Cluster-Robust Jackknife and Bootstrap Inference for Logistic Regression Models

James G. MacKinnon (Queen's University and ACE)
Morten Ørregaard Nielsen (Aarhus University and ACE)
Matthew D. Webb (Carleton University)

Vanderbilt University, April 16, 2025

Queen's Economics Department

Introduction

Introduction

- We show that existing methods for cluster-robust inference in logistic regression models have mediocre finite-sample properties.

Introduction

- We show that existing methods for cluster-robust inference in logistic regression models have mediocre finite-sample properties.
- We propose alternative procedures based on the **cluster jackknife** and/or the **wild cluster bootstrap**.

Introduction

- We show that existing methods for cluster-robust inference in logistic regression models have mediocre finite-sample properties.
- We propose alternative procedures based on the **cluster jackknife** and/or the **wild cluster bootstrap**.
- Our bootstrap procedures are computationally simple because they are based on empirical score vectors at the cluster level.

Introduction

- We show that existing methods for cluster-robust inference in logistic regression models have mediocre finite-sample properties.
- We propose alternative procedures based on the **cluster jackknife** and/or the **wild cluster bootstrap**.
- Our bootstrap procedures are computationally simple because they are based on empirical score vectors at the cluster level.
- First-order conditions are linearized to obtain approximations to the delete-one-cluster estimates needed for the jackknife.

Introduction

- We show that existing methods for cluster-robust inference in logistic regression models have mediocre finite-sample properties.
- We propose alternative procedures based on the **cluster jackknife** and/or the **wild cluster bootstrap**.
- Our bootstrap procedures are computationally simple because they are based on empirical score vectors at the cluster level.
- First-order conditions are linearized to obtain approximations to the delete-one-cluster estimates needed for the jackknife.
- We also propose four wild cluster bootstrap tests based on the same linear approximation.

Introduction

- We show that existing methods for cluster-robust inference in logistic regression models have mediocre finite-sample properties.
- We propose alternative procedures based on the **cluster jackknife** and/or the **wild cluster bootstrap**.
- Our bootstrap procedures are computationally simple because they are based on empirical score vectors at the cluster level.
- First-order conditions are linearized to obtain approximations to the delete-one-cluster estimates needed for the jackknife.
- We also propose four wild cluster bootstrap tests based on the same linear approximation.
- Two of these transform the scores before bootstrapping, as in MacKinnon, Nielsen, and Webb (JAE 2023).

Introduction

- We show that existing methods for cluster-robust inference in logistic regression models have mediocre finite-sample properties.
- We propose alternative procedures based on the **cluster jackknife** and/or the **wild cluster bootstrap**.
- Our bootstrap procedures are computationally simple because they are based on empirical score vectors at the cluster level.
- First-order conditions are linearized to obtain approximations to the delete-one-cluster estimates needed for the jackknife.
- We also propose four wild cluster bootstrap tests based on the same linear approximation.
- Two of these transform the scores before bootstrapping, as in MacKinnon, Nielsen, and Webb (JAE 2023).
- Two are based on restricted scores, and two are based on unrestricted scores.

Related Literature

Related Literature

The wild cluster bootstrap for linear regression models was proposed in Cameron, Gellbach, and Miller (ReStat 2008).

Related Literature

The wild cluster bootstrap for linear regression models was proposed in Cameron, Gellbach, and Miller (ReStat 2008).

- Its asymptotic validity was proved in Djogbenou, MacKinnon, and Nielsen (JoE 2019).

Related Literature

The wild cluster bootstrap for linear regression models was proposed in Cameron, Gellbach, and Miller (ReStat 2008).

- Its asymptotic validity was proved in Djogbenou, MacKinnon, and Nielsen (JoE 2019).
- Its finite-sample properties were studied in MacKinnon and Webb (JAE 2017, TPM 2017, EctsJ 2018).

Related Literature

The wild cluster bootstrap for linear regression models was proposed in Cameron, Gellbach, and Miller (ReStat 2008).

- Its asymptotic validity was proved in Djogbenou, MacKinnon, and Nielsen (JoE 2019).
- Its finite-sample properties were studied in MacKinnon and Webb (JAE 2017, TPM 2017, EctsJ 2018).
- The relationship with randomization inference was explored in Canay, Santo, and Shaikh (REStat 2021).

Related Literature

The wild cluster bootstrap for linear regression models was proposed in Cameron, Gellbach, and Miller (ReStat 2008).

- Its asymptotic validity was proved in Djogbenou, MacKinnon, and Nielsen (JoE 2019).
- Its finite-sample properties were studied in MacKinnon and Webb (JAE 2017, TPM 2017, EctsJ 2018).
- The relationship with randomization inference was explored in Canay, Santo, and Shaikh (REStat 2021).
- Improved versions related to the cluster jackknife were proposed in MacKinnon, Nielsen, and Webb (JAE 2023).

Related Literature

The wild cluster bootstrap for linear regression models was proposed in Cameron, Gellbach, and Miller (ReStat 2008).

- Its asymptotic validity was proved in Djogbenou, MacKinnon, and Nielsen (JoE 2019).
- Its finite-sample properties were studied in MacKinnon and Webb (JAE 2017, TPM 2017, EctsJ 2018).
- The relationship with randomization inference was explored in Canay, Santo, and Shaikh (REStat 2021).
- Improved versions related to the cluster jackknife were proposed in MacKinnon, Nielsen, and Webb (JAE 2023).

A computationally efficient Stata package called `boottest` is described in Roodman, MacKinnon, Nielsen, and Webb (SJ 2019). Computational issues are discussed in MacKinnon (E&S 2023).

Using the delete-one jackknife to estimate variances was studied in Efron and Stein (Ann. Stat. 1981).

Using the delete-one jackknife to estimate variances was studied in Efron and Stein (Ann. Stat. 1981).

An early application to heteroskedasticity-robust estimation was the (original) HC_3 estimator of MacKinnon and White (JoE 1985).

Using the delete-one jackknife to estimate variances was studied in Efron and Stein (Ann. Stat. 1981).

An early application to heteroskedasticity-robust estimation was the (original) HC_3 estimator of MacKinnon and White (JoE 1985).

Using the cluster jackknife was proposed in Bell and McCaffrey (SM 2002), but they computed it like HC_3.

Using the delete-one jackknife to estimate variances was studied in Efron and Stein (Ann. Stat. 1981).

An early application to heteroskedasticity-robust estimation was the (original) HC_3 estimator of MacKinnon and White (JoE 1985).

Using the cluster jackknife was proposed in Bell and McCaffrey (SM 2002), but they computed it like HC_3.

Better computational methods for not-small clusters were discussed in MacKinnon, Nielsen, and Webb (SJ 2023).

Using the delete-one jackknife to estimate variances was studied in Efron and Stein (Ann. Stat. 1981).

An early application to heteroskedasticity-robust estimation was the (original) HC_3 estimator of MacKinnon and White (JoE 1985).

Using the cluster jackknife was proposed in Bell and McCaffrey (SM 2002), but they computed it like HC_3.

Better computational methods for not-small clusters were discussed in MacKinnon, Nielsen, and Webb (SJ 2023).

- It provides a `Stata` package called `summclust`, which computes the CV_3 variance matrix as well as cluster-level measurs of leverage and influence.

Using the delete-one jackknife to estimate variances was studied in Efron and Stein (Ann. Stat. 1981).

An early application to heteroskedasticity-robust estimation was the (original) HC_3 estimator of MacKinnon and White (JoE 1985).

Using the cluster jackknife was proposed in Bell and McCaffrey (SM 2002), but they computed it like HC_3.

Better computational methods for not-small clusters were discussed in MacKinnon, Nielsen, and Webb (SJ 2023).

- It provides a `Stata` package called `summclust`, which computes the CV_3 variance matrix as well as cluster-level measurs of leverage and influence.

Hansen (2024, JAE 2025) proves interesting results about CV_3 and proposes an inferential procedure based on adjusting the standard error and computing a degrees-of-freedom parameter.

Using the delete-one jackknife to estimate variances was studied in Efron and Stein (Ann. Stat. 1981).

An early application to heteroskedasticity-robust estimation was the (original) HC_3 estimator of MacKinnon and White (JoE 1985).

Using the cluster jackknife was proposed in Bell and McCaffrey (SM 2002), but they computed it like HC_3.

Better computational methods for not-small clusters were discussed in MacKinnon, Nielsen, and Webb (SJ 2023).

- It provides a `Stata` package called `summclust`, which computes the CV_3 variance matrix as well as cluster-level measurs of leverage and influence.

Hansen (2024, JAE 2025) proves interesting results about CV_3 and proposes an inferential procedure based on adjusting the standard error and computing a degrees-of-freedom parameter.

- Hansen provides a `Stata` package called `jregress`.

Logistic Regression Models

Logistic Regression Models

There are N observations divided among G clusters, with the g^{th} cluster containing N_g of them.

Logistic Regression Models

There are N observations divided among G clusters, with the g^{th} cluster containing N_g of them.

Let y_{gi} (binary) be the response for observation i in cluster g.

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

Logistic Regression Models

There are N observations divided among G clusters, with the g^{th} cluster containing N_g of them.

Let y_{gi} (binary) be the response for observation i in cluster g.

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

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

Logistic Regression Models

There are N observations divided among G clusters, with the g^{th} cluster containing N_g of them.

Let y_{gi} (binary) be the response for observation i in cluster g.

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

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

In (1), $\Lambda(\cdot)$ is the logistic function,

$$\Lambda(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}, \tag{2}$$

Logistic Regression Models

There are N observations divided among G clusters, with the g^{th} cluster containing N_g of them.

Let y_{gi} (binary) be the response for observation i in cluster g.

