Queen's Economics Department Working Paper No. 1515

Cluster-Robust Jackknife and Bootstrap Inference for Binary Response Models

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Aarhus University

Matthew D. Webb
Carleton University

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

5-2024

Cluster-Robust Jackknife and Bootstrap Inference for Binary Response Models*

James G. MacKinnon†
Queen's University
mackinno@queensu.ca

Morten Ørregaard Nielsen
Aarhus University
mon@econ.au.dk

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

May 30, 2024

Abstract

We study cluster-robust inference for binary response models. Inference based on the most commonly-used cluster-robust variance matrix estimator (CRVE) can be very unreliable. We study several alternatives. Conceptually the simplest of these, but also the most computationally demanding, involves jackknifing at the cluster level. We also propose a linearized version of the cluster-jackknife variance matrix estimator as well as linearized versions of the wild cluster bootstrap. The linearizations are based on empirical scores and are computationally efficient. Throughout we use the logit model as a leading example. We also discuss a new `Stata` software package called `logitjack` which implements these procedures. Simulation results strongly favor the new methods, and two empirical examples suggest that it can be important to use them in practice.

Keywords: logit model, logistic regression, clustered data, grouped data, cluster-robust variance estimator, CRVE, cluster jackknife, robust inference, wild cluster bootstrap, linearization

JEL Codes: C12, C15, C21, C23.

*MacKinnon and Webb thank the Social Sciences and Humanities Research Council of Canada (SSHRC grant 435-2021-0396) for financial support. Nielsen thanks the Danish National Research Foundation for financial support (DNRF Chair grant number DNRF154). We thank participants at the CEA Annual Meeting, the Canadian Econometrics Study Group, the University of Victoria, the Conference to Celebrate Professor M. H. Pesaran's Achievements at U.S.C., and New York Camp Econometrics. Code and data files may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/logitjack/

†Corresponding author. Address: Department of Economics, 94 University Avenue, Queen's University, Kingston, Ontario K7L 3N6, Canada. Email: mackinno@queensu.ca. Tel. 613-533-2293. Fax 613-533-6668.

1 Introduction

Cluster-robust inference has been studied extensively over the past decade. A recent guide to this literature is MacKinnon, Nielsen, and Webb (2023a). Other surveys include Cameron and Miller (2015), MacKinnon (2019), Esarey and Menger (2019), and MacKinnon and Webb (2020). Conley, Gonçalves, and Hansen (2018) surveys a broader class of methods for various types of dependent data. Although the literature has grown enormously, a very large fraction of it concerns linear regression models estimated by ordinary least squares. With the important exception of Hansen and Lee (2019), it has largely ignored nonlinear models. For linear regression models, several different cluster-robust variance matrix estimators (CRVEs) are available, along with a number of bootstrap methods. The finite-sample properties of these methods can vary greatly, and quite a lot is known about most of them. However, there exist almost no comparable results for nonlinear models.

To study the finite-sample properties of methods for cluster-robust inference for nonlinear models, it is essential to specify a particular class of such models. It seems natural to start with binary response models because they are widely used with the sort of cross-section and panel datasets where cluster-robust inference is often needed. As a leading example, we focus on the logit (or logistic regression) model.

As we show in Section 6, the only existing CRVE for logit models that is widely used can have poor finite-sample properties. We therefore propose several alternative procedures based on the cluster jackknife or the wild cluster bootstrap. The first cluster-jackknife procedures that we introduce are similar to the ones for linear models discussed in MacKinnon, Nielsen, and Webb (2023c, b) and Hansen (2023), but they are more challenging computationally because nonlinear estimation is needed. We therefore introduce computationally simpler procedures based on score vectors at the cluster level. These procedures, which appear to be new, involve linearizing the first-order conditions so as to compute approximations to the delete-one-cluster estimates needed for the jackknife. The linearized cluster jackknife estimators appear to be feasible for large samples with either few large clusters or many small ones.

The same linearization methods make it possible to apply what is essentially the wild cluster bootstrap (Cameron, Gelbach, and Miller 2008; Djogbenou, MacKinnon, and Nielsen 2019) to binary response models. We propose several new wild bootstrap methods which can be computed using almost the same code as similar wild cluster bootstrap methods for OLS regression. The methods that seem to work best in many cases are very similar to the WCR-S and WCU-S bootstraps proposed in MacKinnon et al. (2023b); see Section 3.

