$ , not that  $\beta_j = \beta_{0j}$ .

We can then compute the **symmetric bootstrap  $P$  value**

$$\hat{P}^* = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(|t_j^{*b}| > |t_j|). \quad (14)$$

When  $\hat{P}^* < \alpha$ , we can reject the null hypothesis at level  $\alpha$ .

To compute **studentized bootstrap confidence intervals**, sort the  $t_j^{*b}$  from smallest to largest and find  $c_{1-\alpha/2}^*$  and  $c_{\alpha/2}^*$ .

The studentized bootstrap confidence interval at level  $1 - \alpha$  is then

$$[\hat{\beta}_j - \text{se}(\hat{\beta}_j)c_{1-\alpha/2}^*, \quad \hat{\beta}_j - \text{se}(\hat{\beta}_j)c_{\alpha/2}^*], \quad (15)$$

where the cluster-robust standard error  $\text{se}(\hat{\beta}_j)$  is the same function of the actual data as  $\text{se}(\hat{\beta}_j^{*b})$  is a function of the bootstrap data.

`vce(bootstrap)` with `cluster(cvar)` calculates PCB standard errors, but not  $P$  values like (14) or intervals like (15). It is very expensive, and it can be unreliable if many bootstrap samples are omitted.

# The Wild Cluster Bootstrap

The **wild cluster bootstrap** (WCB) often works much better than the PCB. It was proposed in [Cameron, Gelbach, and Miller \(2008\)](#), proved to be valid in [Djogbenou, MacKinnon, and Nielsen \(2019\)](#), and improved in [MacKinnon, Nielsen, and Webb \(JAE 2023b\)](#).

Consider the unrestricted empirical score vectors

$$\hat{\mathbf{s}}_g = \mathbf{X}_g^\top \hat{\mathbf{u}}_g = \mathbf{X}_g^\top \mathbf{y}_g - \mathbf{X}_g^\top \mathbf{X}_g \hat{\boldsymbol{\beta}}, \quad g = 1, \dots, G. \quad (16)$$

To generate the  $b^{\text{th}}$  bootstrap sample, we multiply each of these vectors by a scalar **auxiliary random variable**  $v_g^{*b}$  with mean 0 and variance 1.

This should usually be the Rademacher distribution. When  $G$  is very small (say  $G \leq 12$ ), it is better to use Webb's six-point distribution.

We then obtain bootstrap estimates of the vector  $\boldsymbol{\delta} \equiv \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$ :

$$\hat{\boldsymbol{\delta}}^{*b} = (\mathbf{X}^\top \mathbf{X})^{-1} \sum_{g=1}^G \mathbf{s}_g^{*b}, \quad \mathbf{s}_g^{*b} = v_g^{*b} \hat{\mathbf{s}}_g. \quad (17)$$

Next, we compute the bootstrap  $t$ -statistic

$$t_j^{*b} = \frac{\hat{\delta}_j^{*b}}{\text{se}(\hat{\delta}_j^{*b})}, \quad (18)$$

where  $\text{se}(\hat{\delta}_j^{*b})$  is the square root of the  $j^{\text{th}}$  diagonal element of the  $\text{CV}_1$  matrix (7), with the vectors  $\hat{\mathbf{s}}_g$  replaced by the vectors

$$\hat{\mathbf{s}}_g^{*b} = \mathbf{s}_g^{*b} - \mathbf{X}_g^\top \mathbf{X}_g \hat{\boldsymbol{\delta}}^{*b}. \quad (19)$$

The  $t_j^{*b}$  can then be used to compute bootstrap  $P$  values, using (14), and studentized bootstrap confidence intervals, using (15).

This variant of the WCB is the **WCU-C bootstrap**. The “U” indicates that the bootstrap scores are based on the unrestricted empirical score vectors (16), and “-C” means “classic.”

Imposing restrictions on the bootstrap samples makes bootstrapping more complicated, but it often improves performance.

Suppose we obtain OLS estimates  $\tilde{\beta}$  and residual vectors  $\tilde{u}_g$  subject to the restriction that  $\beta_j = \beta_{j0}$ . Then the analog of (16) is

$$\tilde{s}_g = \mathbf{X}_g^\top \tilde{u}_g = \mathbf{X}_g^\top \mathbf{y}_g - \mathbf{X}_g^\top \mathbf{X}_g \tilde{\beta}, \quad g = 1, \dots, G. \quad (20)$$

Now when we “regress” the  $\tilde{s}_g$  on  $\mathbf{X}_g$ ,  $\hat{\delta}^{*b}$  is an estimate of  $\beta - \hat{\beta}$ .

This variant of the WCB is called the **WCR-C** variant, where “R” stands for “restricted.” It is probably the most widely-used bootstrap procedure for linear models with clustering.

We can still compute bootstrap  $P$  values using (14). However, we cannot construct studentized bootstrap confidence intervals using (15).

Instead, we have to “invert” the bootstrap test statistic, finding two values of  $\beta_j$ , one on each side of  $\hat{\beta}_j$ , such that the equal-tail  $P$  values for tests that  $\beta_j$  equals each of these values are approximately  $\alpha$ .

Happily, `boottest` does this incredibly quickly. See [Roodman, MacKinnon, Nielsen, and Webb \(2019\)](#) and [MacKinnon \(2023\)](#).