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

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

In (1), $\Lambda(\cdot)$ is the logistic function,

$$\Lambda(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}, \tag{2}$$

which has first derivative

$$\lambda(x) = \frac{e^x}{(1 + e^x)^2} = \Lambda(x)\Lambda(-x). \tag{3}$$

The pseudo-loglikelihood function for (1) is

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \big(y_{gi} \log \Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta}) + (1 - y_{gi}) \log \Lambda(-\boldsymbol{X}_{gi}\boldsymbol{\beta})\big). \tag{4}$$

The pseudo-loglikelihood function for (1) is

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \big(y_{gi} \log \Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta}) + (1 - y_{gi}) \log \Lambda(-\boldsymbol{X}_{gi}\boldsymbol{\beta}) \big). \tag{4}$$

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

$$\boldsymbol{s}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \boldsymbol{s}_{gi}(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \big(y_{gi} - \Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta}) \big) \boldsymbol{X}_{gi}. \tag{5}$$

The pseudo-loglikelihood function for (1) is

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \big(y_{gi} \log \Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta}) + (1 - y_{gi}) \log \Lambda(-\boldsymbol{X}_{gi}\boldsymbol{\beta}) \big). \tag{4}$$

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

$$\boldsymbol{s}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \boldsymbol{s}_{gi}(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \big(y_{gi} - \Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta}) \big) \boldsymbol{X}_{gi}. \tag{5}$$

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

$$\hat{\boldsymbol{s}} = \sum_{g=1}^{G} \hat{\boldsymbol{s}}_g = \sum_{g=1}^{G} \boldsymbol{s}_g(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}. \tag{6}$$

The pseudo-loglikelihood function for (1) is

$$\ell(\boldsymbol{y},\boldsymbol{\beta}) = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\big(y_{gi}\log\Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta}) + (1-y_{gi})\log\Lambda(-\boldsymbol{X}_{gi}\boldsymbol{\beta})\big). \tag{4}$$

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

$$\boldsymbol{s}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g}\boldsymbol{s}_{gi}(\boldsymbol{\beta}) = \sum_{i=1}^{N_g}\big(y_{gi} - \Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta})\big)\boldsymbol{X}_{gi}. \tag{5}$$

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

$$\hat{\boldsymbol{s}} = \sum_{g=1}^{G}\hat{\boldsymbol{s}}_g = \sum_{g=1}^{G}\boldsymbol{s}_g(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{6}$$

When the observations are independent,

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{a}{=} \Big(\operatorname{plim} N^{-1}\boldsymbol{H}(\boldsymbol{\beta}_0)\Big)^{-1} N^{-1/2}\sum_{i=1}^{N}\boldsymbol{s}_i(\boldsymbol{\beta}_0). \tag{7}$$

In the absence of clustering, (7) leads to the variance matrix estimator

$$\hat{V}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}, \tag{8}$$

In the absence of clustering, (7) leads to the variance matrix estimator

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}, \tag{8}$$

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

$$\Upsilon_i(\boldsymbol{\beta}) = \Lambda(\boldsymbol{X}_i\boldsymbol{\beta})\Lambda(-\boldsymbol{X}_i\boldsymbol{\beta}); \tag{9}$$

In the absence of clustering, (7) leads to the variance matrix estimator

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}, \tag{8}$$

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

$$\Upsilon_i(\boldsymbol{\beta}) = \Lambda(\boldsymbol{X}_i\boldsymbol{\beta})\Lambda(-\boldsymbol{X}_i\boldsymbol{\beta}); \tag{9}$$

Note that, for the logit model, $\boldsymbol{X}^\top \mathbf{Y}(\boldsymbol{\beta})\boldsymbol{X} = -\boldsymbol{H}(\boldsymbol{\beta})$. This is not true for the probit model.

In the absence of clustering, (7) leads to the variance matrix estimator

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}, \tag{8}$$

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

$$\Upsilon_i(\boldsymbol{\beta}) = \Lambda(\boldsymbol{X}_i\boldsymbol{\beta})\Lambda(-\boldsymbol{X}_i\boldsymbol{\beta}); \tag{9}$$

Note that, for the logit model, $\boldsymbol{X}^\top \mathbf{Y}(\boldsymbol{\beta})\boldsymbol{X} = -\boldsymbol{H}(\boldsymbol{\beta})$. This is not true for the probit model.

The usual cluster-robust variance matrix (**CRVE**) is

$$\mathrm{CV}_{1\mathcal{I}}: \quad \hat{\boldsymbol{V}}_{1\mathcal{I}} = \frac{G}{G-1}\frac{N-1}{N-k}\,(\boldsymbol{X}^\top \hat{\mathbf{Y}} \boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top \right) (\boldsymbol{X}^\top \hat{\mathbf{Y}} \boldsymbol{X})^{-1}. \tag{10}$$

In the absence of clustering, (7) leads to the variance matrix estimator

$$\hat{V}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}, \tag{8}$$

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

$$\mathrm{Y}_i(\boldsymbol{\beta}) = \Lambda(\boldsymbol{X}_i\boldsymbol{\beta})\Lambda(-\boldsymbol{X}_i\boldsymbol{\beta}); \tag{9}$$

Note that, for the logit model, $\boldsymbol{X}^\top \mathbf{Y}(\boldsymbol{\beta})\boldsymbol{X} = -\boldsymbol{H}(\boldsymbol{\beta})$. This is not true for the probit model.

The usual cluster-robust variance matrix (**CRVE**) is

$$\text{CV}_{1\mathcal{I}}: \quad \hat{V}_{1\mathcal{I}} = \frac{G}{G-1}\frac{N-1}{N-k}\,(\boldsymbol{X}^\top \hat{\mathbf{Y}}\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top\right)(\boldsymbol{X}^\top \hat{\mathbf{Y}}\boldsymbol{X})^{-1}. \tag{10}$$

The empirical score vectors here are

$$\boldsymbol{s}_g(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N_g} \big(y_{gi} - \Lambda(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})\big)\boldsymbol{X}_{gi}, \;\; g = 1, \ldots, G. \tag{11}$$

The Cluster Jackknife

The Cluster Jackknife

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

$$\text{CV}_3: \qquad \hat{\boldsymbol{V}}_3(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^{\top}. \tag{12}$$

The Cluster Jackknife

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

$$\text{CV}_3: \qquad \hat{\boldsymbol{V}}_3(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^{\top}. \tag{12}$$

Another cluster jackknife CRVE uses $\bar{\boldsymbol{\beta}}$ instead of $\hat{\boldsymbol{\beta}}$.

The Cluster Jackknife

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

$$\text{CV}_3: \qquad \hat{\boldsymbol{V}}_3(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^\top. \tag{12}$$

Another cluster jackknife CRVE uses $\bar{\boldsymbol{\beta}}$ instead of $\hat{\boldsymbol{\beta}}$.

Computing CV_3 requires $G+1$ nonlinear estimations.

The Cluster Jackknife

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

$$\text{CV}_3: \qquad \hat{\boldsymbol{V}}_3(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^{G} (\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^\top. \tag{12}$$

Another cluster jackknife CRVE uses $\bar{\boldsymbol{\beta}}$ instead of $\hat{\boldsymbol{\beta}}$.

Computing CV_3 requires $G+1$ nonlinear estimations.

We focus on t-statistics of the form

$$t_a = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a})^{1/2}}. \tag{13}$$

The Cluster Jackknife

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

$$\text{CV}_3: \qquad \hat{\boldsymbol{V}}_3(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^\top. \tag{12}$$

Another cluster jackknife CRVE uses $\bar{\boldsymbol{\beta}}$ instead of $\hat{\boldsymbol{\beta}}$.

Computing CV_3 requires $G+1$ nonlinear estimations.

We focus on t-statistics of the form

$$t_a = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\boldsymbol{a}^\top\hat{\boldsymbol{V}}\boldsymbol{a})^{1/2}}. \tag{13}$$

For the restriction $\beta_k = 0$, we have $t_a = \hat{\beta}_k/\hat{s}_k$, where $\hat{s}_k$ is the square root of the k^{th} diagonal element of $\hat{\boldsymbol{V}}$.

The Cluster Jackknife

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

$$\text{CV}_3: \qquad \hat{\boldsymbol{V}}_3(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^{G} (\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}})^{\top}. \tag{12}$$

Another cluster jackknife CRVE uses $\bar{\boldsymbol{\beta}}$ instead of $\hat{\boldsymbol{\beta}}$.

Computing CV_3 requires $G+1$ nonlinear estimations.

We focus on t-statistics of the form

$$t_a = \frac{\boldsymbol{a}^{\top}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\boldsymbol{a}^{\top}\hat{\boldsymbol{V}}\boldsymbol{a})^{1/2}}. \tag{13}$$

For the restriction $\beta_k = 0$, we have $t_a = \hat{\beta}_k / \hat{s}_k$, where $\hat{s}_k$ is the square root of the k^{th} diagonal element of $\hat{\boldsymbol{V}}$.

It is customary to compare t_a with the $t(G-1)$ distribution.

Methods Based on Linearization

Methods Based on Linearization

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

$$\boldsymbol{J}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \Lambda_{gi}(\boldsymbol{\beta})\Lambda_{gi}(-\boldsymbol{\beta})\boldsymbol{X}_{gi}(\boldsymbol{\beta})^\top \boldsymbol{X}_{gi}(\boldsymbol{\beta}), \quad g = 1, \ldots, G. \tag{14}$$

Methods Based on Linearization

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

$$\boldsymbol{J}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \Lambda_{gi}(\boldsymbol{\beta})\Lambda_{gi}(-\boldsymbol{\beta})\boldsymbol{X}_{gi}(\boldsymbol{\beta})^\top \boldsymbol{X}_{gi}(\boldsymbol{\beta}), \;\; g = 1, \ldots, G. \tag{14}$$

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

$$\boldsymbol{b}(\boldsymbol{\beta}) = \left(\sum_{g=1}^{G} \boldsymbol{J}_g(\boldsymbol{\beta})\right)^{-1} \sum_{g=1}^{G} \boldsymbol{s}_g(\boldsymbol{\beta}) = \boldsymbol{J}(\boldsymbol{\beta})^{-1}\boldsymbol{s}(\boldsymbol{\beta}). \tag{15}$$

Methods Based on Linearization

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

$$J_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \Lambda_{gi}(\boldsymbol{\beta})\Lambda_{gi}(-\boldsymbol{\beta})\boldsymbol{X}_{gi}(\boldsymbol{\beta})^\top \boldsymbol{X}_{gi}(\boldsymbol{\beta}), \quad g = 1, \ldots, G. \tag{14}$$

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

$$\boldsymbol{b}(\boldsymbol{\beta}) = \left(\sum_{g=1}^{G} \boldsymbol{J}_g(\boldsymbol{\beta})\right)^{-1} \sum_{g=1}^{G} \boldsymbol{s}_g(\boldsymbol{\beta}) = \boldsymbol{J}(\boldsymbol{\beta})^{-1}\boldsymbol{s}(\boldsymbol{\beta}). \tag{15}$$

When the $\boldsymbol{s}_g(\boldsymbol{\beta})$ and $\boldsymbol{J}_g(\boldsymbol{\beta})$ are evaluated at $\boldsymbol{\beta}_0$, the vector $\boldsymbol{b}(\boldsymbol{\beta}_0)$ provides a linear approximation to $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$.