In Section 2, we introduce the class of binary response models to which our procedures apply, along with sandwich CRVEs for models with G clusters. These are special cases of the

conventional CRVEs discussed in Hansen and Lee (2019). We also discuss two CRVEs based on the cluster jackknife, in which each cluster in turn is deleted from the sample so as to obtain G vectors of parameter estimates. Although the cluster jackknife is not new, it does not seem to have been studied in this context. Then, in Section 3, we discuss a linearization procedure and show how it can be used as the key part of computationally efficient jackknife and wild bootstrap procedures, which appear to be new.

In Section 4, we discuss how to deal with cluster fixed effects. These are very commonly encountered in models with clustered data, and all the jackknife methods need to be modified to handle them. The paper mainly focuses on hypothesis tests, but Section 5 discusses confidence intervals, where computational issues are important. Section 6 presents the results of a large number of simulation experiments. Section 7 discusses two empirical examples which illustrate the application of our proposed methods. Finally, Section 8 concludes.

2 Sandwich CRVEs for Binary Response Models

We are concerned with binary response models of the form

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

Here y_{gi}, which equals either 0 or 1, is the response for observation i in cluster g. There are N observations, with the g^{th} cluster containing N_g of them. The continuous, monotonically increasing function $F(\cdot)$ maps from the real line to the 0-1 interval. The best-known examples are the logistic function and the cumulative normal distribution function. In the former case, (1) leads to the logit model, and in the latter case to the probit model. The first derivative of $F(\cdot)$ is denoted $f(\cdot)$. The row vector $\boldsymbol{X}_{gi}$ contains the values of k explanatory variables, and the k-vector $\boldsymbol{\beta}$ is to be estimated. In many cases, one element of $\boldsymbol{\beta}$ is of particular interest, and we wish to test a hypothesis about it or form a confidence interval. Without loss of generality, we assume that this is the k^{th} element. Then $\boldsymbol{\beta}$ can be divided into a $(k-1)$-vector $\boldsymbol{\beta}_1$ and a scalar β_k.

As specified in (1), the binary response model may or may not involve any intra-cluster correlation. That will depend on just how the y_{gi} are obtained from the probabilities given by $F(\boldsymbol{X}_{gi}\boldsymbol{\beta})$; see Section 6. For the rest of this section, we merely allow for the possibility that intra-cluster correlation exists.

For the logit and probit models, and many others, $F(-x) = 1 - F(x)$ for any argument x. In what follows, we assume that this is the case. Some of our results would need to be modified if it were not.

If $\boldsymbol{y}$ is an N-vector with typical element y_{gi}, the log-likelihood function for (1) can be written as

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \Big(y_{gi} \log F(\boldsymbol{X}_{gi}\boldsymbol{\beta}) + (1 - y_{gi}) \log F(-\boldsymbol{X}_{gi}\boldsymbol{\beta}) \Big). \tag{2}$$

The first-order conditions for $\hat{\boldsymbol{\beta}}$ are then

$$\sum_{g=1}^{G} \sum_{i=1}^{N_g} \frac{\Big(y_{gi} - F(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})\Big) f(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}}) X_{gij}}{F(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}}) F(-\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})} = 0, \quad j = 1, \ldots, k, \tag{3}$$

where X_{gij} is the j^{th} element of $\boldsymbol{X}_{gi}$. The score vector for the g^{th} cluster is

$$\boldsymbol{s}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \boldsymbol{s}_{gi}(\boldsymbol{\beta}), \quad \text{where} \quad \boldsymbol{s}_{gi}(\boldsymbol{\beta}) = \frac{\Big(y_{gi} - F(\boldsymbol{X}_{gi}\boldsymbol{\beta})\Big) f(\boldsymbol{X}_{gi}\boldsymbol{\beta}) \boldsymbol{X}_{gi}}{F(\boldsymbol{X}_{gi}\boldsymbol{\beta}) F(-\boldsymbol{X}_{gi}\boldsymbol{\beta})}. \tag{4}$$

Thus the k equations in (3) can be rewritten as a single equation for the entire parameter vector $\hat{\boldsymbol{\beta}}$, namely, $\hat{\boldsymbol{s}} = \sum_{g=1}^{G} \hat{\boldsymbol{s}}_g = \sum_{g=1}^{G} \boldsymbol{s}_g(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. Of course, if the scores were assumed to be independent within clusters, it would be more natural to write this as the summation of the N empirical score vectors $\boldsymbol{s}_{gi}(\hat{\boldsymbol{\beta}})$. But we are merely assuming independence across clusters, with potentially arbitrary patterns of intra-cluster dependence.