MNW (JAE 2023b) proposes six new variants of the WCB. Two of these are known as **WCU-S** and **WCR-S**, where “S” stands for “score.”

The idea of these new wild cluster bootstraps is to replace the empirical score vectors  $\hat{s}_g$  or  $\tilde{s}_g$  in the bootstrap DGP by modified score vectors that (partly) correct for the distortions caused by least squares.

Recall that  $\hat{u} = M_X u$ . We can partly undo the ill effects of this by using

$$\hat{s}_g = X_g^\top M_{gg}^{-1} \hat{u}_g, \quad (21)$$

where  $M_{gg}^{-1}$  is the inverse of the  $g^{\text{th}}$  diagonal block of  $M_X$ . **But this can be insanely expensive!**

It is shown in MNW (JAE 2023b) that

$$\hat{s}_g = X^\top X (\hat{\beta} - \hat{\beta}^{(g)}), \quad g = 1, \dots, G. \quad (22)$$

Thus computing the  $\hat{s}_g$  is almost costless once the jackknife estimates, which are needed for  $CV_3$ , have been computed.

The modified score vectors  $\hat{s}_g$  are used in the bootstrap DGP for the WCU-S bootstrap. In all other respects, the WCU-S and WCU-C bootstraps are computed in exactly the same way.

A similar procedure can be used to compute restricted score vectors  $\hat{s}$  that “correct” for the distortions caused by estimating the restricted model. The  $\hat{s}$  are used in the bootstrap DGP for the WCR-S bootstrap, which otherwise is computed just like the WCR-C bootstrap.

- It may seem odd to use the  $CV_1$  standard error in  $t_j$  and the  $t_j^{*b}$ , when the transformation (21) is based on the cluster jackknife.
- Another variant uses the  $CV_3$  standard error, but it is a lot more expensive to compute, and it does not consistently perform better.

WCR[U/R]-[C/S] are implemented in current versions of `boottest`. It computes both bootstrap confidence intervals and bootstrap  $P$  values, including for tests of several linear restrictions. In most cases, this is inexpensive, even when  $B$  is chosen to be a large number like 99,999.

## Why Cluster-Robust Inference May Be Unreliable

Consider a hypothesis test at nominal level  $\alpha$  or a confidence interval with nominal coverage  $1 - \alpha$ . When is such a procedure “reliable”?

A “reliable” test should yield a rejection frequency in  $[\alpha_l, \alpha_u]$ , or coverage in  $[1 - \alpha_u, 1 - \alpha_l]$ , for values of  $\alpha_l < \alpha < \alpha_u$  that an investigator is comfortable with.

What are reasonable values of  $\alpha_l$  and  $\alpha_u$ ? Perhaps  $0.9\alpha$  and  $1.1\alpha$  (or  $0.045$  and  $0.055$  when  $\alpha = 0.05$ ).

If an investigator chooses  $\alpha_l = 0.999\alpha$  and  $\alpha_u = 1.001\alpha$ , then there probably exists no method for cluster-robust inference that is reliable!

The number of clusters  $G$  is important, because the expressions for  $CV_1$ ,  $CV_2$ , and  $CV_3$  involve summations over  $G$  terms.

We need  $\frac{1}{G} \sum_{g=1}^G \hat{\mathbf{s}}_g \hat{\mathbf{s}}_g^\top$  in (7), and its analog for the other CRVEs, to converge to the same matrix as  $\frac{1}{G} \sum_{g=1}^G \mathbf{s}_g \mathbf{s}_g^\top$ .

The sample size  $N$  also matters, but not much once  $N/G$  is moderately large. We cannot simply hold  $G$  fixed and let  $N \rightarrow \infty$ .

Any sort of heterogeneity in the  $\mathbf{X}_g^\top \mathbf{X}_g$  and the  $\mathbf{X}_g^\top \mathbf{y}_g$  matters.

Key sources of heterogeneity are

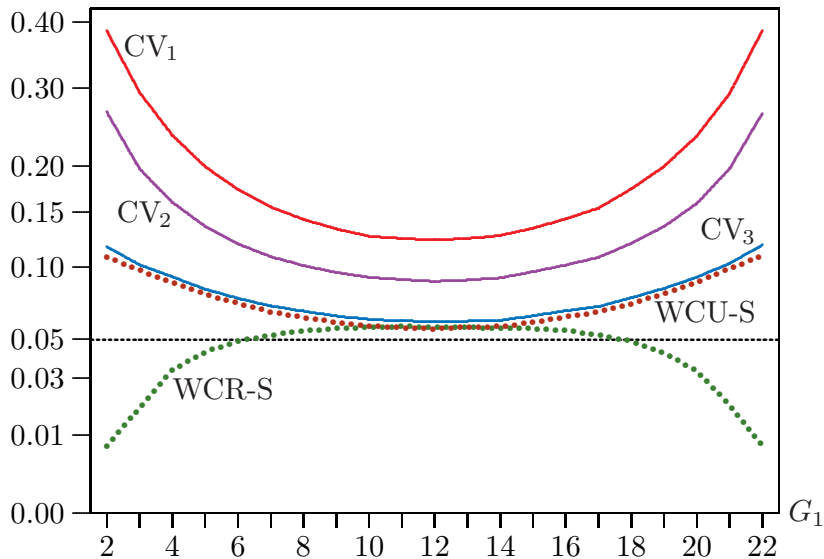
- variation in the  $N_g$  across clusters;
- variation in the distributions of the  $\mathbf{X}_g$  across clusters;
- heteroskedasticity within and across clusters;
- variation in the patterns of within-cluster correlation.

