A guide to cluster robust inference using boottest and summclust in Stata

James G. MacKinnon [1] Morten Ørregaard Nielsen [2]
Matthew D. Webb [3]

[1]Queen's University

[2]Aarhus University and CREATES

[3]Carleton University and Ottawa-Carleton Graduate School of Economics

November 18, 2021

2021 Stata Economics Virtual Symposium

Introduction

- This talk is *very* loosely based on MacKinnon, Nielsen and Webb (2021a).
- Brief overview of the cluster robust variance estimator and the wild cluster bootstrap.
- Simulation results for difficult cases.
- Overview of some diagnostic tools, especially `summclust` command.
- Quick summary of the `boottest` command.
- We focus on what Abadie, Athey, Imbens and Wooldridge (2017) calls the "model-based" approach, according to which every sample can be thought of as a random outcome, or drawing, from some meta-population.

Background on Cluster Robust Inference

Consider the following model:

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

If we assume that the data are generated by (1) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, then the OLS estimator of $\boldsymbol{\beta}$ is

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

it follows that:

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

where $\boldsymbol{s}_g = \boldsymbol{X}_g^\top\boldsymbol{u}_g$ denotes the $k \times 1$ score vector corresponding to the g^{th} cluster.

Variance Estimator

Dividing the sample into clusters only becomes meaningful if we further assume that

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

An estimator of the variance of $\hat{\beta}$ should be based on the usual sandwich formula,

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

The natural way to estimate (4) is to replace the $\boldsymbol{\Sigma}_g$ matrices by their empirical counterparts, which yields the cluster-robust variance estimator, or CRVE,

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

What Can Go Wrong

- The CRVE can work well, but the asymptotics depend on G, the number of clusters.
- The CRVE can work poorly when there are few clusters.
- The CRVE also runs into problems when the clusters are heterogeneous:
  - differing size clusters.
  - Unequal distribution of X; cluster specific treatment is an extreme example.
- The wild cluster bootstrap (Cameron, Gelbach and Miller, 2008; Djogbenou, MacKinnon and Nielsen, 2019) often, but not always, works better than the CRVE.

The Wild Cluster Bootstrap

The restricted version of the wild cluster bootstrap (WCR) works as follows:

- Suppose that $\tilde{\beta}$ denotes the OLS estimate of β subject to the restriction $\boldsymbol{a}^\top\beta = \boldsymbol{a}^\top\beta_0$. Then $\tilde{\boldsymbol{u}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\tilde{\beta}$ denotes the vector of restricted residuals for the g^{th} cluster. The Bootstrap DGP is

$$\boldsymbol{y}_g^{*b} = \boldsymbol{X}_g\tilde{\beta} + \boldsymbol{u}_g^{*b}, \quad \boldsymbol{u}_g^{*b} = v_g^{*b}\tilde{\boldsymbol{u}}_g, \quad g = 1, \ldots, G, \tag{6}$$

  where the v_g^{*b} are independent realizations of an auxiliary random variable v^* .
- Typically, the best choice for v^* is the Rademacher distribution, in which case v^* equals 1 or -1 with equal probabilities Davidson and Flachaire (2008), Djogbenou et al. (2019).
- Then B bootstrap samples are generated, the full model is estimated with the bootstrap samples and either a bootstrap P value or C.I. is calculated.

Figure: Rejection frequencies as G changes, $\gamma = 3$, $\rho = 0.10$

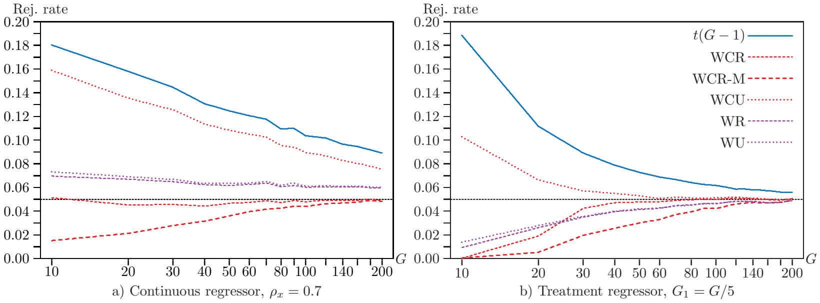


Figure from previous working paper version of Djogbenou, MacKinnon and Nielsen (2019)