Methods Based on Linearization

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

$$\boldsymbol{J}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \Lambda_{gi}(\boldsymbol{\beta})\Lambda_{gi}(-\boldsymbol{\beta})\boldsymbol{X}_{gi}(\boldsymbol{\beta})^\top \boldsymbol{X}_{gi}(\boldsymbol{\beta}), \;\; g = 1, \ldots, G. \tag{14}$$

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

$$\boldsymbol{b}(\boldsymbol{\beta}) = \Big(\sum_{g=1}^{G} \boldsymbol{J}_g(\boldsymbol{\beta})\Big)^{-1} \sum_{g=1}^{G} \boldsymbol{s}_g(\boldsymbol{\beta}) = \boldsymbol{J}(\boldsymbol{\beta})^{-1}\boldsymbol{s}(\boldsymbol{\beta}). \tag{15}$$

When the $\boldsymbol{s}_g(\boldsymbol{\beta})$ and $\boldsymbol{J}_g(\boldsymbol{\beta})$ are evaluated at $\boldsymbol{\beta}_0$, the vector $\boldsymbol{b}(\boldsymbol{\beta}_0)$ provides a linear approximation to $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$.

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

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

$$\hat{\boldsymbol{b}}^{(g)} = (\hat{\boldsymbol{J}} - \hat{\boldsymbol{J}}_g)^{-1}(\hat{\boldsymbol{s}} - \hat{\boldsymbol{s}}_g), \quad g = 1, \ldots, G. \tag{16}$$

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

$$\hat{\boldsymbol{b}}^{(g)} = (\hat{\boldsymbol{J}} - \hat{\boldsymbol{J}}_g)^{-1}(\hat{\boldsymbol{s}} - \hat{\boldsymbol{s}}_g), \quad g = 1, \ldots, G. \tag{16}$$

We can use these approximations to compute cluster-jackknife variance matrices. The one comparable to (12) is

$$\text{CV}_{3\text{L}}: \qquad \hat{\boldsymbol{V}}_{3\text{L}}(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^{G} \hat{\boldsymbol{b}}^{(g)} \hat{\boldsymbol{b}}^{(g)\top}. \tag{17}$$

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

$$\hat{\boldsymbol{b}}^{(g)} = (\hat{\boldsymbol{J}} - \hat{\boldsymbol{J}}_g)^{-1}(\hat{\boldsymbol{s}} - \hat{\boldsymbol{s}}_g), \quad g = 1, \ldots, G. \tag{16}$$

We can use these approximations to compute cluster-jackknife variance matrices. The one comparable to (12) is

$$\mathrm{CV}_{3\mathrm{L}}: \qquad \hat{\boldsymbol{V}}_{3\mathrm{L}}(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^{G} \hat{\boldsymbol{b}}^{(g)} \hat{\boldsymbol{b}}^{(g)\top}. \tag{17}$$

The linear approximation (15) can also be used to compute **wild cluster linearized**, or **WCL**, bootstraps.

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

$$\hat{\boldsymbol{b}}^{(g)} = (\hat{\boldsymbol{J}} - \hat{\boldsymbol{J}}_g)^{-1}(\hat{\boldsymbol{s}} - \hat{\boldsymbol{s}}_g), \quad g = 1, \ldots, G. \tag{16}$$

We can use these approximations to compute cluster-jackknife variance matrices. The one comparable to (12) is

$$\text{CV}_{3\text{L}}: \qquad \hat{\boldsymbol{V}}_{3\text{L}}(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^{G} \hat{\boldsymbol{b}}^{(g)} \hat{\boldsymbol{b}}^{(g)\top}. \tag{17}$$

The linear approximation (15) 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.

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

$$\hat{\boldsymbol{b}}^{(g)} = (\hat{\boldsymbol{J}} - \hat{\boldsymbol{J}}_g)^{-1}(\hat{\boldsymbol{s}} - \hat{\boldsymbol{s}}_g), \quad g = 1, \ldots, G. \tag{16}$$

We can use these approximations to compute cluster-jackknife variance matrices. The one comparable to (12) is

$$\text{CV}_{3\text{L}}: \qquad \hat{\boldsymbol{V}}_{3\text{L}}(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^{G} \hat{\boldsymbol{b}}^{(g)} \hat{\boldsymbol{b}}^{(g)\top}. \tag{17}$$

The linear approximation (15) 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.

The same linearization can also be used to obtain $\text{CV}_{2\text{L}}$.

Four WCL Bootstrap Methods

Four WCL Bootstrap Methods

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

$$\ddot{\mathbf{s}}_g^{*b} = v_g^{*b}\ddot{\mathbf{s}}_g, \quad g = 1, \ldots, G. \tag{18}$$

Four WCL Bootstrap Methods

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

$$\ddot{\boldsymbol{s}}_g^{*b} = v_g^{*b}\ddot{\boldsymbol{s}}_g, \quad g = 1, \ldots, G. \tag{18}$$

Then the bootstrap model is estimated by OLS, yielding

$$\ddot{\boldsymbol{b}}^{*b} = \left(\sum_{g=1}^{G} \ddot{\boldsymbol{J}}_g\right)^{-1} \sum_{g=1}^{G} \ddot{\boldsymbol{s}}_g^{*b}. \tag{19}$$

Four WCL Bootstrap Methods

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

$$\ddot{\mathbf{s}}_g^{*b} = v_g^{*b}\ddot{\mathbf{s}}_g, \quad g = 1, \ldots, G. \tag{18}$$

Then the bootstrap model is estimated by OLS, yielding

$$\ddot{\boldsymbol{b}}^{*b} = \Bigl(\sum_{g=1}^{G} \ddot{\boldsymbol{J}}_g\Bigr)^{-1} \sum_{g=1}^{G} \ddot{\mathbf{s}}_g^{*b}. \tag{19}$$

The empirical bootstrap score vectors are

$$\ddot{\boldsymbol{w}}_g^{*b} = \ddot{\mathbf{s}}_g^{*b} - \ddot{\boldsymbol{J}}_g \ddot{\boldsymbol{b}}^{*b}, \quad g = 1, \ldots, G. \tag{20}$$

Four WCL Bootstrap Methods

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

$$\ddot{\boldsymbol{s}}_g^{*b} = v_g^{*b}\ddot{\boldsymbol{s}}_g, \quad g = 1, \ldots, G. \tag{18}$$

Then the bootstrap model is estimated by OLS, yielding

$$\ddot{\boldsymbol{b}}^{*b} = \Big(\sum_{g=1}^{G} \ddot{\boldsymbol{J}}_g\Big)^{-1} \sum_{g=1}^{G} \ddot{\boldsymbol{s}}_g^{*b}. \tag{19}$$

The empirical bootstrap score vectors are

$$\ddot{\boldsymbol{w}}_g^{*b} = \ddot{\boldsymbol{s}}_g^{*b} - \ddot{\boldsymbol{J}}_g \ddot{\boldsymbol{b}}^{*b}, \quad g = 1, \ldots, G. \tag{20}$$

The CV_1 bootstrap variance matrix is

$$\ddot{\boldsymbol{V}}_b^{*} = \frac{G(N-1)}{(G-1)(N-k)}\, \ddot{\boldsymbol{J}}^{-1}\Big(\sum_{g=1}^{G} \ddot{\boldsymbol{w}}_g^{*b}(\ddot{\boldsymbol{w}}_g^{*b})^\top\Big)\ddot{\boldsymbol{J}}^{-1}. \tag{21}$$

- When $\ddot{\boldsymbol{s}}_g = \tilde{\boldsymbol{s}}_g$ and $\ddot{\boldsymbol{J}}_g = \tilde{\boldsymbol{J}}_g$, we have the WCLR-C bootstrap.

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

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

These are analogous to the classic WCR-C and WCU-C bootstraps for linear regression models.

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

These are analogous to the classic WCR-C and WCU-C bootstraps for linear regression models.

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

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

These are analogous to the classic WCR-C and WCU-C bootstraps for linear regression models.

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

The transformed scores are

$$\dot{\boldsymbol{s}}_g = \tilde{\boldsymbol{s}}_g - \tilde{\boldsymbol{J}}_{1g}\tilde{\boldsymbol{b}}_1^{(g)} \quad \text{and} \quad \acute{\boldsymbol{s}}_g = \hat{\boldsymbol{s}}_g - \hat{\boldsymbol{J}}_g\hat{\boldsymbol{b}}^{(g)}, \quad g = 1, \ldots, G. \tag{22}$$

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

These are analogous to the classic WCR-C and WCU-C bootstraps for linear regression models.

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

The transformed scores are

$$\dot{\boldsymbol{s}}_g = \tilde{\boldsymbol{s}}_g - \tilde{\boldsymbol{J}}_{1g}\tilde{\boldsymbol{b}}_1^{(g)} \quad \text{and} \quad \acute{\boldsymbol{s}}_g = \hat{\boldsymbol{s}}_g - \hat{\boldsymbol{J}}_g\hat{\boldsymbol{b}}^{(g)}, \quad g = 1, \ldots, G. \tag{22}$$

- When $\ddot{\boldsymbol{s}}_g = \dot{\boldsymbol{s}}_g$ and $\ddot{\boldsymbol{J}}_g = \tilde{\boldsymbol{J}}_g$, we have the WCLR-S bootstrap.

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

These are analogous to the classic WCR-C and WCU-C bootstraps for linear regression models.

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

The transformed scores are

$$\dot{\boldsymbol{s}}_g = \tilde{\boldsymbol{s}}_g - \tilde{\boldsymbol{J}}_{1g}\tilde{\boldsymbol{b}}_1^{(g)} \quad \text{and} \quad \acute{\boldsymbol{s}}_g = \hat{\boldsymbol{s}}_g - \hat{\boldsymbol{J}}_g\hat{\boldsymbol{b}}^{(g)}, \quad g = 1, \ldots, G. \tag{22}$$

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

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

These are analogous to the classic WCR-C and WCU-C bootstraps for linear regression models.

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

The transformed scores are

$$\dot{\boldsymbol{s}}_g = \tilde{\boldsymbol{s}}_g - \tilde{\boldsymbol{J}}_{1g}\tilde{\boldsymbol{b}}_1^{(g)} \quad \text{and} \quad \acute{\boldsymbol{s}}_g = \hat{\boldsymbol{s}}_g - \hat{\boldsymbol{J}}_g\hat{\boldsymbol{b}}^{(g)}, \quad g = 1, \ldots, G. \tag{22}$$

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

These are analogous to the WCR-S and WCU-S bootstraps for linear regression models.

The Linear Probability Model (LPM)

The Linear Probability Model (LPM)

Instead of linearizing a logit model, we could just estimate the **LPM**

$$y_{gi} = \boldsymbol{X}_{gi}\delta + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{23}$$

where u_{gi} is a disturbance term with rather odd properties, and then use the classic wild cluster bootstrap or the new -S variants.