When there are treatment dummies, including DiD models, numbers of treated clusters ( $G_1$ ) and control clusters ( $G_0$ ) matter greatly.

- $G_1$  should not be too small, nor  $G_0$  in pure treatment case.
- $G_1/G$  should not be too close to 0 or 1.

Otherwise, cluster-robust standard errors are too small, pairs cluster and WCU bootstraps over-reject, and WCR bootstraps under-reject.

See [MacKinnon and Webb \(JAE 2017, EJ 2018\)](#).

Figure 1. Rejection frequencies as functions of  $G_1$  for  $G = 24$ 

## How Should We Cluster?

Before we begin to compute standard errors, we need to decide just how the scores are clustered.

- At what level should we cluster?
- Should we cluster in one, two, or more than two dimensions?
- If either the disturbances or the (partialled-out) regressor of interest is not clustered, then the scores are not clustered.

### Testing the Level of Clustering

Suppose there is more than one level at which we could cluster.

- Perhaps there is a fine level (say, schools) and a coarser level (say school districts).
- When there are many fine clusters, inference is likely to be reliable if fine clustering is appropriate.
- But it will be invalid if coarse clustering is appropriate.

Several tests for the level of clustering have been proposed.

MacKinnon, Nielsen, and Webb (JoE, 2023d) proposes both asymptotic and wild bootstrap tests based on elements of the score vectors after regressors that are not of primary interest have been partialled out. These are called **score-variance tests**.

- Score-variance tests depend on which coefficient(s) are of interest. It is easiest to deal with a single coefficient.
- Asymptotic score-variance tests may not perform well. Bootstrap tests (based on the WCU-C bootstrap) seem to perform better.
- There is a package called `mnwsvt`, but it is not on SSC and probably needs some work.
- Even without testing, we should cluster at the coarse level if  $se_c$ , the coarse standard error, is noticeably larger than  $se_f$ , the fine one.
- However, we should probably not cluster at the coarse level if  $se_c$  is smaller, even “significantly” smaller, than  $se_f$ .

## Two-Way Clustering

There are circumstances in which there may be two, or even more than two, clustering dimensions. Obvious ones are space and time.

In the two-way case, every observation is assumed to belong to one cluster in each of the two dimensions.

If an observation belongs to cluster  $g$  in the spatial dimension and cluster  $t$  in the time dimension, then it may be correlated with other observations that belong either to cluster  $g$  or to cluster  $t$ .

The idea of two-way clustering was independently discovered by [Miglioretti and Heagerty \(2006\)](#), [Cameron, Gelbach, and Miller \(2011\)](#), and [Thompson \(2011\)](#).

Recent work includes [MacKinnon, Nielsen, and Webb \(2021, 2024\)](#), [Chiang, Hansen, and Sasaki \(2024\)](#), and [Davezies, d'Haultefoeuille, and Gyonvarch \(2025\)](#).

Anything that causes inference to be unreliable for one-way clustering also causes it to be unreliable for two-way clustering.

For two-way clustering, the filling in the sandwich for the true variance matrix is

$$\Sigma = \sum_{g=1}^G \Sigma_g + \sum_{h=1}^H \Sigma_h - \sum_{g=1}^G \sum_{h=1}^H \Sigma_{gh}. \quad (23)$$

Here the  $\Sigma_g$  and  $\Sigma_h$  are the variance matrices of the score vectors for each of the two clustering dimensions, and the  $\Sigma_{gh}$  are the variance matrices for the intersections of the two dimensions.

The third term in (23) has to be subtracted to avoid double counting. Although (23) is positive definite, its empirical analogs are not. In consequence, standard errors may be undefined or extremely small. If two-way clustering yields smaller standard errors than one-way clustering in either dimension, then the former should not be believed.

MacKinnon, Nielsen, and Webb (2024) and Davezies et al. (2025) suggested computing both one-way standard errors, along with two-way ones (if defined) and using whichever is largest.

# Measures of Cluster Heterogeneity

The smaller the number of clusters  $G$ , the less reliable can we expect all types of cluster-robust inference to be.

- There is no simple rule about how large  $G$  needs to be. It depends in complicated ways on many features of the model and data.
- In general, for a given  $G$ , the more the data vary across clusters, the less reliable inference tends to be.

One important source of variation is cluster sizes. When the  $N_g$  vary greatly, we cannot expect any method to work really well.

Problems occur when there are a few excessively large clusters, not a few excessively small ones. Imagine the following two samples:

**Sample 1:** 19 clusters each with  $N_g = 500$ , 1 cluster with  $N_{20} = 10$ . This is almost like a sample with 19 equal-sized clusters; use  $t(18)$ .

**Sample 2:** 19 clusters each with  $N_g = 500$ , 1 cluster with  $N_{20} = 10,000$ . Thus  $N_{20} > \sum_{g=1}^{19} N_g$ . Inference is likely to be dreadful!