Figure: Rejection frequencies for continuous regressor, $G = 20$, $N = 4000$, $\rho = 0.10$

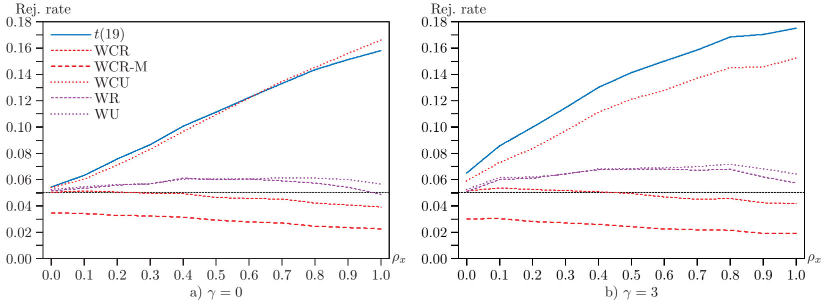


Figure from previous working paper version of Djogbenou, MacKinnon and Nielsen (2019)

Figure: Rejection frequencies for treatment dummy, $G = 20$, $N = 4000$, $\rho = 0.10$

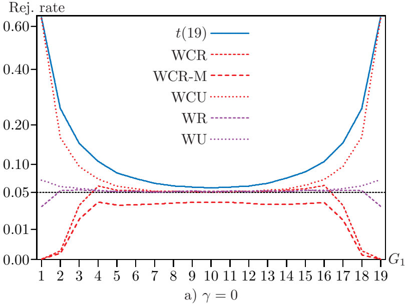


Figure from previous working paper version of Djogbenou, MacKinnon and Nielsen (2019)

When Will CRVEs be Unreliable?

- There are a few diagnostics one can examine to check whether CV_1 is likely to be reliable.
- Carter, Schnepel and Steigerwald (2017) propose the effective number of clusters, G^*.
- This can be calculated using the Stata package `clusteff` described in Lee and Steigerwald (2018).
- The forthcoming `summclust` package calculates G^* more efficiently.
- When G^* differs significantly from G then inference based on $t \sim t(G-1)$ is likely to be unreliable.
- In those situations you can alternatively use WCR or $t \sim t(G^*-1)$; see MacKinnon and Webb (2017) for details.
- The following directly will host summclust on github shortly.

Cluster Level Leverage

MacKinnon, Nielsen and Webb (2021b) proposes a cluster level measure of leverage.

If we drop the g^{th} cluster when we estimate $\boldsymbol{\beta}$, the g^{th} residual vector changes from $\hat{\boldsymbol{u}}_g$ to $(\mathbf{I} - \boldsymbol{H}_g)^{-1}\hat{\boldsymbol{u}}_g$, where

$$\boldsymbol{H}_g = \boldsymbol{X}_g(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_g^\top \tag{7}$$

is the $N_g \times N_g$ diagonal block of the hat matrix that corresponds to cluster g.

As a measure of leverage, we can instead use a matrix norm of the $\boldsymbol{H}_g$.

$$L_g = \mathrm{Tr}(\boldsymbol{H}_g) = \mathrm{Tr}\big(\boldsymbol{X}_g^\top\boldsymbol{X}_g(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big). \tag{8}$$

Partial Leverage

The partial leverage of observation i is simply the i^{th} diagonal element of the matrix $\acute{\boldsymbol{x}}_j(\acute{\boldsymbol{x}}_j^\top\acute{\boldsymbol{x}}_j)^{-1}\acute{\boldsymbol{x}}_j^\top$, which is just $\acute{x}_{ji}^2/(\acute{\boldsymbol{x}}_j^\top\acute{\boldsymbol{x}}_j)$.
The analogous measure of partial leverage for cluster g is

$$L_{gj} = \frac{\acute{\boldsymbol{x}}_{gj}^\top\acute{\boldsymbol{x}}_{gj}}{\acute{\boldsymbol{x}}_j^\top\acute{\boldsymbol{x}}_j}, \tag{9}$$

where $\acute{\boldsymbol{x}}_{gj}$ is the subvector of $\acute{\boldsymbol{x}}_j$ corresponding to the g^{th} cluster.
The average of the L_{gj} is evidently $1/G$, so that if cluster h has $L_{hj} >> 1/G$, it has high partial leverage for the j^{th} coefficient.