Most treatments of binary response models assume that the observations are independent or, equivalently, that each cluster contains just one observation. In that case, the asymptotic variance matrix is readily obtained from the result that

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} \Big(\text{plim}\, N^{-1} \boldsymbol{H}(\boldsymbol{\beta}_0)\Big)^{-1} N^{1/2} \sum_{i=1}^{N} \boldsymbol{s}_i(\boldsymbol{\beta}_0), \tag{5}$$

where "$\stackrel{a}{=}$" denotes asymptotic equality, $\boldsymbol{H}(\boldsymbol{\beta})$ is the Hessian, $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, and $\boldsymbol{s}_i(\boldsymbol{\beta}_0)$ is $\boldsymbol{s}_g(\boldsymbol{\beta}_0)$ for the special case in which clusters and observations coincide. This leads to the variance matrix estimators

$$\hat{\boldsymbol{V}}_H(\hat{\boldsymbol{\beta}}) = -\boldsymbol{H}(\hat{\boldsymbol{\beta}})^{-1} \quad \text{and} \quad \hat{\boldsymbol{V}}_{\mathcal{I}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})^{-1}, \tag{6}$$

where $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})$ denotes the information matrix evaluated at $\hat{\boldsymbol{\beta}}$. Asymptotically, the plim of $N^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ equals minus the plim of $N^{-1}\boldsymbol{H}(\boldsymbol{\beta})$ by the information matrix equality. For the logit model, $\hat{\boldsymbol{V}}_H(\hat{\boldsymbol{\beta}})$ actually equals $\hat{\boldsymbol{V}}_{\mathcal{I}}(\hat{\boldsymbol{\beta}})$ (Section 3.1), but this is not true in general.

When there is no clustering, it is not hard to show that

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^{\top} \boldsymbol{\Upsilon}(\boldsymbol{\beta}) \boldsymbol{X}, \tag{7}$$

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

$$\Upsilon_i(\boldsymbol{\beta}) = \frac{f^2(\boldsymbol{X}_i\boldsymbol{\beta})}{F(\boldsymbol{X}_i\boldsymbol{\beta})F(-\boldsymbol{X}_i\boldsymbol{\beta})}; \tag{8}$$

see, among many others, Davidson and MacKinnon (2004, Section 11.3).

The asymptotic equality (5) does not hold when there is clustering, because the rate at which $\hat{\boldsymbol{\beta}}$ tends to $\boldsymbol{\beta}_0$ is, in general, not $N^{-1/2}$; see Djogbenou et al. (2019). Nevertheless, it is possible to make inferences based on the CRVEs

$$\boldsymbol{H}(\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})\boldsymbol{H}(\hat{\boldsymbol{\beta}})^{-1} \quad \text{and} \quad \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})^{-1}, \tag{9}$$

where $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})$ is an estimator of the expectation of the sum of the outer products of the $\boldsymbol{s}_g(\boldsymbol{\beta})$ with themselves; see Hansen and Lee (2019, Theorem 10). Conventional CRVEs have this familiar sandwich form, with a matrix based on the outer product of the scores sandwiched between two instances of something that estimates the inverse of the information matrix.