Chiang, Sasaki, and Wang (2025) suggests using weighted least squares, with weights  $N_g^{-1/2}$ , instead of OLS.

- This might well be a good idea for Sample 2, but it would be a terrible idea for Sample 1, where OLS should usually work well.
- More sophisticated forms of WLS might be worth investigating.

Inference becomes less reliable as the **leverage** and **partial leverage** of the clusters vary more; see MacKinnon, Nielsen, and Webb (SJ, 2023c).

Suppose that we are interested in  $x_j$ . Partialing it out, we obtain  $\hat{x}_j$ . The measure of partial leverage for regressor  $j$  for the  $g^{\text{th}}$  cluster is then

$$L_{gj} = \frac{\hat{x}_{gj}^\top \hat{x}_{gj}}{\hat{x}_j^\top \hat{x}_j}, \quad (24)$$

where  $\hat{x}_{gj}$  is the subvector of  $\hat{x}_j$  corresponding to the  $g^{\text{th}}$  cluster.

Since the  $L_{gj}$  sum to unity, their average is  $1/G$ . Thus, if cluster  $h$  has  $L_{hj} \gg 1/G$ , it has high partial leverage for the  $j^{\text{th}}$  coefficient.

When  $G$  is small, look at all the  $L_{gj}$  for every coefficient of interest.

When  $G$  is not small, graph them or report summary measures of how much they vary across clusters. One such measure is the scaled variance

$$V_s^j = \frac{G^2}{(G-1)} \sum_{g=1}^G (L_{gj} - 1/G)^2. \quad (25)$$

The square root of  $V_s^j$  is the **coefficient of variation** of the  $L_{gj}$ .

It will be 0 whenever every cluster has the same partial leverage and large whenever  $\text{Var}(L_{gj})$  is large relative to  $1/G^2$ .

**Carter, Schnepel, and Steigerwald (2017)** proposes a family of measures  $G_j^*(\rho)$  called the “effective number of clusters.”

- These measures depend on a parameter  $\rho$ , the intra-cluster correlation of the disturbances in a random-effects model.
- When there are cluster fixed effects, only  $\rho = 0$  makes sense.

MacKinnon, Nielsen, and Webb (SJ, 2023c) shows that  $G_j^*(0)$  is simply a monotonically decreasing function of the scaled variance (25).

Thus, when  $V_s^j$  is large,  $G_j^*(0)$  is necessarily much smaller than  $G$ .

$V_s^j$  and  $G_j^*(0)$  convey exactly the same information, but the latter is easier to interpret, because it is bounded above by  $G$ .

- If  $G = 50$  and  $G^*(0) = 48.7$ , we would expect many methods to yield fairly reliable inferences.
- But if  $G_j^*(0) = 11.7$ , we would expect cluster-robust inference to be challenging.

The package `summclust` calculates both leverage and partial leverage at the cluster level, as well as measures of how much they vary, the effective number of clusters, and several other useful diagnostics.

It also calculates both  $CV_1$  and  $CV_3$  variance matrices, along with the associated  $P$  values and confidence intervals.

# Heteroskedasticity

Both the extent of heteroskedasticity across clusters, if any, and the patterns of intra-cluster correlation can be important.

The average variance of the residuals for each cluster is

$$\hat{\sigma}_g^2 = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (\hat{u}_{gi} - \bar{u}_g)^2, \quad g = 1, \dots, G, \quad (26)$$

where  $\bar{u}_g$  is the average residual for cluster  $g$ .

If there are cluster fixed effects, we can compute the  $\hat{\sigma}_g^2$  by regressing the squared residuals on a full set of cluster dummies. When they vary substantially across clusters, inference may be unreliable.

Simulations in [Hansen \(2025\)](#) and [Chiang, Sasaki, and Wang \(2025\)](#) make  $\text{Var}(u_{gi})$  much larger for treated than control observations.

This typically harms performance of every method, especially all variants of the wild cluster bootstrap.

To check whether the data display this sort of heteroskedasticity, run the regression

$$\hat{u}_{gi}^2 = \eta_1 + \eta_2 T_{gi} + e_{gi}, \quad g = 1, \dots, G, \quad i = 1, \dots, N_g. \quad (27)$$

where  $T_{gi}$  is the treatment dummy for observation  $gi$ .

In (27),  $\eta_1$  is the average variance for control observations, and  $\eta_1 + \eta_2$  is the average variance for treated observations.

A cluster-robust test for  $\eta_2 = 0$  tests whether the disturbances for control and treated clusters have the same average variance.

If  $\hat{\eta}_1 + \hat{\eta}_2$  is much larger than  $\hat{\eta}_1$ , then results from inferential methods that might otherwise perform well should be interpreted with caution.

A diagnostic that depends on both  $\mathbf{y}$  and  $\mathbf{X}$  is the variability in the omit-one-cluster estimates  $\hat{\beta}^{(g)}$ ; `summlust` optionally reports these.

When there are one or two clusters where  $\hat{\beta}^{(g)}$  differs greatly from  $\hat{\beta}$ , an investigator should be cautious.

# Targeted Monte Carlo Experiments

A targeted Monte Carlo experiment can provide estimates of how accurate  $P$  values and confidence intervals are likely to be.