The Linear Probability Model (LPM)

Instead of linearizing a logit model, we could just estimate the **LPM**

$$y_{gi} = \boldsymbol{X}_{gi}\delta + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{23}$$

where u_{gi} is a disturbance term with rather odd properties, and then use the classic wild cluster bootstrap or the new -S variants.

For the WCR-C bootstrap, the score vector is

$$\sum_{i=1}^{N_g}(y^*_{gi} - \boldsymbol{X}_{gi}\tilde{\delta})\boldsymbol{X}_{gi} = \begin{cases} \sum_{i=1}^{N_g}(y_{gi} - \boldsymbol{X}_{gi}\tilde{\delta})\boldsymbol{X}_{gi} \text{ with prob. } 1/2, \\ \sum_{i=1}^{N_g}(\boldsymbol{X}_{gi}\tilde{\delta} - y_{gi})\boldsymbol{X}_{gi} \text{ with prob. } 1/2. \end{cases} \tag{24}$$

The Linear Probability Model (LPM)

Instead of linearizing a logit model, we could just estimate the **LPM**

$$y_{gi} = \boldsymbol{X}_{gi}\delta + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{23}$$

where u_{gi} is a disturbance term with rather odd properties, and then use the classic wild cluster bootstrap or the new -S variants.

For the WCR-C bootstrap, the score vector is

$$\sum_{i=1}^{N_g} (y^*_{gi} - \boldsymbol{X}_{gi}\tilde{\delta})\boldsymbol{X}_{gi} = \begin{cases} \sum_{i=1}^{N_g} (y_{gi} - \boldsymbol{X}_{gi}\tilde{\delta})\boldsymbol{X}_{gi} \text{ with prob. } 1/2, \\ \sum_{i=1}^{N_g} (\boldsymbol{X}_{gi}\tilde{\delta} - y_{gi})\boldsymbol{X}_{gi} \text{ with prob. } 1/2. \end{cases} \tag{24}$$

This is not very different from the WCLR-C bootstrap score vector

$$\sum_{i=1}^{N_g} (y^*_{gi} - \tilde{\Lambda}_{gi})\boldsymbol{X}_{gi} = \begin{cases} \sum_{i=1}^{N_g} (y_{gi} - \tilde{\Lambda}_{gi})\boldsymbol{X}_{gi} \text{ with prob. } 1/2, \\ \sum_{i=1}^{N_g} (\tilde{\Lambda}_{gi} - y_{gi})\boldsymbol{X}_{gi} \text{ with prob. } 1/2. \end{cases} \tag{25}$$

Cluster Fixed Effects

Cluster Fixed Effects

Cluster fixed effects create important computational issues. Now

$$\Pr(y_{gi} = 1) = \Lambda\left(\boldsymbol{X}_{gi}\boldsymbol{\beta} + \sum_{h=1}^{G} \delta_h D_{gi}^h\right), \tag{26}$$

where the D_{gi}^h are cluster fixed-effect dummies. There is no constant term so there are $G + k - 1$ parameters to estimate.

Cluster Fixed Effects

Cluster fixed effects create important computational issues. Now

$$\Pr(y_{gi} = 1) = \Lambda\Big(\boldsymbol{X}_{gi}\boldsymbol{\beta} + \sum_{h=1}^{G} \delta_h D_{gi}^h\Big), \tag{26}$$

where the D_{gi}^h are cluster fixed-effect dummies. There is no constant term so there are $G + k - 1$ parameters to estimate.

- When cluster h is omitted, it is impossible to identify δ_h, because $D_{gi}^h = 0$ for all $g \neq h$.

Cluster Fixed Effects

Cluster fixed effects create important computational issues. Now

$$\Pr(y_{gi} = 1) = \Lambda\left(\boldsymbol{X}_{gi}\boldsymbol{\beta} + \sum_{h=1}^{G} \delta_h D_{gi}^h\right), \tag{26}$$

where the D_{gi}^h are cluster fixed-effect dummies. There is no constant term so there are $G + k - 1$ parameters to estimate.

- When cluster h is omitted, it is impossible to identify δ_h, because $D_{gi}^h = 0$ for all $g \neq h$.
- For a linear model, we could first partial out the fixed effects. But, because (26) is nonlinear, we cannot do that here.

Cluster Fixed Effects

Cluster fixed effects create important computational issues. Now

$$\Pr(y_{gi} = 1) = \Lambda\left(\boldsymbol{X}_{gi}\boldsymbol{\beta} + \sum_{h=1}^{G} \delta_h D_{gi}^h\right), \tag{26}$$

where the D_{gi}^h are cluster fixed-effect dummies. There is no constant term so there are $G + k - 1$ parameters to estimate.

- When cluster h is omitted, it is impossible to identify δ_h, because $D_{gi}^h = 0$ for all $g \neq h$.
- For a linear model, we could first partial out the fixed effects. But, because (26) is nonlinear, we cannot do that here.
- We can rely on a generalized inverse if the logit routine uses one.

Cluster Fixed Effects

Cluster fixed effects create important computational issues. Now

$$\Pr(y_{gi} = 1) = \Lambda\Bigl(\boldsymbol{X}_{gi}\boldsymbol{\beta} + \sum_{h=1}^{G} \delta_h D_{gi}^h\Bigr), \tag{26}$$

where the D_{gi}^h are cluster fixed-effect dummies. There is no constant term so there are $G + k - 1$ parameters to estimate.

- When cluster h is omitted, it is impossible to identify δ_h, because $D_{gi}^h = 0$ for all $g \neq h$.
- For a linear model, we could first partial out the fixed effects. But, because (26) is nonlinear, we cannot do that here.
- We can rely on a generalized inverse if the logit routine uses one.
- We can estimate a different model for each omitted cluster, each with just $k + G - 2$ coefficients, in order to obtain the $\hat{\boldsymbol{\beta}}^{(g)}$.

With cluster fixed effects, we can estimate slope coefficients and make inferences about them. But this is insufficient for inference about predicted probabilities and marginal effects.

With cluster fixed effects, we can estimate slope coefficients and make inferences about them. But this is insufficient for inference about predicted probabilities and marginal effects.

- For every cluster, the constant term is δ_g. We cannot estimate it when we omit cluster g, because it is only identified by the observations in that cluster.

With cluster fixed effects, we can estimate slope coefficients and make inferences about them. But this is insufficient for inference about predicted probabilities and marginal effects.

- For every cluster, the constant term is δ_g. We cannot estimate it when we omit cluster g, because it is only identified by the observations in that cluster.
- Without the variance of $\hat{\delta}_g$ and its covariances with the slope coefficients, we cannot obtain the standard error of $\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g$, which is needed for the standard error of $\Lambda(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g)$.

With cluster fixed effects, we can estimate slope coefficients and make inferences about them. But this is insufficient for inference about predicted probabilities and marginal effects.

- For every cluster, the constant term is δ_g. We cannot estimate it when we omit cluster g, because it is only identified by the observations in that cluster.
- Without the variance of $\hat{\delta}_g$ and its covariances with the slope coefficients, we cannot obtain the standard error of $\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g$, which is needed for the standard error of $\Lambda(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g)$.

We also need the full variance matrix in order to obtain the standard errors of the marginal effects.

With cluster fixed effects, we can estimate slope coefficients and make inferences about them. But this is insufficient for inference about predicted probabilities and marginal effects.

- For every cluster, the constant term is δ_g. We cannot estimate it when we omit cluster g, because it is only identified by the observations in that cluster.
- Without the variance of $\hat{\delta}_g$ and its covariances with the slope coefficients, we cannot obtain the standard error of $\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g$, which is needed for the standard error of $\Lambda(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g)$.

We also need the full variance matrix in order to obtain the standard errors of the marginal effects.

We could use CV_1, but the elements corresponding to the δ_g will be severely biased downwards, since each of the fixed-effect dummy variables is simply a treatment dummy for a single treated cluster.

With cluster fixed effects, we can estimate slope coefficients and make inferences about them. But this is insufficient for inference about predicted probabilities and marginal effects.

- For every cluster, the constant term is δ_g. We cannot estimate it when we omit cluster g, because it is only identified by the observations in that cluster.
- Without the variance of $\hat{\delta}_g$ and its covariances with the slope coefficients, we cannot obtain the standard error of $\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g$, which is needed for the standard error of $\Lambda(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}} + \hat{\delta}_g)$.

We also need the full variance matrix in order to obtain the standard errors of the marginal effects.

We could use CV_1, but the elements corresponding to the δ_g will be severely biased downwards, since each of the fixed-effect dummy variables is simply a treatment dummy for a single treated cluster.

Further work is needed!

Confidence Intervals

Confidence Intervals

A conventional confidence interval has the form

$$\big[\hat{\beta}_j - c_{1-\alpha/2}\,\mathrm{se}(\hat{\beta}_j),\;\; \hat{\beta}_j + c_{1-\alpha/2}\,\mathrm{se}(\hat{\beta}_j)\big], \tag{27}$$

usually with $c_{1-\alpha/2}$ a quantile of $t(G-1)$.

Confidence Intervals

A conventional confidence interval has the form

$$\big[\hat{\beta}_j - c_{1-\alpha/2}\,\mathrm{se}(\hat{\beta}_j),\;\; \hat{\beta}_j + c_{1-\alpha/2}\,\mathrm{se}(\hat{\beta}_j)\big], \tag{27}$$

usually with $c_{1-\alpha/2}$ a quantile of $t(G-1)$.

We can instead use bootstrap standard errors in (27). These are

$$\mathrm{se}_{\mathrm{boot}}(\hat{\beta}_j) = \left(\frac{1}{B-1}\sum_{b=1}^{B}\big(\hat{\beta}_j^{*b} - \bar{\beta}_j^{*}\big)^2\right)^{1/2}. \tag{28}$$

Confidence Intervals

A conventional confidence interval has the form

$$\big[\hat\beta_j - c_{1-\alpha/2}\,\mathrm{se}(\hat\beta_j),\;\; \hat\beta_j + c_{1-\alpha/2}\,\mathrm{se}(\hat\beta_j)\big], \tag{27}$$

usually with $c_{1-\alpha/2}$ a quantile of $t(G-1)$.

We can instead use bootstrap standard errors in (27). These are

$$\mathrm{se}_{\mathrm{boot}}(\hat\beta_j) = \left(\frac{1}{B-1}\sum_{b=1}^{B}\big(\hat\beta_j^{*b} - \bar\beta_j^{*}\big)^2\right)^{1/2}. \tag{28}$$

Alternatively, we can use the **studentized bootstrap interval**

$$\big[\hat\beta_j - c^{*}_{1-\alpha/2}\,\mathrm{se}_1(\hat\beta_j),\;\; \hat\beta_j - c^{*}_{\alpha/2}\,\mathrm{se}_1(\hat\beta_j)\big]. \tag{29}$$

These are both easy to construct using an unrestricted bootstrap DGP.