The most natural CRVE based on (9) is probably

$$\text{CV}_{1\text{H}}\text{:} \qquad \hat{\boldsymbol{V}}_{1\text{H}}(\hat{\boldsymbol{\beta}}) = \frac{G}{G-1}\frac{N-1}{N-k}\boldsymbol{H}(\hat{\boldsymbol{\beta}})^{-1}\left(\sum_{g=1}^{G}\hat{\boldsymbol{s}}_g\hat{\boldsymbol{s}}_g^\top\right)\boldsymbol{H}(\hat{\boldsymbol{\beta}})^{-1}. \tag{10}$$

The filling in the sandwich here is the obvious estimator of $\text{E}\left(\boldsymbol{s}_g(\boldsymbol{\beta})\boldsymbol{s}_g(\boldsymbol{\beta})^\top\right)$. The degrees-of-freedom factor is optional, but it seems reasonable to include it by analogy with the usual CV_1 estimator for linear regression models. The estimator in (10) is almost the same as the one used by `Stata`, which omits the factor of $(N-1)/(N-k)$. We refer to it as $\text{CV}_{1\text{H}}$ because it is analogous to the CV_1 estimator and employs the estimated Hessian.

For the model (1), the contribution to the Hessian made by the gi^{th} observation depends on the value of y_{gi}. Specifically,

$$H_{gi}(\boldsymbol{\beta}) = \frac{f'(-\boldsymbol{X}_{gi}\boldsymbol{\beta})F(-\boldsymbol{X}_{gi}\boldsymbol{\beta}) - f^2(-\boldsymbol{X}_{gi}\boldsymbol{\beta})}{F^2(-\boldsymbol{X}_{gi}\boldsymbol{\beta})}\boldsymbol{X}_{gi}^\top\boldsymbol{X}_{gi} \qquad \text{if } y_{gi} = 0, \tag{11}$$

$$H_{gi}(\boldsymbol{\beta}) = \frac{f'(\boldsymbol{X}_{gi}\boldsymbol{\beta})F(\boldsymbol{X}_{gi}\boldsymbol{\beta}) - f^2(\boldsymbol{X}_{gi}\boldsymbol{\beta})}{F^2(\boldsymbol{X}_{gi}\boldsymbol{\beta})}\boldsymbol{X}_{gi}^\top\boldsymbol{X}_{gi} \qquad \text{if } y_{gi} = 1. \tag{12}$$

The $k \times k$ matrices with typical elements given by (11) or (12) are summed over all the observations for which y_{gi} equals 0 and 1, respectively, to obtain $\boldsymbol{H}(\hat{\boldsymbol{\beta}})$.

The estimated Hessian $\boldsymbol{H}(\hat{\boldsymbol{\beta}})$ can be replaced by its expectation. In order to take the expectations of (11) and (12), we use the fact that $\text{E}(y_{gi}) = F(\boldsymbol{X}_{gi}\boldsymbol{\beta})$. Thus we multiply (11) by $F(-\boldsymbol{X}_{gi}\boldsymbol{\beta})$ and (12) by $F(\boldsymbol{X}_{gi}\boldsymbol{\beta})$, and then add them. The first terms in each of the two numerators cancel out, since $f'(-\boldsymbol{X}_{gi}\boldsymbol{\beta}) + f'(\boldsymbol{X}_{gi}\boldsymbol{\beta}) = 0$ by the symmetry of the function

$f(\cdot)$. This leaves just the second terms, of which the weighted sum is

$$-\left(\frac{f^2(-\boldsymbol{X}_{gi}\boldsymbol{\beta})}{F(-\boldsymbol{X}_{gi}\boldsymbol{\beta})}+\frac{f^2(\boldsymbol{X}_{gi}\boldsymbol{\beta})}{F(\boldsymbol{X}_{gi}\boldsymbol{\beta})}\right)\boldsymbol{X}_{gi}^\top\boldsymbol{X}_{gi}. \tag{13}$$

When we cross-multiply and make use of the facts that $F(\boldsymbol{X}_{gi}\boldsymbol{\beta})+F(-\boldsymbol{X}_{gi}\boldsymbol{\beta})=1$ and that $f(\boldsymbol{X}_{gi}\boldsymbol{\beta})=f(-\boldsymbol{X}_{gi}\boldsymbol{\beta})$, (13) simplifies to

$$\frac{-f^2(\boldsymbol{X}_{gi}\boldsymbol{\beta})}{F(\boldsymbol{X}_{gi}\boldsymbol{\beta})F(-\boldsymbol{X}_{gi}\boldsymbol{\beta})}\boldsymbol{X}_{gi}^\top\boldsymbol{X}_{gi}, \tag{14}$$

