Jackknife methods for improved cluster–robust inference

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

[1]Queen's University

[2]Aarhus University

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

August 3, 2023

2023 - Canadian Stata Conference

Introduction

- Cluster robust inference can be a challenge
- Reliable inference requires at least two things:
  - Getting the level of clustering correct (Ibragimov and Müller, 2016; MacKinnon et al., 2020)
  - Determining whether the asymptotic requirements are satisfied and changing the approach to inference when they are not
- See our guide for an overview "Cluster-robust inference: A guide to empirical practice." (MacKinnon, Nielsen and Webb, 2022a)
- This talk will focus on results in:
  - "Leverage, Influence, and the Jackknife in Clustered Regression Models: Reliable Inference Using summclust" MacKinnon, Nielsen and Webb (2022c)
  - "Fast and reliable jackknife and bootstrap methods for cluster-robust inference." (MacKinnon, Nielsen and Webb, 2022b)

When will the conventional estimator be unreliable?

When is the conventional `reg y x, cluster(clustervarname)` going to be unreliable?

- When there are few clusters
- When the clusters are unbalanced
- When some clusters have high leverage
- When some clusters are highly influential
- When the effective number of clusters G^* is small, and differs from G

What can you do to improve inferences?

- Estimate leverage, influence, and G^* using `summclust`
  - `summclust y x, cluster(clustervarname)`
  - `summclust` will also quickly calculate CV_3
- Alternatively consider the wild cluster bootstrap (Cameron et al., 2008):
  - Available natively in **Stata 18** using:
    - `wildbootstrap reg y x, cluster(clustervarname)`
  - `boottest` is an ado program with a few added features (Roodman et al., 2019):
    - `reg y x`
    - `boottest x, cluster(clustervarname)`
  - One new feature in bootstrap is the WCR-S variant proposed in MacKinnon et al. (2022b)
    - `reg y x`
    - `boottest x, cluster(clustervarname) jackknife`

Nunn and Wanthechekon example - summclust output

```
. summclust trust_neighbors exports ${CTRL}, cluster(eth)  gstar


SUMMCLUST - MacKinnon, Nielsen, and Webb

Cluster summary statistics for exports when clustered by eth.
There are 20027 observations within 185 eth clusters.

Regression Output
```

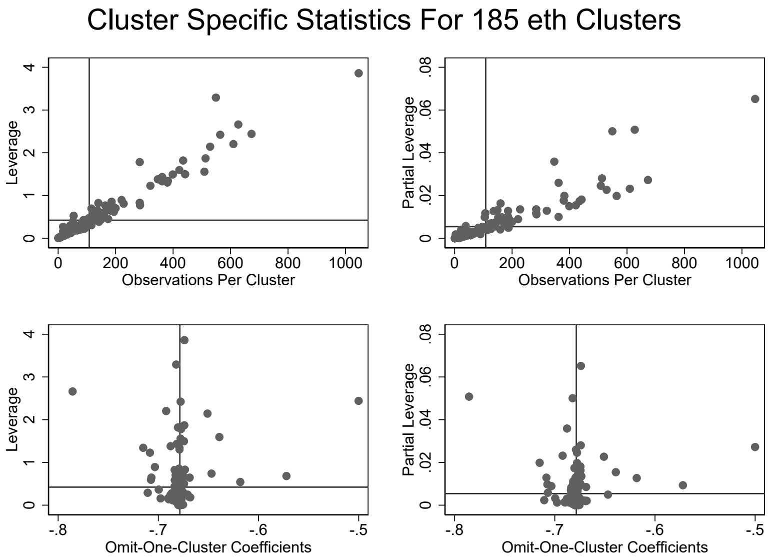
s.e.	Coeff	Sd. Err.	t-stat	P value	CI-lower	CI-upper
CV1	-0.679136	0.142233	-4.7748	0.0000	-0.959752	-0.398520
CV3	-0.679136	0.263065	-2.5816	0.0106	-1.198148	-0.160124

Cluster Variability

Statistic	Ng	Leverage	Partial L.	beta no g
min	1.00	0.002350	0.000000	-0.785572
q1	16.00	0.061718	0.000479	-0.679851
median	44.00	0.179111	0.001866	-0.679144
mean	108.25	0.421622	0.005405	-0.678541
q3	133.00	0.526073	0.006497	-0.678548
max	1046.00	3.861242	0.065182	-0.500066
coefvar	1.44	1.462991	1.676173	0.028645

```
Effective Number of Clusters
----------------------------
G*(0)  = 12.315
G*(1)  =  4.767
----------------------------
```