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C.

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

In theory, studentized bootstrap intervals should perform better than ones that use bootstrap standard errors.

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

In theory, studentized bootstrap intervals should perform better than ones that use bootstrap standard errors. Not always!

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

In theory, studentized bootstrap intervals should perform better than ones that use bootstrap standard errors. Not always!

- (29) is based on an asymptotically pivotal test statistic, and it allows the t-statistic to have an asymmetric distribution.

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

In theory, studentized bootstrap intervals should perform better than ones that use bootstrap standard errors. Not always!

- (29) is based on an asymptotically pivotal test statistic, and it allows the t-statistic to have an asymmetric distribution.
- (27) is not based on an asymptotically pivotal quantity, and it imposes symmetry on the distribution.

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

In theory, studentized bootstrap intervals should perform better than ones that use bootstrap standard errors. Not always!

- (29) is based on an asymptotically pivotal test statistic, and it allows the *t*-statistic to have an asymmetric distribution.
- (27) is not based on an asymptotically pivotal quantity, and it imposes symmetry on the distribution.

Why not invert a bootstrap test based on a restricted bootstrap DGP, such as the WCLR-S bootstrap?

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

In theory, studentized bootstrap intervals should perform better than ones that use bootstrap standard errors. Not always!

- (29) is based on an asymptotically pivotal test statistic, and it allows the t-statistic to have an asymmetric distribution.
- (27) is not based on an asymptotically pivotal quantity, and it imposes symmetry on the distribution.

Why not invert a bootstrap test based on a restricted bootstrap DGP, such as the WCLR-S bootstrap?

- The logit model has to be estimated many times, with β_j equal to each candidate value for the limits of the interval.

It may seem odd to use the CV_1 standard error in (29), but it is essential to use the same standard error as in the WCLU bootstrap itself.

It seems plausible that intervals based on WCLU-S should outperform ones based on WCLU-C. They do!

In theory, studentized bootstrap intervals should perform better than ones that use bootstrap standard errors. Not always!

- (29) is based on an asymptotically pivotal test statistic, and it allows the t-statistic to have an asymmetric distribution.
- (27) is not based on an asymptotically pivotal quantity, and it imposes symmetry on the distribution.

Why not invert a bootstrap test based on a restricted bootstrap DGP, such as the WCLR-S bootstrap?

- The logit model has to be estimated many times, with β_j equal to each candidate value for the limits of the interval.
- We sometimes encountered numerical problems, making it infeasible to perform simulation experiments.

Simulation Design

Simulation Design

There are $N = 500G$ observations, with G often 24 and $N = 12{,}000$.

$$\mathrm{E}(y_{gi}) = \Lambda\Bigl(\beta_1 + \sum_{j=2}^{k-1} \beta_j X_{gij} + \beta_k T_{gi}\Bigr). \tag{30}$$

Simulation Design

There are $N = 500G$ observations, with G often 24 and $N = 12{,}000$.

$$\mathrm{E}(y_{gi}) = \Lambda\Big(\beta_1 + \sum_{j=2}^{k-1} \beta_j X_{gij} + \beta_k T_{gi}\Big). \tag{30}$$

The X_{gij} are binary random variables which vary at the cluster level.

Simulation Design

There are $N = 500G$ observations, with G often 24 and $N = 12{,}000$.

$$\mathrm{E}(y_{gi}) = \Lambda\Big(\beta_1 + \sum_{j=2}^{k-1} \beta_j X_{gij} + \beta_k T_{gi}\Big). \tag{30}$$

The X_{gij} are binary random variables which vary at the cluster level.

T_{gi} is a treatment dummy, which equals 1 for G_1 out of G clusters. The hypothesis under test is $\beta_k = 0$.

Simulation Design

There are $N = 500G$ observations, with G often 24 and $N = 12{,}000$.

$$\mathrm{E}(y_{gi}) = \Lambda\Big(\beta_1 + \sum_{j=2}^{k-1} \beta_j X_{gij} + \beta_k T_{gi}\Big). \tag{30}$$

The X_{gij} are binary random variables which vary at the cluster level.

T_{gi} is a treatment dummy, which equals 1 for G_1 out of G clusters. The hypothesis under test is $\beta_k = 0$.

The unconditional expectation of y_{gi} is π, which depends on the regressors and parameters in (30). We change it by varying β_1.

Simulation Design

There are $N = 500G$ observations, with G often 24 and $N = 12{,}000$.

$$\mathrm{E}(y_{gi}) = \Lambda\Big(\beta_1 + \sum_{j=2}^{k-1} \beta_j X_{gij} + \beta_k T_{gi}\Big). \tag{30}$$

The X_{gij} are binary random variables which vary at the cluster level.

T_{gi} is a treatment dummy, which equals 1 for G_1 out of G clusters. The hypothesis under test is $\beta_k = 0$.

The unconditional expectation of y_{gi} is π, which depends on the regressors and parameters in (30). We change it by varying β_1.

Intra-cluster correlation is determined by a parameter ϕ, which is often set to 0.10 so that it is moderate.

Cluster sizes depend on a parameter γ as in MacKinnon and Webb (JAE 2017).

Cluster sizes depend on a parameter γ as in MacKinnon and Webb (JAE 2017).

The N observations are divided among the G clusters using the formula

$$N_g = \left\lfloor N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right\rfloor, \quad g = 1, \ldots, G-1, \tag{31}$$

The value of N_G is then set to $N - \sum_{g=1}^{G-1} N_g$.

Cluster sizes depend on a parameter γ as in MacKinnon and Webb (JAE 2017).

The N observations are divided among the G clusters using the formula

$$N_g = \left\lfloor N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right\rfloor, \quad g = 1, \ldots, G-1, \tag{31}$$

The value of N_G is then set to $N - \sum_{g=1}^{G-1} N_g$.

For $G = 24$, $N_g = 500$ for all g when $\gamma = 0$.

Cluster sizes depend on a parameter γ as in MacKinnon and Webb (JAE 2017).

The N observations are divided among the G clusters using the formula

$$N_g = \left\lfloor N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right\rfloor, \quad g = 1, \ldots, G-1, \tag{31}$$

The value of N_G is then set to $N - \sum_{g=1}^{G-1} N_g$.

For $G = 24$, $N_g = 500$ for all g when $\gamma = 0$.

For $G = 24$, the N_g vary from 163 to 1120 when $\gamma = 2$.

Cluster sizes depend on a parameter γ as in MacKinnon and Webb (JAE 2017).

The N observations are divided among the G clusters using the formula

$$N_g = \left\lfloor N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right\rfloor, \quad g = 1, \ldots, G-1, \tag{31}$$

The value of N_G is then set to $N - \sum_{g=1}^{G-1} N_g$.

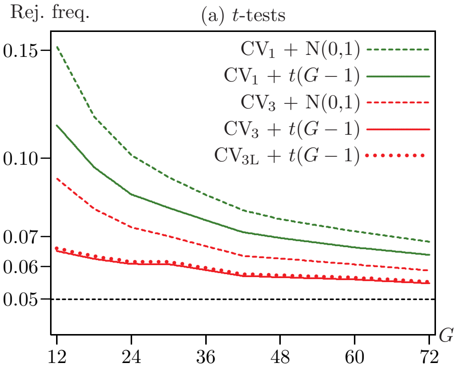
For $G = 24$, $N_g = 500$ for all g when $\gamma = 0$.

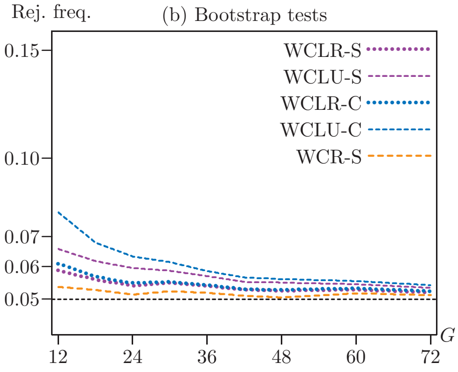
For $G = 24$, the N_g vary from 163 to 1120 when $\gamma = 2$.

For $G = 24$, the N_g vary from 40 to 1889 when $\gamma = 4$.

Simulation Results

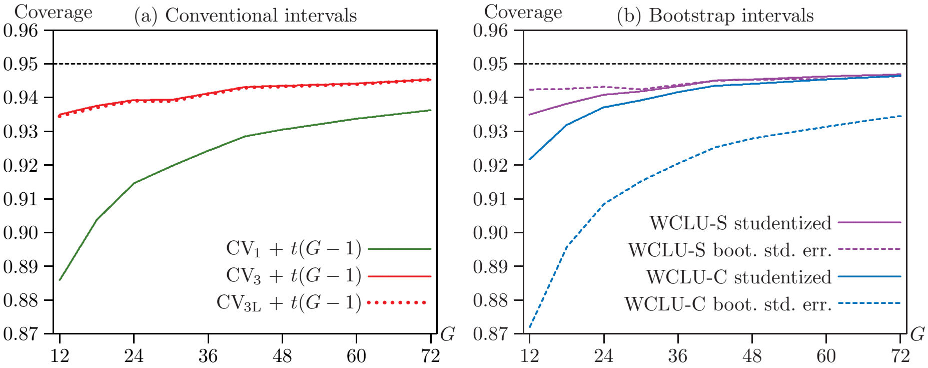
Figure 1. Rejection frequencies for tests at the .05 level as functions of G





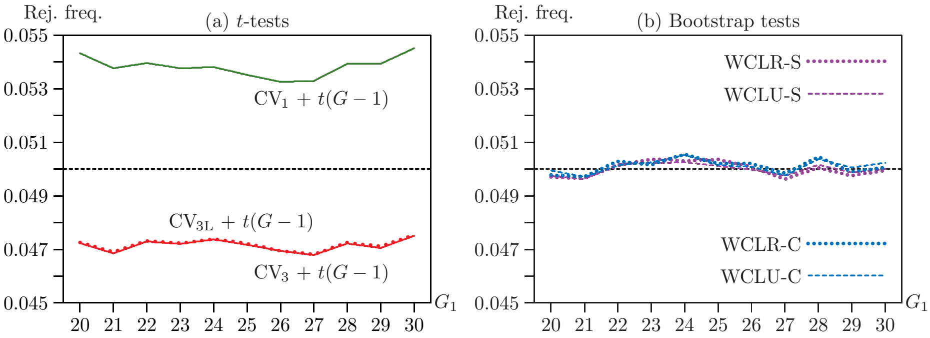
$N = 500G$, $G_1 = G/3$, $k = 7$, $\gamma = 2$, $\phi = 0.10$, $\pi = 0.31$, $B = 999$
100,000 replications