- Use the actual matrix  $X$  and the actual clusters, along with  $\beta = \hat{\beta}$  or  $\beta = \mathbf{0}$ . Inexpensive since  $X^\top X$  and  $X_g^\top X_g$  are fixed.
- The difficulty is deciding precisely how to generate the  $u_g$ .
- The easiest approach is to use the random-effects model

$$u_{gi} = v_g + \epsilon_{gi}, \quad v_g \sim N(0, \sigma_v^2), \quad \epsilon_{gi} \sim N(0, \sigma_\epsilon^2), \quad (28)$$

which implies that the disturbances within each cluster are homoskedastic and equi-correlated with  $\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_\epsilon^2)$ .

- But if the regressors include cluster fixed effects, they completely explain the  $v_g$ , so there is no within-cluster correlation.
- There are many ways to generate the  $u_{gi}$  for models with cluster fixed effects, and it is not clear how much it matters.

# Placebo Regressions

These are like Monte Carlo experiments, but instead of varying  $y$ , we vary one column of  $X$  across replications.

For each replication, we add a random extra regressor which looks like, but does not replace, the regressor of interest.

$$y = X\beta + z\gamma + u \quad (29)$$

Find the fraction of replications for which a test of  $\gamma = 0$  rejects the null, or a confidence interval includes 0.

If we replaced the regressor of interest, we would be assuming the model is correctly specified without it.

How we generate the  $z$  matters. For DiD models, they should look like **placebo laws**, as in **Bertrand, Duflo, and Mullainathan (2004)**.

Can we model heteroskedasticity related to treatment?

# Logistic Regression Models

If the binary variable  $y_{gi}$  is the response for observation  $i$  in cluster  $g$ ,

$$\Pr(y_{gi} = 1 \mid \mathbf{X}_{gi}) = \Lambda(\mathbf{X}_{gi}\boldsymbol{\beta}), \quad g = 1, \dots, G, \quad i = 1, \dots, N_g. \quad (30)$$

Here  $\mathbf{X}_{gi}$  contains  $k$  explanatory variables, with  $\boldsymbol{\beta}$  to be estimated.

In (30),  $\Lambda(\cdot)$  is the logistic function,

$$\Lambda(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}. \quad (31)$$

The pseudo-loglikelihood function for (30) is

$$\ell(\mathbf{y}, \boldsymbol{\beta}) = \sum_{g=1}^G \sum_{i=1}^{N_g} (y_{gi} \log \Lambda(\mathbf{X}_{gi}\boldsymbol{\beta}) + (1 - y_{gi}) \log \Lambda(-\mathbf{X}_{gi}\boldsymbol{\beta})). \quad (32)$$

There are other ways to write this.

Using the fact that the first derivative of  $\Lambda(x)$  is  $\Lambda(x)\Lambda(-x)$ , the score vector for the  $g^{\text{th}}$  cluster is simply

$$\mathbf{s}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \mathbf{s}_{gi}(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} (y_{gi} - \Lambda(\mathbf{X}_{gi}\boldsymbol{\beta}))\mathbf{X}_{gi}. \quad (33)$$

Thus, the first-order condition for  $\hat{\boldsymbol{\beta}}$  can be written as

$$\hat{\mathbf{s}} = \sum_{g=1}^G \hat{\mathbf{s}}_g = \sum_{g=1}^G \mathbf{s}_g(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \quad (34)$$

When the observations are independent, we obtain the variance matrix estimator

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}, \quad (35)$$

where  $\mathbf{Y}(\boldsymbol{\beta})$  is an  $N \times N$  diagonal matrix with typical diagonal element

$$Y_i(\boldsymbol{\beta}) = \Lambda(\mathbf{X}_i\boldsymbol{\beta})\Lambda(-\mathbf{X}_i\boldsymbol{\beta}). \quad (36)$$

For the logit model,  $\mathbf{X}^\top \mathbf{Y}(\boldsymbol{\beta})\mathbf{X} = -H(\boldsymbol{\beta})$ .

The usual CRVE is

$$\text{CV}_{1\mathcal{I}}: \hat{\mathbf{V}}_{1\mathcal{I}} = \frac{G}{G-1} \frac{N-1}{N-k} (\mathbf{X}^\top \hat{\mathbf{Y}}\mathbf{X})^{-1} \left( \sum_{g=1}^G \hat{\mathbf{s}}_g \hat{\mathbf{s}}_g^\top \right) (\mathbf{X}^\top \hat{\mathbf{Y}}\mathbf{X})^{-1}. \quad (37)$$

The empirical score vectors here are

$$\mathbf{s}_g(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N_g} (y_{gi} - \Lambda(\mathbf{X}_{gi}\hat{\boldsymbol{\beta}})) \mathbf{X}_{gi}, \quad g = 1, \dots, G. \quad (38)$$

If  $\hat{\boldsymbol{\beta}}^{(g)}$  is the vector of delete-one estimates when cluster  $g$  is deleted, we obtain the cluster-jackknife CRVE

$$\text{CV}_3: \hat{\mathbf{V}}_3(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^G (\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^\top. \quad (39)$$

Computing  $\text{CV}_3$  requires  $G + 1$  nonlinear estimations.

## Methods Based on Linearization

MacKinnon, Nielsen, and Webb (2025) discuss jackknife and bootstrap methods based on linearizing the logit model.