Figure: Nunn and Wanthechekon example - default summclust figure



Background on Cluster Robust Inference

- Consider the linear regression model

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

  where the data have been divided into G disjoint clusters.
- The $\boldsymbol{y}_g$, $\boldsymbol{X}_g$, and $\boldsymbol{u}_g$ may be stacked into N-vectors $\boldsymbol{y}$, $\boldsymbol{X}$,and $\boldsymbol{u}$, so that (1) can be rewritten as $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$.
- This division is meaningful if we make assumptions about the errors, and the score vectors $\boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g$.
- For a correctly specified model, $\mathrm{E}(\boldsymbol{s}_g) = 0$ for all g. We further assume that

$$\mathrm{E}(\boldsymbol{s}_g \boldsymbol{s}_g^\top) = \boldsymbol{\Sigma}_g \quad \text{and} \quad \mathrm{E}(\boldsymbol{s}_g \boldsymbol{s}_{g'}^\top) = 0, \quad g, g' = 1, \ldots, G, \quad g' \neq g, \tag{2}$$

OLS Estimator of β

- The OLS estimator of β is

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \beta_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u},$$

- It follows that

$$\hat{\beta} - \beta_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{u}_g = \Big(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{X}_g \Big)^{-1} \sum_{g=1}^{G} \boldsymbol{s}_g . \qquad (3)$$

- Inference is usually done by replacing the score vectors $\boldsymbol{s}_g$ with the empirical score vectors $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g$

Variance of $\hat{\beta}$

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

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

- However, we need an estimate of the $\boldsymbol{\Sigma}_g$
- The most common approach is

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

- This is known as CV_1
- The default for "clustered" errors in Stata

Two Other Cluster Robust Variances Estimators

- Bell and McCaffrey (2002) proposed two other estimators CV_2 and CV_3
- CV_2 collapses to HC_2 with singleton clusters
- CV_3 collapses to HC_3 with singleton clusters

$$\text{CV}_3\text{:} \qquad \frac{G-1}{G}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\Big(\sum_{g=1}^{G} \acute{\boldsymbol{s}}_g \acute{\boldsymbol{s}}_g^\top\Big)(\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{6}$$

where $\acute{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \boldsymbol{M}_{gg}^{-1} \hat{\boldsymbol{u}}_g$ and $\boldsymbol{M}_{gg} = \mathrm{I}_{N_g} - \boldsymbol{X}_g(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}_g^\top$.

Issues With CV_2 and CV_3

- Despite CV_2 and CV_3 being proposed two decades ago, and being endorsed in Pustejovsky and Tipton (2018) and Imbens and Kolesár (2016) they have not been used often
- A major limitation is that the $\boldsymbol{M}_{gg}$ can be very large matrices, so storing/inverting these can lead to memory issues
- Two recent papers show how to calculate CV_3 without constructing $\boldsymbol{M}_{gg}^{-1}$ (Niccodemi et al., 2020; Niccodemi and Wansbeek, 2022)
- **Stata 18** has a fast version of CV_2, implemented using:
  - `reg y x, vce(hc2 clustervarname)`
- We instead show how to calculate CV_3 as a jackknife

Two Cluster-Jackknife Variance Estimators

- A cluster-jackknife estimator of $\text{Var}(\hat{\beta})$ is

$$\text{CV}_{3\text{J}}: \qquad \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\beta}^{(g)} - \bar{\beta})(\hat{\beta}^{(g)} - \bar{\beta})^\top, \qquad (7)$$