Figure 2. Coverage for 95% confidence intervals as functions of G



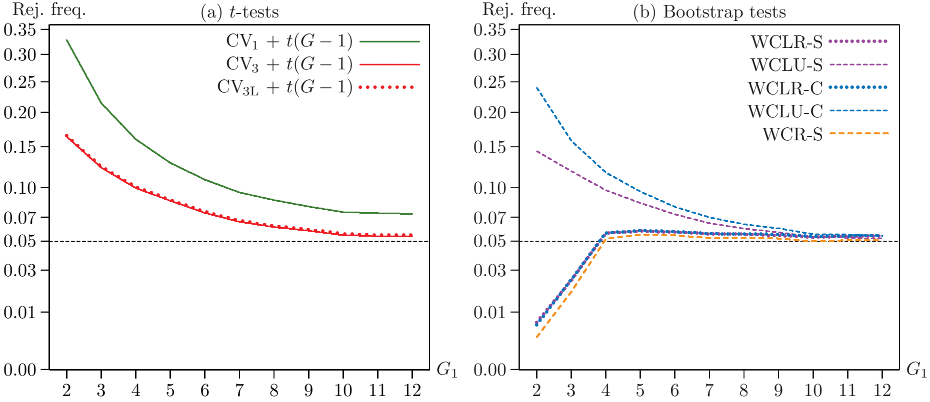
$N = 500G$, $G_1 = G/3$, $k = 7$, $\gamma = 2$, $\phi = 0.10$, $\pi = 0.31$, $B = 999$
100,000 replications

Figure 3. Rejection frequencies for .05-level tests in an almost ideal case



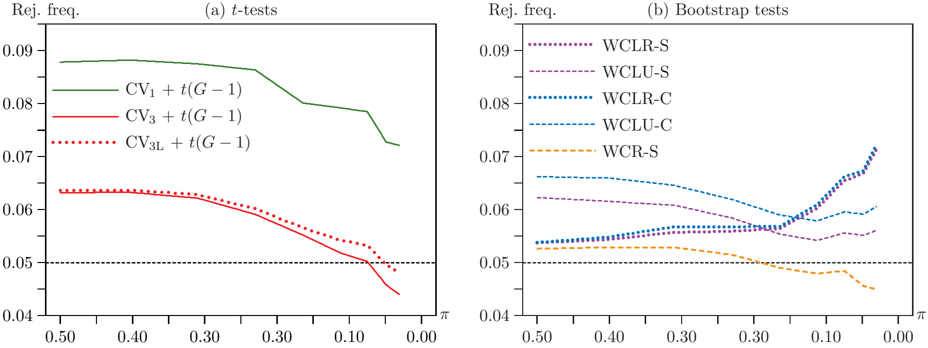
$G = 50$, $N = 25{,}000$, $k = 7$, $\gamma = 0$, $\phi = 0$, $\pi = 0.5$, $B = 999$
400,000 replications.

Figure 4. Rejection frequencies for .05-level tests as functions of G_1



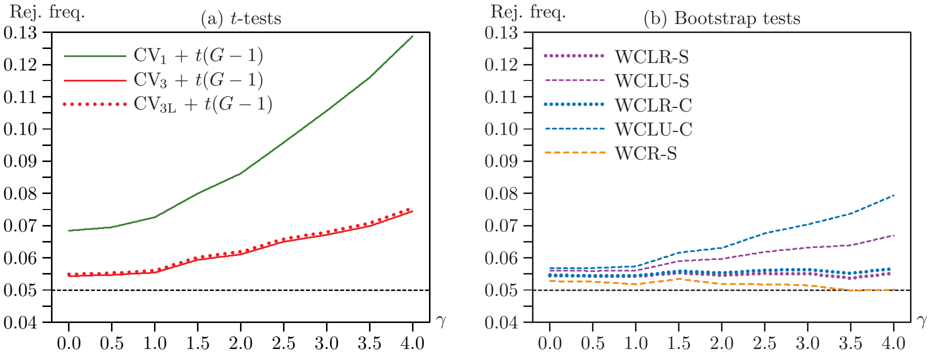
$N = 12{,}000$, $G = 24$, $k = 7$, $\gamma = 2$, $\phi = 0.10$, $\pi = 0.31$, $B = 999$
100,000 replications

Figure 5. Rejection frequencies for tests at the .05 level as functions of π



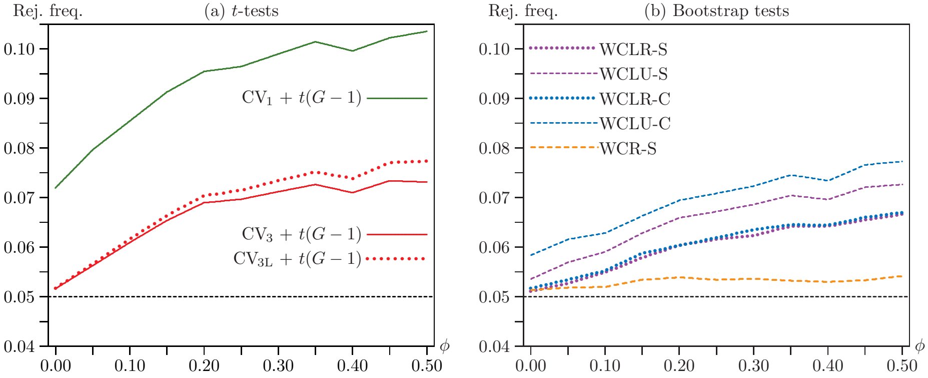
$N = 12{,}000$, $G = 24$, $G_1 = 8$, $k = 7$, $\gamma = 2$, $\phi = 0.10$, $B = 999$
100,000 replications

Figure 6. Rejection frequencies for tests at the .05 level as functions of γ



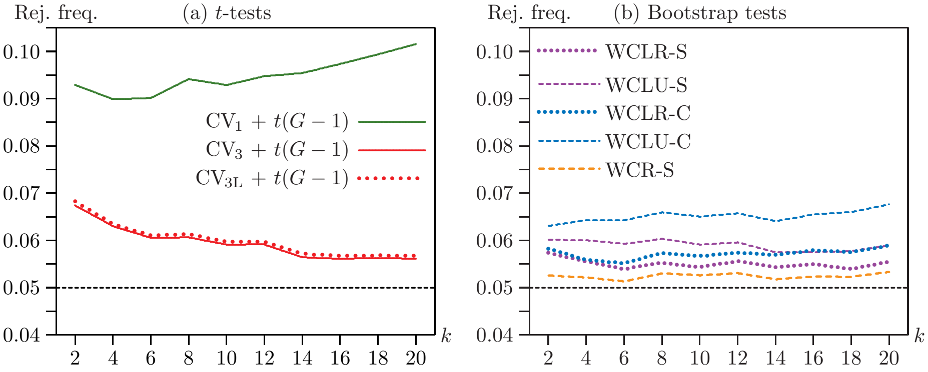
$N = 12{,}000$, $G = 24$, $G_1 = 8$, $k = 7$, $\phi = 0.10$, $\pi = 0.31$, $B = 999$
100,000 replications

Figure 7. Rejection frequencies for tests at the .05 level as functions of ϕ



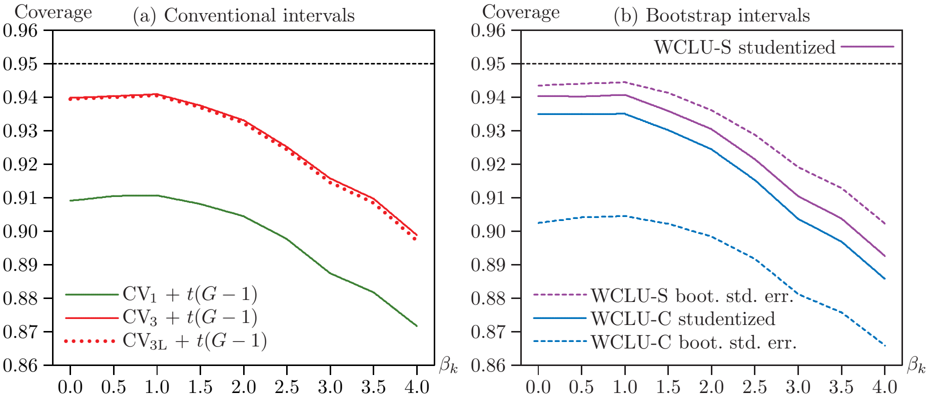
$N = 12{,}000$, $G = 24$, $G_1 = 8$, $k = 7$, $\pi = 0.31$, $B = 999$
100,000 replications

Figure 8. Rejection frequencies for tests at the .05 level as functions of k



$N = 12{,}000$, $G = 24$, $G_1 = 8$, $\phi = 0.10$, $\pi = 0.31$, $B = 999$
100,000 replications

Figure 9. Coverage for 95% confidence intervals as functions of β_k.



$N = 12{,}000$, $G = 24$, $G_1 = 8$, $\phi = 0.10$, $\pi = 0.31$, $B = 999$
100,000 replications

Conclusions from Simulations

Conclusions from Simulations

- Conventional t-tests based on CV_1 and $t(G-1)$ generally over-reject, often severely. CV_3 t-tests perform better.

Conclusions from Simulations

- Conventional t-tests based on CV_1 and $t(G-1)$ generally over-reject, often severely. CV_3 t-tests perform better.
- CV_3 t-tests can either under-reject or over-reject, the latter especially when G_1/G is small, π is far from 0.5, or ϕ is high.

Conclusions from Simulations

- Conventional t-tests based on CV_1 and $t(G-1)$ generally over-reject, often severely. CV_3 t-tests perform better.
- CV_3 t-tests can either under-reject or over-reject, the latter especially when G_1/G is small, π is far from 0.5, or ϕ is high.
- Linearized cluster jackknife, or $\mathrm{CV}_{3\mathrm{L}}$, standard errors are much cheaper to compute than CV_3 ones, and usually very similar.

Conclusions from Simulations

- Conventional t-tests based on CV_1 and $t(G-1)$ generally over-reject, often severely. CV_3 t-tests perform better.
- CV_3 t-tests can either under-reject or over-reject, the latter especially when G_1/G is small, π is far from 0.5, or ϕ is high.
- Linearized cluster jackknife, or $\text{CV}_{3\text{L}}$, standard errors are much cheaper to compute than CV_3 ones, and usually very similar.
- The WCLR-S bootstrap often performs well. Problems can arise when π is extreme or there is a lot of intra-cluster correlation.