For the logit model, the contributions to the information matrix are

$$J_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \Lambda_{gi}(\boldsymbol{\beta}) \Lambda_{gi}(-\boldsymbol{\beta}) \mathbf{X}_{gi}(\boldsymbol{\beta})^\top \mathbf{X}_{gi}(\boldsymbol{\beta}), \quad g = 1, \dots, G. \quad (40)$$

The estimates from linearizing the model around  $\boldsymbol{\beta}$  are then

$$\mathbf{b}(\boldsymbol{\beta}) = \left( \sum_{g=1}^G J_g(\boldsymbol{\beta}) \right)^{-1} \sum_{g=1}^G \mathbf{s}_g(\boldsymbol{\beta}) = \mathbf{J}(\boldsymbol{\beta})^{-1} \mathbf{s}(\boldsymbol{\beta}). \quad (41)$$

When the  $\mathbf{s}_g(\boldsymbol{\beta})$  and  $J_g(\boldsymbol{\beta})$  are evaluated at  $\boldsymbol{\beta}_0$ , the vector  $\mathbf{b}(\boldsymbol{\beta}_0)$  is a linear approximation to  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$  (Davidson and MacKinnon, 1984).

After we estimate the logit model, we form the cluster-level vectors  $\hat{\mathbf{s}}_g = \mathbf{s}_g(\hat{\boldsymbol{\beta}})$  and matrices  $\hat{J}_g = J_g(\hat{\boldsymbol{\beta}})$  for  $g = 1, \dots, G$ .

The linear approximations to  $\hat{\beta}^{(g)} - \hat{\beta}$  when each cluster is omitted in turn are then

$$\hat{\mathbf{b}}^{(g)} = (\hat{\mathbf{J}} - \hat{\mathbf{J}}_g)^{-1}(\hat{\mathbf{s}} - \hat{\mathbf{s}}_g), \quad g = 1, \dots, G. \quad (42)$$

We can use these approximations to compute cluster-jackknife variance matrices. The one comparable to (10) is

$$\text{CV}_{3L}: \quad \hat{\mathbf{V}}_{3L}(\hat{\beta}) = \frac{G-1}{G} \sum_{g=1}^G \hat{\mathbf{b}}^{(g)} \hat{\mathbf{b}}^{(g)\top}. \quad (43)$$

The linear approximation (41) can also be used to compute **wild cluster linearized**, or **WCL**, bootstraps.

Once the logit model has been estimated (possibly subject to the restrictions to be tested) and linearized, computations are identical to those for the WCR/WCU bootstraps for linear regression models.

Let  $\ddot{x}$  denote  $\hat{x}$  or  $\tilde{x}$ , and  $v_g^{*b}$  be random variates with mean 0 and variance 1 (probably Rademacher). Bootstrap scores are generated by

$$\ddot{s}_g^{*b} = v_g^{*b} \ddot{s}_g, \quad g = 1, \dots, G. \quad (44)$$

Then the bootstrap model is estimated by OLS, yielding

$$\ddot{b}^{*b} = \left( \sum_{g=1}^G \ddot{J}_g \right)^{-1} \sum_{g=1}^G \ddot{s}_g^{*b}. \quad (45)$$

The empirical bootstrap score vectors are

$$\ddot{w}_g^{*b} = \ddot{s}_g^{*b} - \ddot{J}_g \ddot{b}^{*b}, \quad g = 1, \dots, G. \quad (46)$$

The  $CV_1$  bootstrap variance matrix is

$$\ddot{V}_b^* = \frac{G(N-1)}{(G-1)(N-k)} \ddot{J}^{-1} \left( \sum_{g=1}^G \ddot{w}_g^{*b} (\ddot{w}_g^{*b})^\top \right) \ddot{J}^{-1}. \quad (47)$$

- When  $\ddot{s}_g = \tilde{s}_g$  and  $\ddot{j}_g = \tilde{j}_g$ , we have the **WCLR-C bootstrap**.
- When  $\ddot{s}_g = \hat{s}_g$  and  $\ddot{j}_g = \hat{j}_g$ , we have the **WCLU-C bootstrap**.

These are analogous to the classic WCR-C and WCU-C bootstraps.

We can also transform the empirical scores, as proposed in **MacKinnon, Nielsen, and Webb (JAE 2023b)**, to undo some of the deleterious effects of ML estimation.

The transformed scores are

$$\dot{s}_g = \tilde{s}_g - \tilde{j}_{1g} \tilde{b}_1^{(g)} \quad \text{and} \quad \dot{s}_g = \hat{s}_g - \hat{j}_g \hat{b}^{(g)}, \quad g = 1, \dots, G. \quad (48)$$

- When  $\ddot{s}_g = \dot{s}_g$  and  $\ddot{j}_g = \tilde{j}_g$ , we have the **WCLR-S bootstrap**.
- When  $\ddot{s}_g = \dot{s}_g$  and  $\ddot{j}_g = \hat{j}_g$ , we have the **WCLU-S bootstrap**.

These are analogous to the WCR-S and WCU-S bootstraps.

The `logitjack` package computes  $CV_3$ ,  $CV_{3L}$ , all four bootstrap  $P$  values, and confidence intervals based on WCLU-C and WCLU-S.