which is just $-\Upsilon_{gi}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_{gi}^\top\boldsymbol{X}_{gi}$, where $\Upsilon_{gi}(\boldsymbol{\beta})$ was defined in (8). This leads to the CRVE

$$\text{CV}_{1\mathcal{I}}\text{:}\quad \hat{\boldsymbol{V}}_{1\mathcal{I}}=\frac{G}{G-1}\frac{N-1}{N-k}(\boldsymbol{X}^\top\hat{\boldsymbol{\Upsilon}}\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\hat{\boldsymbol{s}}_g\hat{\boldsymbol{s}}_g^\top\right)(\boldsymbol{X}^\top\hat{\boldsymbol{\Upsilon}}\boldsymbol{X})^{-1}, \tag{15}$$

where $\hat{\boldsymbol{\Upsilon}}=\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})$. The matrix that is inverted twice here is the empirical counterpart of (7). We refer to this estimator as $\text{CV}_{1\mathcal{I}}$ because it uses the information matrix $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})$ instead of the Hessian $\boldsymbol{H}(\hat{\boldsymbol{\beta}})$.

Cluster-jackknife variance matrix estimators for the linear regression model are discussed in MacKinnon et al. (2023c, b) and Hansen (2023). Each cluster is deleted in turn, yielding the vector of delete-one estimates $\hat{\boldsymbol{\beta}}^{(g)}$ when the g^{th} cluster is deleted. The jackknife CRVE is

$$\text{CV}_{3\text{J}}\text{:}\qquad \hat{\boldsymbol{V}}_{3\text{J}}(\hat{\boldsymbol{\beta}})=\frac{G-1}{G}\sum_{g=1}^{G}(\hat{\boldsymbol{\beta}}^{(g)}-\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(g)}-\bar{\boldsymbol{\beta}})^\top, \tag{16}$$

where $\bar{\boldsymbol{\beta}}$ is the arithmetic mean of the $\hat{\boldsymbol{\beta}}^{(g)}$. An alternative jackknife CRVE is

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

which differs from (16) only because it computes the variance around $\hat{\boldsymbol{\beta}}$ instead of $\bar{\boldsymbol{\beta}}$.

The notation in (16) and (17) is descended from the use of HC_3 in MacKinnon and White (1985) to denote a heteroskedasticity-consistent variance matrix estimator based on the jackknife. Bell and McCaffrey (2002) discusses both (16) and (17), but they are computed in a completely different way so that they have the usual sandwich form. That method would be computationally attractive if all the N_g were very small, but it can be extremely expensive, or even infeasible, when any of them is large (MacKinnon et al. 2023b, Section 4). Simulation evidence in Bell and McCaffrey (2002) and MacKinnon et al. (2023b) suggests that, for linear regression models, $\text{CV}_{3\text{J}}$ and CV_3 tend to be extremely similar. The former is always at least slightly smaller than the latter, however, because the variation of the $\hat{\boldsymbol{\beta}}^{(g)}$

around their mean of $\bar{\boldsymbol{\beta}}$ cannot exceed their variation around any other vector, including $\hat{\boldsymbol{\beta}}$.

It is inevitably costlier to compute $\mathrm{CV}_{3\mathrm{J}}$ or CV_3 for a binary response model than for a linear regression model with similar numbers of parameters, clusters, and observations, because in the former case we need to perform $G+1$ nonlinear optimizations. Much of the time, however, $\hat{\boldsymbol{\beta}}$ should provide a good starting point for obtaining each of the $\hat{\boldsymbol{\beta}}^{(g)}$. Thus the cost of computing $G+1$ sets of estimates should be less than $G+1$ times as great as the cost of computing $\hat{\boldsymbol{\beta}}$ by itself. Moreover, unless G is extremely large, computing $G+1$ sets of estimates will be much cheaper than any bootstrap method that requires nonlinear estimation for every bootstrap sample. Note, however, that the bootstrap methods introduced in Section 3 do not require any nonlinear estimation within the bootstrap procedure.