Conclusions from Simulations

- Conventional t-tests based on CV_1 and $t(G-1)$ generally over-reject, often severely. CV_3 t-tests perform better.
- CV_3 t-tests can either under-reject or over-reject, the latter especially when G_1/G is small, π is far from 0.5, or ϕ is high.
- Linearized cluster jackknife, or $\text{CV}_{3\text{L}}$, standard errors are much cheaper to compute than CV_3 ones, and usually very similar.
- The WCLR-S bootstrap often performs well. Problems can arise when π is extreme or there is a lot of intra-cluster correlation.
- All methods can be somewhat unreliable when the binary outcomes are unbalanced, with most equal to either 0 or 1.

Conclusions from Simulations

- Conventional t-tests based on CV_1 and $t(G-1)$ generally over-reject, often severely. CV_3 t-tests perform better.
- CV_3 t-tests can either under-reject or over-reject, the latter especially when G_1/G is small, π is far from 0.5, or ϕ is high.
- Linearized cluster jackknife, or $\text{CV}_{3\text{L}}$, standard errors are much cheaper to compute than CV_3 ones, and usually very similar.
- The WCLR-S bootstrap often performs well. Problems can arise when π is extreme or there is a lot of intra-cluster correlation.
- All methods can be somewhat unreliable when the binary outcomes are unbalanced, with most equal to either 0 or 1.
- WCLU-S often performs much better than WCLU-C, and WCLR-S generally performs even better.

Conclusions from Simulations

- Conventional t-tests based on CV_1 and $t(G-1)$ generally over-reject, often severely. CV_3 t-tests perform better.
- CV_3 t-tests can either under-reject or over-reject, the latter especially when G_1/G is small, π is far from 0.5, or ϕ is high.
- Linearized cluster jackknife, or $\mathrm{CV}_{3\mathrm{L}}$, standard errors are much cheaper to compute than CV_3 ones, and usually very similar.
- The WCLR-S bootstrap often performs well. Problems can arise when π is extreme or there is a lot of intra-cluster correlation.
- All methods can be somewhat unreliable when the binary outcomes are unbalanced, with most equal to either 0 or 1.
- WCLU-S often performs much better than WCLU-C, and WCLR-S generally performs even better.
- Bootstrap standard errors should always be based on the WCLU-S bootstrap, and these can lead to good confidence intervals.

Empirical Example 1 – Angrist and Lavy

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination?

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

- The mean of the dependent variable is 0.287.

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

- The mean of the dependent variable is 0.287.
- There are 10 regressors plus a constant term.

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

- The mean of the dependent variable is 0.287.
- There are 10 regressors plus a constant term.
- Cluster sizes vary from 12 to 146, and partial leverage varies a lot.

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

- The mean of the dependent variable is 0.287.
- There are 10 regressors plus a constant term.
- Cluster sizes vary from 12 to 146, and partial leverage varies a lot.
- We compute 20 P values. The bootstrap ones with asterisks are based on t-statistics using bootstrap standard errors.

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

- The mean of the dependent variable is 0.287.
- There are 10 regressors plus a constant term.
- Cluster sizes vary from 12 to 146, and partial leverage varies a lot.
- We compute 20 P values. The bootstrap ones with asterisks are based on t-statistics using bootstrap standard errors.
- Most P values are less than 0.05. For the logit model, they vary between 0.0264 (WCLU-C*) and 0.0578 (WCLU-S*).

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

- The mean of the dependent variable is 0.287.
- There are 10 regressors plus a constant term.
- Cluster sizes vary from 12 to 146, and partial leverage varies a lot.
- We compute 20 P values. The bootstrap ones with asterisks are based on t-statistics using bootstrap standard errors.
- Most P values are less than 0.05. For the logit model, they vary between 0.0264 (WCLU-C*) and 0.0578 (WCLU-S*).
- Bootstrap P values use the Rademacher distribution.

Empirical Example 1 – Angrist and Lavy

The first example is based on Angrist and Lavy (2009) and concerns cash incentives for high-school students in Israel. Do they increase the chance of passing a high-stakes examination? Maybe!

There were 1861 students in 34 schools, of which 16 were treated.

- The mean of the dependent variable is 0.287.
- There are 10 regressors plus a constant term.
- Cluster sizes vary from 12 to 146, and partial leverage varies a lot.
- We compute 20 P values. The bootstrap ones with asterisks are based on t-statistics using bootstrap standard errors.
- Most P values are less than 0.05. For the logit model, they vary between 0.0264 (WCLU-C*) and 0.0578 (WCLU-S*).
- Bootstrap P values use the Rademacher distribution.
- They are based on 9,999,999 bootstrap samples to ensure that the random number generator plays almost no role.

We also perform **placebo regressions**. For each of 400,000 replications, we add one additional regressor to the original model and test the hypothesis that its coefficient equals 0.

We also perform **placebo regressions**. For each of 400,000 replications, we add one additional regressor to the original model and test the hypothesis that its coefficient equals 0.

The placebo regressor equals 1 for 16 randomly chosen schools and 0 for the remaining 18 schools.

We also perform **placebo regressions**. For each of 400,000 replications, we add one additional regressor to the original model and test the hypothesis that its coefficient equals 0.

The placebo regressor equals 1 for 16 randomly chosen schools and 0 for the remaining 18 schools.

There are ${}_{34}C_{16} = 2{,}203{,}961{,}430$ ways to choose the placebo regressor, so we sample with replacement because it is easier.

We also perform **placebo regressions**. For each of 400,000 replications, we add one additional regressor to the original model and test the hypothesis that its coefficient equals 0.

The placebo regressor equals 1 for 16 randomly chosen schools and 0 for the remaining 18 schools.

There are ${}_{34}C_{16} = 2{,}203{,}961{,}430$ ways to choose the placebo regressor, so we sample with replacement because it is easier.

Note that we do not omit the original treatment regressor. Doing that would increase rejection frequencies for the placebo regressions.

We also perform **placebo regressions**. For each of 400,000 replications, we add one additional regressor to the original model and test the hypothesis that its coefficient equals 0.

The placebo regressor equals 1 for 16 randomly chosen schools and 0 for the remaining 18 schools.

There are ${}_{34}C_{16} = 2{,}203{,}961{,}430$ ways to choose the placebo regressor, so we sample with replacement because it is easier.

Note that we do not omit the original treatment regressor. Doing that would increase rejection frequencies for the placebo regressions.

- Rejection frequencies for placebo regressions with 400,000 replications vary from 0.0364 (WCLU-S*) to 0.0836 (WCLU-C*).

We also perform **placebo regressions**. For each of 400,000 replications, we add one additional regressor to the original model and test the hypothesis that its coefficient equals 0.

The placebo regressor equals 1 for 16 randomly chosen schools and 0 for the remaining 18 schools.

There are ${}_{34}C_{16} = 2{,}203{,}961{,}430$ ways to choose the placebo regressor, so we sample with replacement because it is easier.

Note that we do not omit the original treatment regressor. Doing that would increase rejection frequencies for the placebo regressions.

- Rejection frequencies for placebo regressions with 400,000 replications vary from 0.0364 (WCLU-S*) to 0.0836 (WCLU-C*).
- Other methods that reject less than 4% of the time are CV_3 (0.0373) and $\mathrm{CV}_{3\mathrm{L}}$ (0.0387).

We also perform **placebo regressions**. For each of 400,000 replications, we add one additional regressor to the original model and test the hypothesis that its coefficient equals 0.

The placebo regressor equals 1 for 16 randomly chosen schools and 0 for the remaining 18 schools.

There are ${}_{34}\mathrm{C}_{16} = 2{,}203{,}961{,}430$ ways to choose the placebo regressor, so we sample with replacement because it is easier.

Note that we do not omit the original treatment regressor. Doing that would increase rejection frequencies for the placebo regressions.

- Rejection frequencies for placebo regressions with 400,000 replications vary from 0.0364 (WCLU-S*) to 0.0836 (WCLU-C*).
- Other methods that reject less than 4% of the time are CV_3 (0.0373) and $\mathrm{CV}_{3\mathrm{L}}$ (0.0387).
- Reassuringly, the methods that over-reject most significantly are the ones that yield the smallest P values for the actual dataset.

Table 1: Effects of Cash Incentives on Passing the Bagrut

Model	Method	Coef.	Std. error	t stat.	P value	Placebo
Logit	CV_1	0.7164	0.3149	2.2746	0.0296	0.0794
Logit	CV_{2L}	0.7164	0.3303	2.1687	0.0374	0.0607
Logit	CV_3	0.7164	0.3609	1.9850	0.0555	0.0373
Logit	CV_{3L}	0.7164	0.3592	1.9941	0.0545	0.0387
Logit	WCLR-C	0.7164		2.2746	0.0523	0.0464
Logit	WCLR-S	0.7164		2.2746	0.0564	0.0426
Logit	WCLU-C	0.7164		2.2746	0.0457	0.0529
Logit	WCLU-C*	0.7164	0.3095	2.3142	0.0264	0.0846
Logit	WCLU-S	0.7164		2.2476	0.0487	0.0476
Logit	WCLU-S*	0.7164	0.3645	1.9655	0.0578	0.0364
LPM	CV_1	0.1047	0.0444	2.3572	0.0245	0.0866
LPM	CV_2	0.1047	0.0466	2.2483	0.0314	0.0681
LPM	CV_3	0.1047	0.0506	2.0695	0.0464	0.0454
LPM	WCR-C	0.1047		2.3572	0.0393	0.0530
LPM	WCR-S	0.1047		2.3572	0.0418	0.0497

Empirical Example 2 – Tuition Fees

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada?

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada? Probably not! P values vary greatly across methods.

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada? Probably not! P values vary greatly across methods.

We use Labour Force Survey data for 2009–2019 for males aged 20 and 21 who reside in one of the ten provinces.

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada? Probably not! P values vary greatly across methods.

We use Labour Force Survey data for 2009–2019 for males aged 20 and 21 who reside in one of the ten provinces.

We do not report results for females, because even the least reliable methods provide no evidence that tuition fees matter.

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada? Probably not! P values vary greatly across methods.

We use Labour Force Survey data for 2009–2019 for males aged 20 and 21 who reside in one of the ten provinces.

We do not report results for females, because even the least reliable methods provide no evidence that tuition fees matter.

The sample excludes immigrants in Canada for less than 10 years, because they may pay higher fees.

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada? Probably not! P values vary greatly across methods.

We use Labour Force Survey data for 2009–2019 for males aged 20 and 21 who reside in one of the ten provinces.

We do not report results for females, because even the least reliable methods provide no evidence that tuition fees matter.

The sample excludes immigrants in Canada for less than 10 years, because they may pay higher fees.

- There are 127,518 observations.

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada? Probably not! P values vary greatly across methods.