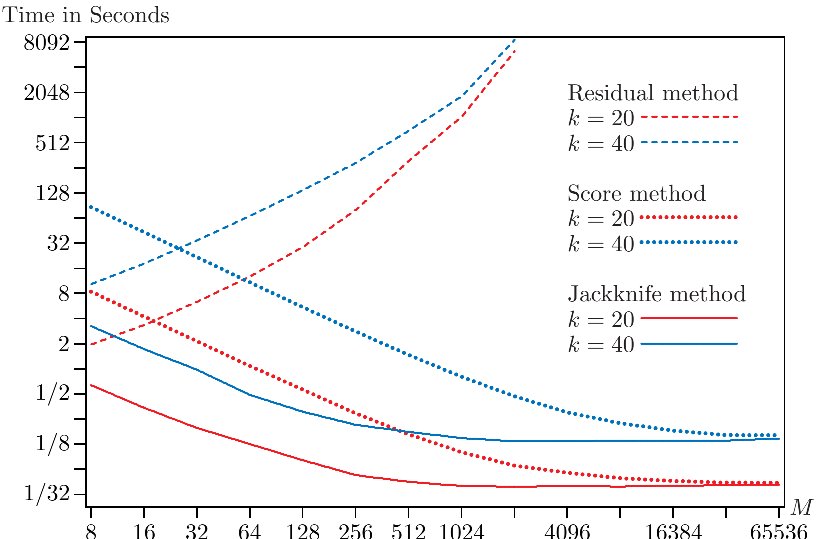
- where $\beta^{(g)}$ are the leave out cluster g estimates of β (more on these later)
- $\bar{\beta}$ is the sample mean of the $\hat{\beta}^{(g)}$
- We can estimate CV_3 in (6) if we replace $\bar{\beta}$ in (7) by $\hat{\beta}$

$$\text{CV}_3: \qquad \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\beta}^{(g)} - \hat{\beta})(\hat{\beta}^{(g)} - \hat{\beta})^\top. \qquad (8)$$

- Brute force versions of these can be estimated in Stata using the `jackknife` prefix, or `vce(jackknife)` or `vce(jackknife, mse)`
- **NB** cluster fixed effects and singular sub-samples cause problems for the native Stata routines

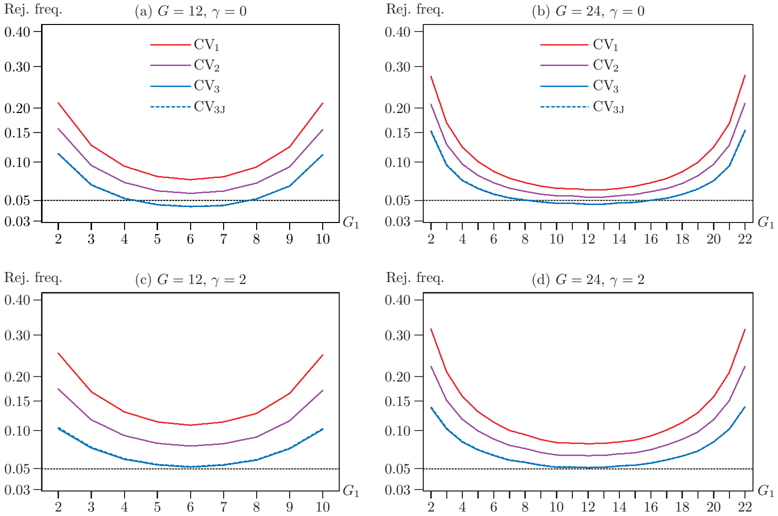
The Jackknife is faster and is feasible for large samples

Figure: Figure from "bootknife" MacKinnon, Nielsen and Webb (2022b)



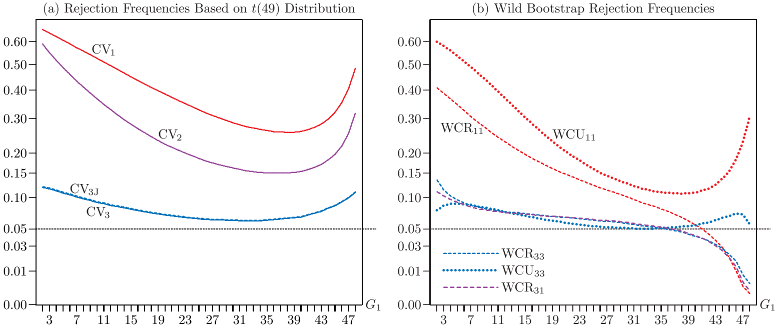
The Jackknife is more reliable

Figure: Figure from "bootknife" MacKinnon, Nielsen and Webb (2022b)