Another advantage of jackknife methods is that they can readily be adapted to make inferences about smooth functions of $\boldsymbol{\beta}$. For example, if we care about $\delta = \beta_2/\beta_3$, we simply need to calculate $\hat{\delta}$ for the entire sample and $\hat{\delta}^{(g)}$ for each vector of delete-one estimates and then use the analog of (16) or (17) to calculate its jackknife variance. Bootstrap methods also have this useful feature.

The jackknife methods we propose do, however, suffer from a potentially important computational problem. Suppose there exists some linear combination of the $\boldsymbol{X}_{gi}$, say $\boldsymbol{X}_{gi}\boldsymbol{\beta}^\bullet$, with the property that

$$y_{gi} = 0 \quad \text{whenever} \quad \boldsymbol{X}_{gi}\boldsymbol{\beta}^\bullet < 0, \quad \text{and} \tag{18}$$

$$y_{gi} = 1 \quad \text{whenever} \quad \boldsymbol{X}_{gi}\boldsymbol{\beta}^\bullet > 0. \tag{19}$$

Then it is possible to make the value of the log-likelihood function (2), which is always negative, arbitrarily close to 0 by setting $\boldsymbol{\beta} = \gamma\boldsymbol{\beta}^\bullet$ and letting $\gamma \to \infty$. This is precisely what a numerical optimization routine will attempt to do, although it will normally stop with an error message long before any element of $\hat{\boldsymbol{\beta}}$ becomes infinitely large. In this case, the vector $\boldsymbol{X}\boldsymbol{\beta}^\bullet$, which of course is not unique, is said to be a perfect classifier, since it allows us to predict y_{gi} perfectly for every observation in the sample.

When there is a perfect classifier, we cannot obtain well-defined estimates of all the parameters by maximizing (2). If this happens for the entire sample, then we either need to drop one or more regressors, obtain additional data, or use some form of regularization. The problem for the jackknife estimators is that, even if there are no perfect classifiers for the entire sample, there might be a perfect classifier for one or more of the subsamples. When this happens, the values of $\mathrm{CV}_{3\mathrm{J}}$ and CV_3 may become extremely large and completely unreliable. Thus any program to compute $\mathrm{CV}_{3\mathrm{J}}$ and CV_3 needs to check whether there is a perfect classifier when any one of the G clusters is dropped. When that happens, it should

either report that the variance matrix could not be computed or omit the offending vector(s) of delete-one estimates and report that it has done so. In the latter case, especially if the deleted cluster is large, $\mathrm{CV}_{3\mathrm{J}}$ is likely to be more reliable than CV_3, because $\bar{\boldsymbol{\beta}}$ for the reduced sample may differ noticeably from $\hat{\boldsymbol{\beta}}$ for the full sample.

It is straightforward to base inference on CV_3 or $\mathrm{CV}_{3\mathrm{J}}$. Suppose there are $r \geq 1$ linear restrictions. These can be written as $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, with $\boldsymbol{R}$ an $r \times k$ matrix and $\boldsymbol{r}$ an r-vector. Tests of these restrictions are commonly based on the Wald statistic

$$W(\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top (\boldsymbol{R}\hat{\boldsymbol{V}}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \tag{20}$$

where $\hat{\boldsymbol{V}}$ could be any of the CRVEs defined in (10), (15), (16), or (17). Asymptotically, as $G \to \infty$, $W(\hat{\boldsymbol{\beta}})$ is distributed as $\chi^2(r)$ under the null hypothesis.

When there is just one restriction, the signed square root of $W(\hat{\boldsymbol{\beta}})$ has the form of a t-statistic. When $\boldsymbol{a}^\top$ is a single row of $\boldsymbol{R}$ and $\boldsymbol{r} = \boldsymbol{0}$, such a t-statistic can be written as

$$t_a = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\boldsymbol{a}^\top \hat{\boldsymbol{V}} \boldsymbol{a})^{1/2}}. \tag{21}$$

With linear models it is customary to compare this with the $t(G-1)$ distribution (Bester, Conley, and Hansen 2011). However, both the `logit` command in `Stata` and the `sandwich` package in `R` compare this with the $N(0,1)$ distribution. In the very common case in which there is a single zero restriction, say that $\beta_k = 0$, (21) reduces to $\hat{\beta}_k / \hat{s}_k$, where $\hat{s}_k$ is the square root of the k^{th} diagonal element of $\hat{\boldsymbol{V}}$.