We use Labour Force Survey data for 2009–2019 for males aged 20 and 21 who reside in one of the ten provinces.

We do not report results for females, because even the least reliable methods provide no evidence that tuition fees matter.

The sample excludes immigrants in Canada for less than 10 years, because they may pay higher fees.

- There are 127,518 observations.
- The ten clusters vary in size from 3,402 (P.E.I.) to 37,109 (Ontario).

Empirical Example 2 – Tuition Fees

The second example concerns university tuition fees in Canada, which vary by province and year.

Do tuition fees affect the probability of university attendance in Canada? Probably not! P values vary greatly across methods.

We use Labour Force Survey data for 2009–2019 for males aged 20 and 21 who reside in one of the ten provinces.

We do not report results for females, because even the least reliable methods provide no evidence that tuition fees matter.

The sample excludes immigrants in Canada for less than 10 years, because they may pay higher fees.

- There are 127,518 observations.
- The ten clusters vary in size from 3,402 (P.E.I.) to 37,109 (Ontario).
- The mean of the dependent variable is 0.4208.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.
- Bootstrap methods use the six-point distribution of Webb (CJE,2023) instead of Rademacher, with 9,999,999 replications.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.
- Bootstrap methods use the six-point distribution of Webb (CJE,2023) instead of Rademacher, with 9,999,999 replications.
- Computing CV_3 was far more expensive than anything else. It cost about 41 times as much as $\mathrm{CV}_{3\mathrm{L}}$, and results were almost identical.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.
- Bootstrap methods use the six-point distribution of Webb (CJE,2023) instead of Rademacher, with 9,999,999 replications.
- Computing CV_3 was far more expensive than anything else. It cost about 41 times as much as $\mathrm{CV}_{3\mathrm{L}}$, and results were almost identical.

Once again, we perform a placebo regression experiment.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.
- Bootstrap methods use the six-point distribution of Webb (CJE,2023) instead of Rademacher, with 9,999,999 replications.
- Computing CV_3 was far more expensive than anything else. It cost about 41 times as much as $\mathrm{CV}_{3\mathrm{L}}$, and results were almost identical.

Once again, we perform a placebo regression experiment.

We generate artificial tuition series by using an AR(1) model, simulated separately for each province.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.
- Bootstrap methods use the six-point distribution of Webb (CJE,2023) instead of Rademacher, with 9,999,999 replications.
- Computing CV_3 was far more expensive than anything else. It cost about 41 times as much as $\mathrm{CV}_{3\mathrm{L}}$, and results were almost identical.

Once again, we perform a placebo regression experiment.

We generate artificial tuition series by using an AR(1) model, simulated separately for each province.

The placebo regressions use 400,000 replications, with $B = 999$.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.
- Bootstrap methods use the six-point distribution of Webb (CJE,2023) instead of Rademacher, with 9,999,999 replications.
- Computing CV_3 was far more expensive than anything else. It cost about 41 times as much as $\mathrm{CV}_{3\mathrm{L}}$, and results were almost identical.

Once again, we perform a placebo regression experiment.

We generate artificial tuition series by using an AR(1) model, simulated separately for each province.

The placebo regressions use 400,000 replications, with $B = 999$.

As before, we include both the actual tuition series and the simulated one in the placebo regessions.

- There are 4 ordinary regressors plus 20 dummies for year and province fixed effects.
- Bootstrap methods use the six-point distribution of Webb (CJE,2023) instead of Rademacher, with 9,999,999 replications.
- Computing CV_3 was far more expensive than anything else. It cost about 41 times as much as $\mathrm{CV}_{3\mathrm{L}}$, and results were almost identical.

Once again, we perform a placebo regression experiment.

We generate artificial tuition series by using an AR(1) model, simulated separately for each province.

The placebo regressions use 400,000 replications, with $B = 999$.

As before, we include both the actual tuition series and the simulated one in the placebo regessions.

The only parameter that seems to matter is the autoregressive coefficient. Reported results are for the random walk case. With smaller values, rejection frequencies were a bit higher.

- Placebo rejection frequencies vary between 0.0485 (WCLU-S*) and 0.1502 (WCU-C*).

- Placebo rejection frequencies vary between 0.0485 (WCLU-S*) and 0.1502 (WCU-C*).
- The rejection frequency for WCLR-S is 0.0527; for CV_{3L}, it is 0.0575. These are methods that might be expected to work well.

- Placebo rejection frequencies vary between 0.0485 (WCLU-S*) and 0.1502 (WCU-C*).
- The rejection frequency for WCLR-S is 0.0527; for $\mathrm{CV}_{3\mathrm{L}}$, it is 0.0575. These are methods that might be expected to work well.
- There is a strong, inverse relationship between the placebo rejection frequencies and the reported P values.

- Placebo rejection frequencies vary between 0.0485 (WCLU-S*) and 0.1502 (WCU-C*).
- The rejection frequency for WCLR-S is 0.0527; for $\mathrm{CV}_{3\mathrm{L}}$, it is 0.0575. These are methods that might be expected to work well.
- There is a strong, inverse relationship between the placebo rejection frequencies and the reported P values.
- All the methods with P values less than 0.05 over-reject approximately 10–15% of the time in the placebo regressions.

- Placebo rejection frequencies vary between 0.0485 (WCLU-S*) and 0.1502 (WCU-C*).
- The rejection frequency for WCLR-S is 0.0527; for $\text{CV}_{3\text{L}}$, it is 0.0575. These are methods that might be expected to work well.
- There is a strong, inverse relationship between the placebo rejection frequencies and the reported P values.
- All the methods with P values less than 0.05 over-reject approximately 10–15% of the time in the placebo regressions.
- Methods that perform reasonably well in the placebo regressions all yield P values greater than 0.13.

- Placebo rejection frequencies vary between 0.0485 (WCLU-S*) and 0.1502 (WCU-C*).
- The rejection frequency for WCLR-S is 0.0527; for $\mathrm{CV}_{3\mathrm{L}}$, it is 0.0575. These are methods that might be expected to work well.
- There is a strong, inverse relationship between the placebo rejection frequencies and the reported P values.
- All the methods with P values less than 0.05 over-reject approximately 10–15% of the time in the placebo regressions.
- Methods that perform reasonably well in the placebo regressions all yield P values greater than 0.13.

Once again, there seems to be substantial agreement between the placebo regressions, which use real data, and our simulation experiments, which do not.

Table 2: Effects of Tuition Fees on University Attendance

Model	Method	Coef.	Std. error	t stat.	P value	Placebo
Logit	CV_1	−0.1302	0.0469	−2.7745	0.0216	0.1298
Logit	CV_3	−0.1302	0.0799	−1.6301	0.1375	0.0574
Logit	CV_{3L}	−0.1302	0.0800	−1.6280	0.1380	0.0575
Logit	WCLR-C	−0.1302		−2.7745	0.1399	0.0639
Logit	WCLR-S	−0.1302		−2.7745	0.1551	0.0527
Logit	WCLU-C	−0.1302		−2.7745	0.0210	0.0993
Logit	WCLU-S	−0.1302		−2.7745	0.0912	0.0724
Logit	WCLU-C*	−0.1302	0.0445	−2.9244	0.0169	0.1464
Logit	WCLU-S*	−0.1302	0.0843	−1.5442	0.1569	0.0485
LPM	CV_1	−0.0296	0.0106	−2.7899	0.0211	0.1332
LPM	CV_3	−0.0296	0.0184	−1.6120	0.1414	0.0601
LPM	WCR-C	−0.0296		−2.7899	0.1414	0.0658
LPM	WCR-S	−0.0296		−2.7899	0.1534	0.0548
LPM	WCU-S*	−0.0296	0.0194	1.5290	0.1606	0.0508

Conclusions

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.
- Cluster jackknife, or CV_3, t-tests reject less often than CV_1 t-tests.

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.
- Cluster jackknife, or CV_3, t-tests reject less often than CV_1 t-tests.
- $\mathrm{CV}_{3\mathrm{L}}$ t-tests usually yield results close to CV_3 t-tests and are very much cheaper to compute.

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.
- Cluster jackknife, or CV_3, t-tests reject less often than CV_1 t-tests.
- $\mathrm{CV}_{3\mathrm{L}}$ t-tests usually yield results close to CV_3 t-tests and are very much cheaper to compute.
- The best test is often WCLR-S. It frequently outperforms WCLR-C, but often not by much.

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.
- Cluster jackknife, or CV_3, t-tests reject less often than CV_1 t-tests.
- CV_{3L} t-tests usually yield results close to CV_3 t-tests and are very much cheaper to compute.
- The best test is often WCLR-S. It frequently outperforms WCLR-C, but often not by much.
- WCLU-S almost always outperforms WCLU-C, often by a lot.

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.
- Cluster jackknife, or CV_3, t-tests reject less often than CV_1 t-tests.
- $\text{CV}_{3\text{L}}$ t-tests usually yield results close to CV_3 t-tests and are very much cheaper to compute.
- The best test is often WCLR-S. It frequently outperforms WCLR-C, but often not by much.
- WCLU-S almost always outperforms WCLU-C, often by a lot.
- Strange things can happen when the fraction of 1s (or 0s) is small and/or when there is a lot of intra-cluster correlation.

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.
- Cluster jackknife, or CV_3, t-tests reject less often than CV_1 t-tests.
- $\mathrm{CV}_{3\mathrm{L}}$ t-tests usually yield results close to CV_3 t-tests and are very much cheaper to compute.
- The best test is often WCLR-S. It frequently outperforms WCLR-C, but often not by much.
- WCLU-S almost always outperforms WCLU-C, often by a lot.
- Strange things can happen when the fraction of 1s (or 0s) is small and/or when there is a lot of intra-cluster correlation.
- When CV_3, $\mathrm{CV}_{3\mathrm{L}}$, WCLR-S, and WCLU-S yield similar results, they can probably be believed.

Conclusions

- Conventional t-tests based on CV_1 should never be used. They always over-reject, even more so if based on N(0,1) critical values.
- Cluster jackknife, or CV_3, t-tests reject less often than CV_1 t-tests.
- $\mathrm{CV}_{3\mathrm{L}}$ t-tests usually yield results close to CV_3 t-tests and are very much cheaper to compute.
- The best test is often WCLR-S. It frequently outperforms WCLR-C, but often not by much.
- WCLU-S almost always outperforms WCLU-C, often by a lot.
- Strange things can happen when the fraction of 1s (or 0s) is small and/or when there is a lot of intra-cluster correlation.
- When CV_3, $\mathrm{CV}_{3\mathrm{L}}$, WCLR-S, and WCLU-S yield similar results, they can probably be believed.
- Use placebo regressions to see which tests are reliable.