Cluster Level Influence

MacKinnon et al. (2021b) also proposes a cluster level measure of influence. As an example, consider using a regression to estimate a sample mean. We can rewrite the expression for $\hat{\beta}$ as

$$\hat{\beta} = \sum_{g=1}^{G} \frac{N_g}{N} \bar{y}_g = \sum_{g=1}^{G} L_g \hat{\beta}_g, \tag{10}$$

so that $\hat{\beta}$ is seen to be a weighted average of the G estimates $\hat{\beta}_g = \bar{y}_g$, with the weight for each cluster equal to its leverage. Similarly, we find that

$$\hat{\beta}^{(g)} = \frac{N}{N - N_g} \sum_{h \neq g} L_h \hat{\beta}_h, \tag{11}$$

Subtracting (10) from (11), we conclude that

$$\hat{\beta}^{(g)} - \hat{\beta} = L_g \big(\hat{\beta}^{(g)} - \hat{\beta}_g\big) = \frac{N_g}{N} \big(\hat{\beta}^{(g)} - \hat{\beta}_g\big). \tag{12}$$

Therefore, cluster g will be influential whenever omitting it yields an estimate $\hat{\beta}^{(g)}$ that differs substantially from the estimate $\hat{\beta}_g$ for cluster g itself, especially when cluster g also has high leverage.

Alternatives to bootstrapping

- While the wild cluster bootstrap works well it can sometimes fail.
- Alternative CRVEs are sometimes reliable but computationally infeasible with large clusters :

$$\text{CV}_2: \qquad (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\Big(\sum_{g=1}^{G} \ddot{\boldsymbol{s}}_g \ddot{\boldsymbol{s}}_g^\top\Big)(\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{13}$$

  In the middle factor here,

$$\ddot{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \boldsymbol{M}_g^{-1/2} \hat{\boldsymbol{u}}_g, \quad \text{where} \quad \boldsymbol{M}_g = \mathbf{I}_{N_g} - \boldsymbol{X}_g(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}_g^\top. \tag{14}$$

  - In Stata, see the `reg_sandwich` package.
  - Tyszler, M., Pustejovsky, J. E., & Tipton, E. 2017.
- See also Randomization Inference and other forms of randomization MacKinnon and Webb (2020), Cai, Canay, Kim and Shaikh (2021), Canay, Romano and Shaikh (2017) and references therein.
  - In Stata, see the `RITEST` package, by Simon Hess.

Multi-way Clustering

- Clustering can occur in more than one dimension.
- Cameron et al. (2011) proposed a variance estimator of $\hat{\beta}$

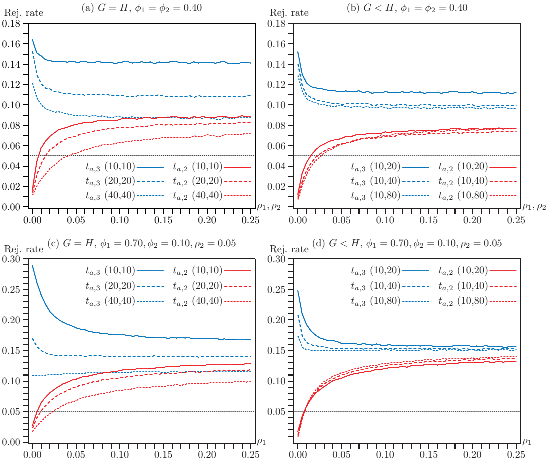
$$\widehat{\text{Var}}(\hat{\beta}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \hat{\Sigma} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

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

- MacKinnon, Nielsen and Webb (2021c) proposes a multi-way cluster bootstrap.
- Multi-way theory is still under active development (Chiang, Kato and Sasaki, 2020; Chiang, Kato, Ma and Sasaki, 2021; Davezies, D'Haultfœuille and Guyonvarch, 2021; Menzel, 2021).

Figure from MacKinnon et al. (2021c)

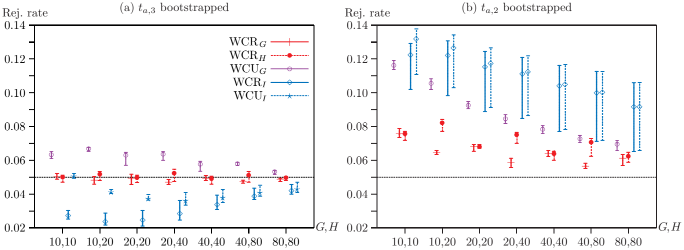
Figure: Rejection frequencies for two-way t-tests



Notes: There are 400,000 replications, and the sample size N is always 6400. All tests are at the 5% nominal level.