Jackknifing the residuals for the Wild Cluster Bootstrap Really Helps

Figure: Cluster sizes based on state of incorporation in the US



Cluster Level Heterogeneity

- Many simulations and theoretical results have shown that CV_1 is most reliable with a large number of homogeneous clusters Djogbenou, MacKinnon and Nielsen (2019)
- At the observation level there are three classic measures of heterogeneity: leverage, partial leverage, and influence (Belsley, Kuh and Welsch, 1980; Chatterjee and Hadi, 1986)
- Measures of leverage at the observation level are based on how much the residual for observation i changes when we drop that observation from the regression
- If h_i denotes the i^{th} diagonal element of the "hat matrix" $\boldsymbol{H} = \boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, then omitting the i^{th} observation changes the i^{th} residual from $\hat{u}_i$ to $\hat{u}_i/(1-h_i)$.

The Influences of Clusters

- Similarly, dropping the g^{th} cluster when we estimate β changes the g^{th} residual vector from $\hat{\boldsymbol{u}}_g$ to $(\mathrm{I} - \boldsymbol{H}_g)^{-1}\hat{\boldsymbol{u}}_g$, where

$$\boldsymbol{H}_g = \boldsymbol{X}_g(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_g^\top \tag{9}$$

  is the $N_g \times N_g$ diagonal block of $\boldsymbol{H}$ that corresponds to cluster g.
- The matrix $\boldsymbol{H}_g$ is the cluster analog of the scalar h_i.
- These matrices can be large, hence we suggest the scalar:

$$L_g = \mathrm{Tr}(\boldsymbol{H}_g) = \mathrm{Tr}\big(\boldsymbol{X}_g^\top\boldsymbol{X}_g(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big). \tag{10}$$

- The average value of L_g is k/G
- When a particular L_g is much larger than k/G, that cluster is said to have high leverage

Partial Leverage

- We might be interested in what happens if we were to only alter the coefficient of a particular regressor when dropping each cluster.
- For individual observations, Cook and Weisberg (1980) introduced the concept of partial leverage,lLet

$$\acute{\boldsymbol{x}}_j = \left(\mathrm{I} - \boldsymbol{X}_{[j]}\left(\boldsymbol{X}_{[j]}^\top \boldsymbol{X}_{[j]}\right)^{-1} \boldsymbol{X}_{[j]}^\top\right)\boldsymbol{x}_j, \tag{11}$$

  where $\boldsymbol{x}_j$ is the vector of observations on the j^{th} regressor, and $\boldsymbol{X}_{[j]}$ is the matrix of observations on all the other regressors.
- 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 $\acute{x}_{ji}^2/(\acute{\boldsymbol{x}}_j^\top \acute{\boldsymbol{x}}_j)$, where $\acute{x}_{ji}^2$ is the i^{th} element of $\acute{\boldsymbol{x}}_j$.

Cluster Partial Leverage

- The analogous measure of partial leverage for cluster g is

$$L_{gj} = \frac{\acute{\boldsymbol{x}}_{gj}^{\top} \acute{\boldsymbol{x}}_{gj}}{\acute{\boldsymbol{x}}_{j}^{\top} \acute{\boldsymbol{x}}_{j}}, \tag{12}$$

  where $\acute{\boldsymbol{x}}_{gj}$ is the subvector of $\acute{\boldsymbol{x}}_{j}$ corresponding to the g^{th} cluster
- The average partial leverage is $1/G$
- A cluster is said to have high partial leverage when $L_{gj} >> 1/G$
- Examining the empirical distribution of L_{gj} is often useful

Cluster Influence

- We may also be interested directly in what happens to the coefficients when we omit a cluster
- We can do this in a computationally efficient manner, by first constructing

$$\boldsymbol{X}_g^\top \boldsymbol{X}_g \quad \text{and} \quad \boldsymbol{X}_g^\top \boldsymbol{y}_g, \quad g = 1, \ldots, G. \tag{13}$$

- We can then get the vector of estimates when cluster g is deleted is then

$$\hat{\boldsymbol{\beta}}^{(g)} = (\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}_g^\top \boldsymbol{X}_g)^{-1} (\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}_g^\top \boldsymbol{y}_g). \tag{14}$$

- If interest is mostly in a single coefficient one could report all the $\hat{\beta}_j^{(g)}$ for $g = 1, \ldots, G$ in either a histogram or a table.
- The `summclust` package uses these $\hat{\beta}_j^{(g)}$ to report CV_3 standard errors.