# An Empirical Application

Porter and Serra (AEJ Applied, 2020) studies female students in Principles of Economics classes in 2015 and 2016.

Some classes in 2016 were exposed to “successful and charismatic women who majored in economics at the same university.”

Dependent variable is 1 if a student took another economics class. Only 21.7% did.

- $N = 627$ ;
- $G = 12$ ;
- $k = 11$ ;
- Four of the classes were “treated” in 2016, so  $G_1 = 4$ ;
- Proportion of observations treated is 0.2073.
- $N_g$  varies from 12 to 104;
- partial leverage varies from 0.016 to 0.155;
- $G^*(0) = 8.490$  and  $G^*(1) = 5.978$ .

Table 1: Effects of Treatment on Taking Another Economics Course

Method	Coef.	Std. error	<i>t</i> -stat.	<i>P</i> value	Lower	Upper
LPM HC <sub>1</sub>	0.1389	0.0673	2.0632	0.0395	0.0067	0.2710
LPM HC <sub>3</sub>	0.1389	0.0680	2.0431	0.0415	0.0054	0.2723
LPM CV <sub>1</sub>	0.1389	0.0518	2.6791	0.0214	0.0248	0.2529
LPM CV <sub>3</sub>	0.1389	0.0646	2.1505	0.0546	-0.0033	0.2810
jregress	0.1389	0.0674	2.0589	0.0504	-0.0004	0.2781
Logit (default)	0.8739	0.4071	2.1467	0.0318	0.0760	1.6717
Logit CV <sub>1</sub> Nml.	0.8739	0.3087	2.8306	0.0046	0.2688	1.4790
Logit CV <sub>1</sub> <i>t</i> (11)	0.8739	0.3112	2.8079	0.0170	0.1889	1.5589
Logit CV <sub>3</sub>	0.8739	0.3905	2.2380	0.0469	0.0144	1.7333
Logit CV <sub>3L</sub>	0.8739	0.3875	2.2554	0.0455	0.0211	1.7266

Table 2: Effects of Treatment Using Bootstrap Methods

Method	Coef.	<i>t</i> -stat.	<i>P</i> value	Lower	Upper
LPM Pairs (stud-boot)	0.1389	2.6791	0.1019	-0.0087	0.3796
LPM Pairs (boot s.e.)	0.1389	2.3108	0.0412	0.0066	0.2711
LPM WCU-C	0.1389	2.6791	0.0332	0.0103	0.2674
LPM WCU-S	0.1389	2.6791	0.0443	0.0034	0.2743
LPM WCR-C	0.1389	2.6791	0.0345	0.0133	0.2617
LPM WCR-S	0.1389	2.6791	0.0404	0.0079	0.2573
Logit WCLU-C (boot s.e.)	0.8739	2.9575	0.0114	0.2235	1.5242
Logit WCLU-S (boot s.e.)	0.8739	2.1602	0.0212	-0.0165	1.7642
Logit WCLR-C	0.8739	2.8079	0.0294		
Logit WCLR-S	0.8739	2.8079	0.0346		

$B = 999,999$ ; Webb (6-point) weights.

LPM results from `boottest`; Logit results from `logitjack`.

Some reported *t*-statistics use bootstrap standard errors, but *P* values are for actual ones.

## Targeted Monte Carlo Experiments

We don't know how the disturbances are generated, so I assume they come from a random-effects model (28) with intra-cluster correlation  $\rho$ . In the experiments, the value of  $\rho$  is varied from 0.00 to 0.50 by 0.05. It matters greatly!

Results for  $\rho = 0$  seem consistent with what we observe.

In fact, the direct estimate of  $\rho$  is  $-0.0042$ .

## Placebo Regressions

The placebo regressor treats the same number of observations in each of the four treated clusters as in the data, but the identities of the treated observations are chosen randomly.

The four treated clusters have 78, 104, 68, and 44 observations.

Of these, 33, 38, 38, and 21 are treated.

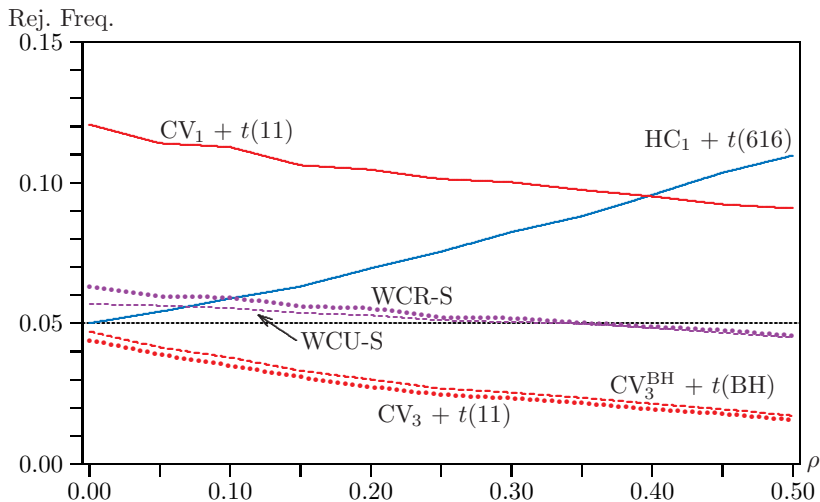
Figure 2. Monte Carlo rejection frequencies as functions of  $\rho$ 