Figure from MacKinnon et al. (2021c)

Figure: Rejection frequencies for wild cluster bootstrap tests



Notes: There are 100,000 replications, and $N = 6400$. All bootstrap tests use $B = 399$ and reject whenever $\hat{P}^*_S < 0.05$. In all cases, $\phi_1 = \phi_2 = 0.40$. For each method and each pair of G, H values, the top of the vertical line shows the largest observed rejection frequency across the cases $\rho_1 = \rho_2 = 0.01, 0.02, \ldots, 0.10$, the bottom of the line shows the smallest one, and the mean over the ten frequencies is shown by a symbol.

boottest

- In Stata, the program `boottest` handles many of these routines.
- Roodman et al. (2019) describes the features of the program and how it achieves computational efficiency.
- `boottest` itself is for estimating bootstrap P values and confidence intervals.
- `waldtest` is contained within boottest and can be used for asymptotic P values and confidence intervals.

waldtest

$$y_i = \alpha + \beta x_i + \gamma w_i + \epsilon_i$$

- Imagine you are interested in estimating the above model.
- You want to test the null hypothesis $H_0 : \beta_0 = 0$ under different assumptions about the level of clustering: city, state, etc .
- It can also handle multi-way clustering, such as state and year.

Example

```
reg y x w, robust
waldtest x, cluster(city)
waldtest x, cluster(state)
waldtest x, cluster(state year)
```

boottest

$$y_i = \alpha + \beta x_i + \gamma w_i + \epsilon_i$$

- Consider the same set up as before, but now you wish to use a bootstrap procedure to test the null hypothesis $H_0 : \beta_0 = 0$.
- The following example shows how to do so for: the wild cluster bootstrap WCR (clustering by state); the wild bootstrap WR clustering by state (MacKinnon and Webb, 2018); and multi-way clustered by state and year.

Example

```
reg y x w, robust
boottest x, cluster(state)
boottest x, cluster(state) bootcluster(obsid)
boottest x, cluster(state year) bootcluster(year)
```

Some Guidance

- For each plausible level of clustering examine the distribution of cluster sizes.
- Settle on a level of clustering, perhaps by testing .
- For key regressions report measures of cluster level influence, leverage, and the effective number of clusters, shortly available with `summclust`.
- Employ the wild cluster bootstrap by default, easily done with `boottest`.
- Consider alternative means of inference with few treated clusters.

Bibliography I

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge (2017) 'When should you adjust standard errors for clustering?' Working Paper 24003, National Bureau of Economic Research

Cai, Yong, Ivan A. Canay, Deborah Kim, and Azeem M. Shaikh (2021) 'A user's guide to approximate randomization tests with a small number of clusters.' ArXiv e-prints 2102.09058

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

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh (2017) 'Randomization tests under an approximate symmetry assumption.' *Econometrica* 85, 1013–1030

Bibliography II

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

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

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2021) 'Empirical process results for exchangeable arrays.' *Annals of Statistics* 49, 845–862

Davidson, Russell, and Emmanuel Flachaire (2008) 'The wild bootstrap, tamed at last.' *Journal of Econometrics* 146, 162–169

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

Bibliography III

Lee, Chang Hyung, and Douglas G. Steigerwald (2018) 'Inference for clustered data.' *Stata Journal* 18(2), 447–460

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., and Matthew D. Webb (2020) 'Randomization inference for difference-in-differences with few treated clusters.' *Journal of Econometrics* 218, 435–450

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2021a) 'Cluster-robust inference: A guide to empirical practice.' QED Working Paper 1456

Bibliography IV

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2021b) 'The summclust package: Leverage and influence in clustered regression models.' QED Working Paper, Queen's University

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2021c) 'Wild bootstrap and asymptotic inference with multiway clustering.' *Journal of Business & Economic Statistics* 39, 505–519

Menzel, Konrad (2021) 'Bootstrap with cluster-dependence in two or more dimensions.' *Econometrica* 89, 2143–2188

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