What to report

- It is helpful to examine several measures of heterogeneity to determine the reliability of CV1
- We suggest inspecting all of the cluster sizes, (partial) leverages, and omit one cluster coefficients
- Inspecting these as a histogram or as scatter plots can be informative
- One could calculate the scaled variance scaled variance

$$V_s(a_\bullet) = \frac{1}{(G-1)\bar{a}^2} \sum_{g=1}^{G} (a_g - \bar{a})^2, \tag{15}$$

- Alternatively, one could look at alternative means, such as harmonic, geometric, and quadratic

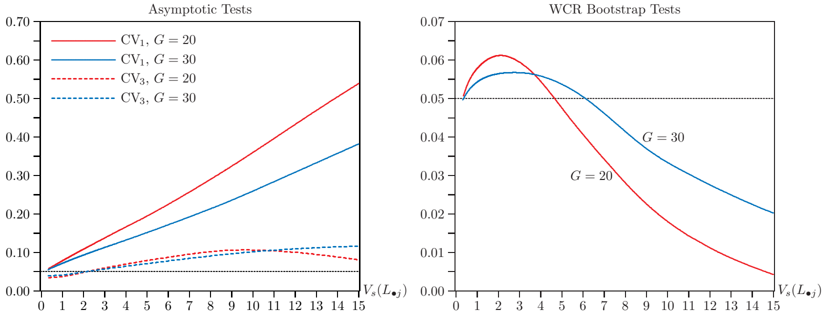
Quick Simulation Experiment

- We are interested in the usefulness of these summary measures in determining when CV1 might be unreliable
- In the simulations there are 2000 (3000) obs divided among 20 (30) clusters, by

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{16}$$

- When $G = 30$, min N_g 7 – 32, max N_g 237 – 396.
- For each sample we calculate the scaled variance of the partial leverages $V_s(L_{\bullet j})$
- We fit the rejection frequency using

$$r_i = \beta_0 + f_1(V_{si}) + f_2(V_{si}^{1/2}) + \beta_1 G_{i0}^* + u_i, \tag{17}$$


Asymptotic Tests
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
CV_1, $G = 20$
CV_1, $G = 30$
CV_3, $G = 20$
CV_3, $G = 30$
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
$V_s(L_{\bullet j})$
WCR Bootstrap Tests
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
$G = 30$
$G = 20$
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
$V_s(L_{\bullet j})$

Conclusion and Future Work

- Determining when cluster robust inference is reliable is challenging
- Inspecting the extent of cluster heterogeneity can help
- We propose cluster level measures of leverage and influence to help detect heterogeneity
- Our measure of influence, allows for rapid calculation of a more reliable variance estimator CV_3 and CV_{3J}
- We also show how to quickly calculate the effective number of clusters
- We developed the `Stata` package `summclust` to make these calculations easy
- Work in progress by us involves extending the cluster jackknife to multi-way clustering and logit models

Bibliography I

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

Belsley, David A., Edwin Kuh, and Roy E. Welsch (1980) *Regression Diagnostics* (New York: Wiley)

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

Chatterjee, Samprit, and Ali S. Hadi (1986) 'Influential observations, high-leverage points, and outliers in linear regression.' *Statistical Science* 1, 379–416

Cook, R. Dennis, and Sanford Weisberg (1980) 'Characterizations of an empirical influence function for detecting influential cases in regression.' *Technometrics* 22, 495–508

Bibliography II

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

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

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

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2022a) 'Cluster-robust inference: A guide to empirical practice.' *Journal of Econometrics* xx, to appear

Bibliography III

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2022b) 'Fast jackknife and bootstrap methods for cluster-robust inference.' QED Working Paper 1485, Queen's University

MacKinnon, James G., Morten Ø. Nielsen, and Matthew D. Webb (2022c) 'Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust.' QED Working Paper 1483, Queen's University

Niccodemi, Gianmaria, and Tom Wansbeek (2022) 'A new estimator for standard errors with few unbalanced clusters.' *Econometrics* 10, 1–7

Niccodemi, Gianmaria, Rob Alessie, Viola Angelini, Jochen Mierau, and Tom Wansbeek (2020) 'Refining clustered standard errors with few clusters.' Working Paper 2020002-EEF, University of Groningen

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 & Economic Statistics* 36, 672–683

Bibliography IV

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