3 Methods Based on Linearization

Computing either $\mathrm{CV}_{3\mathrm{J}}$ or CV_3 requires $G+1$ nonlinear optimizations. However, replacing the $\hat{\boldsymbol{\beta}}^{(g)}$ in (16) and (17) by estimates from a linear approximation yields cluster-jackknife CRVEs that are much cheaper to compute. The linear approximation is based on the artificial regression for binary response models proposed in Davidson and MacKinnon (1984). However, it can be performed without explicitly running a regression. We just need the contribution to the scores, $\boldsymbol{s}_g(\boldsymbol{\beta})$, and the contribution to the information matrix, $\boldsymbol{J}_g(\boldsymbol{\beta})$, made by each of the clusters. The $\boldsymbol{s}_g(\boldsymbol{\beta})$ are given by (4), and

$$\boldsymbol{J}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \frac{f^2(\boldsymbol{X}_{gi}\boldsymbol{\beta})}{F(\boldsymbol{X}_{gi}\boldsymbol{\beta})F(-\boldsymbol{X}_{gi}\boldsymbol{\beta})} \boldsymbol{X}_{gi}^\top \boldsymbol{X}_{gi} \tag{22}$$

from (14). Then the estimates from linearizing the model around $\boldsymbol{\beta}$ are

$$\boldsymbol{b}(\boldsymbol{\beta}) = \left(\sum_{g=1}^{G} \boldsymbol{J}_g(\boldsymbol{\beta}) \right)^{-1} \sum_{g=1}^{G} \boldsymbol{s}_g(\boldsymbol{\beta}) = \boldsymbol{J}(\boldsymbol{\beta})^{-1} \boldsymbol{s}(\boldsymbol{\beta}), \tag{23}$$

where $\boldsymbol{J}(\boldsymbol{\beta}) = \sum_{g=1}^{G} \boldsymbol{J}_g(\boldsymbol{\beta})$. When the $\boldsymbol{s}_g(\boldsymbol{\beta})$ and $\boldsymbol{J}_g(\boldsymbol{\beta})$ are evaluated at the true value $\boldsymbol{\beta}_0$, the estimate $\boldsymbol{b}(\boldsymbol{\beta}_0)$ provides a linear approximation to $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. This gives us almost everything we need to compute linearized jackknife and bootstrap tests.

To compute linear approximations to the delete-one-cluster estimates, we first estimate the model by maximizing (2). Then we form the cluster-level vectors and matrices $\hat{\boldsymbol{s}}_g = \boldsymbol{s}_g(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{J}}_g = \boldsymbol{J}_g(\hat{\boldsymbol{\beta}})$ using (4) and (22). The linear approximations to $\hat{\boldsymbol{\beta}}^{(g)} - \hat{\boldsymbol{\beta}}$ when each cluster is omitted in turn are then

$$\hat{\boldsymbol{b}}^{(g)} = (\hat{\boldsymbol{J}} - \hat{\boldsymbol{J}}_g)^{-1} (\hat{\boldsymbol{s}} - \hat{\boldsymbol{s}}_g), \quad g = 1, \ldots, G. \tag{24}$$

We can use these approximations to compute cluster-jackknife variance matrices. The one comparable to (17) is

$$\text{CV}_{3\text{L}}\text{:} \qquad \hat{\boldsymbol{V}}_{3\text{L}}(\hat{\boldsymbol{\beta}}) = \frac{G-1}{G} \sum_{g=1}^{G} \hat{\boldsymbol{b}}^{(g)} \hat{\boldsymbol{b}}^{(g)\top}. \tag{25}$$

Nothing is subtracted from the $\hat{\boldsymbol{b}}^{(g)}$ here, because when we evaluate (23) at $\hat{\boldsymbol{\beta}}$, the estimate $\hat{\boldsymbol{b}} = \boldsymbol{b}(\hat{\boldsymbol{\beta}})$ is identically zero by the first-order conditions for $\hat{\boldsymbol{\beta}}$. We could instead subtract $\bar{\boldsymbol{b}}$, the arithmetic mean of the $\hat{\boldsymbol{b}}^{(g)}$. If we did so, we would obtain a linearized cluster-jackknife CRVE comparable to (16). The computations in (24) and (25) used to compute $\text{CV}_{3\text{L}}$ are usually far less expensive than the ones needed to compute CV_3; see Section 7.2.