Table 3: *P* Values and Rejection Frequencies

Method	<i>P</i> value	M.C. ( $\rho = 0$ )	M.C. ( $\rho = 0.25$ )	Placebo Reg.
HC <sub>1</sub> + <i>t</i> (616)	0.0395	0.0500	0.0754	0.0528
CV <sub>1</sub> + <i>t</i> (11)	0.0214	0.1206	0.1012	0.1866
CV <sub>3</sub> + <i>t</i> (11)	0.0546	0.0437	0.0243	0.0933
CV <sub>3</sub> <sup>BH</sup> + <i>t</i> (BH)	0.0504	0.0469	0.0266	0.0457
WCU-S	0.0443	0.0567	0.0509	0.0773
WCR-S	0.0404	0.0627	0.0517	0.0613
Pairs Cluster	0.1019	0.0241	0.0158	0.0245
Logit (default)	0.0318	0.0507	0.0769	0.0465
Logit CV <sub>1</sub> <i>t</i> (11)	0.0170	0.0853	0.0745	0.1492
Logit CV <sub>3L</sub> <i>t</i> (11)	0.0455	0.0421	0.0252	0.0895
Logit WCLR-S	0.0346	0.0610	0.0538	0.0558

Rejection frequencies are based on 100,000 replications.

Actual bootstrap tests use  $B = 999,999$ ; simulations use  $B = 999$ .

# Summary

- Unless  $G$  is large and the clusters are well balanced, default cluster-robust standard errors ( $CV_1$ ) can be much too small.
- For treatment models, small values of  $G_1$  or  $G - G_1$  are red flags.
- Luckily, there are several alternatives, including:
  - $CV_3$  standard errors by `summclust` or `vce(jackknife,mse)`;
  - Hansen's modified  $CV_3$  standard errors and critical values computed by `jregress`;
  - Wild cluster bootstrap  $P$  values and confidence intervals, especially WCR-S and WCU-S variants, computed by `boottest`.
- Partial leverages matter even more than cluster sizes. Use `summclust` to compute summary statistics, including  $G^*$ .
- For logit models, `logitjack` provides  $CV_{3L}$  and WCL bootstraps.
- Targeted Monte Carlo experiments and placebo regressions can tell us which  $P$  values or confidence intervals to believe.

# References

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) “How much should we trust differences-in-differences estimates?” *Quarterly Journal of Economics* 119, 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, 137–151

Boot, Tom, Gianmaria Niccodemi, and Tom Wansbeek (2023) “Unbiased estimation of the OLS covariance matrix when the errors are clustered.” *Empirical Economics* 64, 2511–2533

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, 414–427

- Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2011) “Robust inference with multiway clustering.” *Journal of Business & Economic Statistics* 29, 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, 698–709
- Chiang, Harold D., Bruce E. Hansen, and Yuya Sasaki (2024) “Standard errors for two-way clustering with serially correlated time effects.” *Review of Economics and Statistics*, to appear
- Chiang, Harold D., Yuya Sasaki, and Yulong Wang (2025) “Genuinely robust inference for clustered data.” ArXiv e-prints 2308.10138v6
- Davezies, Laurent, Xavier D’Haultfoeuille, and Yannick Guyonvarch (2025) “Analytic inference with two-way clustering.” ArXiv e-prints 2506.20749v1
- Davidson, Russell and James G. MacKinnon (1984), “Convenient specification tests for logit and probit models,” *Journal of Econometrics* 25, 241–262.
- Djogbenou, Antoine A., James G. MacKinnon, and Morten Ø. Nielsen (2019) “Asymptotic theory and wild bootstrap inference with clustered errors.” *Journal of Econometrics* 212, 393–412

Hansen, Bruce E. (2025a) “Jackknife standard errors for clustered regression.” Working Paper, University of Wisconsin

Hansen, Bruce E. (2025b) “Standard errors for difference-in-difference regression.” *Journal of Applied Econometrics* 40, 291–309

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

MacKinnon, James G. (2023) “Fast cluster bootstrap methods for linear regression models.” *Econometrics and Statistics* 26, 52–71

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

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

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2023a) “Cluster-robust inference: A guide to empirical practice.” *Journal of Econometrics* 232, 272–299

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2023b) "Fast jackknife and bootstrap methods for cluster-robust inference." *Journal of Applied Econometrics* 38, 671–694

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2023c) "Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust." *Stata Journal* 23, 942–982

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2023d) "Testing for the appropriate level of clustering in linear regression models." *Journal of Econometrics* 235, 2027–2056

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2024) "Jackknife inference with two-way clustering." arXiv e-prints 2406.08880

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2025) "Cluster-robust jackknife and bootstrap inference for logistic regression models." *Econometric Reviews*, forthcoming.

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

- Miglioretti, Diana L., and Patrick J. Heagerty (2006) "Marginal modeling of nonnested multilevel data using standard software." *American Journal of Epidemiology* 165, 453–463
- Porter, Catherine, and Danila Serra (2020) "Gender differences in the choice of major: The importance of female role models." *American Economic Journal: Applied Economics* 12, 226–254
- 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, 4–60
- Thompson, Samuel B. (2011) "Simple formulas for standard errors that cluster by both firm and time." *Journal of Financial Economics* 99, 1–10