The linearization given by (23) can also be used to compute a $\text{CV}_{2\text{L}}$ variance matrix similar to the CV_2 matrix proposed in Bell and McCaffrey (2002) and referred to there as "bias-reduced linearization." These matrices are generalizations of the HC_2 matrix of MacKinnon and White (1985). There is more than one way to compute them, only one of which (Niccodemi et al. 2020) is feasible for large samples; see Appendix A. Because the simulations in MacKinnon et al. (2023b) suggest that CV_2 very rarely performs better than CV_3 (although it always performs better than CV_1), we do not study $\text{CV}_{2\text{L}}$ further.

It seems plausible that, except perhaps in cases where one or a few clusters are highly influential (MacKinnon et al. 2023c), all four of the cluster-jackknife variance matrices will yield similar inferences. We will investigate this conjecture in Section 6.

The linear approximation (23) can also be used to compute new versions of the wild cluster bootstrap, which we refer to as "wild cluster linearized," or WCL, bootstraps. Like the score bootstraps proposed in Kline and Santos (2012), the WCL bootstraps are based

on restricted or unrestricted empirical scores. However, they differ in one important respect from the Kline and Santos (2012) methods. Both procedures generate bootstrap samples from empirical bootstrap scores, but then our WCL methods multiply those bootstrap scores by the inverse of some version of the $\boldsymbol{J}$ matrix, in order to mimic the estimation step that yields empirical scores for the actual model.

We now describe the bootstrap data-generating processes. To avoid having to give two separate results for the restricted and unrestricted bootstraps, we let "$\ddot{x}$" denote either "$\tilde{x}$" or "$\hat{x}$" for any x. In the first step, we multiply the score vector $\ddot{\boldsymbol{s}}_g$ for cluster g by random variates v_g^{*b} for $b = 1, \ldots, B$ bootstrap samples. The v_g^{*b} must have mean 0 and variance 1. In most cases, it seems best for them to be independent draws from the Rademacher distribution, for which v_g^{*b} equals $+1$ and -1 with equal probabilities; see Djogbenou et al. (2019). Thus the bootstrap score vectors are

$$\ddot{\boldsymbol{s}}_g^{*b} = v_g^{*b} \ddot{\boldsymbol{s}}_g, \quad g = 1, \ldots, G. \tag{26}$$

The next step is to estimate the coefficient vector $\boldsymbol{b}$ by least squares:

$$\ddot{\boldsymbol{b}}^{*b} = \left(\sum_{g=1}^{G} \ddot{\boldsymbol{J}}_g \right)^{-1} \sum_{g=1}^{G} \ddot{\boldsymbol{s}}_g^{*b}. \tag{27}$$

The vector $\ddot{\boldsymbol{b}}^{*b}$ is then used to compute the empirical bootstrap score vectors

$$\ddot{\boldsymbol{w}}_g^{*b} = \ddot{\boldsymbol{s}}_g^{*b} - \ddot{\boldsymbol{J}}_g \ddot{\boldsymbol{b}}^{*b}, \quad g = 1, \ldots, G. \tag{28}$$

These are what the bootstrap score vectors become after the model has been "estimated" using the linearization (23).

The CV_1 bootstrap variance matrix can then be written as

$$\ddot{\boldsymbol{V}}_b^* = \frac{G(N-1)}{(G-1)(N-k)} \ddot{\boldsymbol{J}}^{-1} \left(\sum_{g=1}^{G} \ddot{\boldsymbol{w}}_g^{*b} (\ddot{\boldsymbol{w}}_g^{*b})^\top \right) \ddot{\boldsymbol{J}}^{-1}, \tag{29}$$

and the bootstrap t-statistic that corresponds to (21) is

$$\ddot{t}_a^{*b} = \frac{\boldsymbol{a}^\top \ddot{\boldsymbol{b}}^{*b}}{(\boldsymbol{a}^\top \ddot{\boldsymbol{V}}_b^* \boldsymbol{a})^{1/2}}. \tag{30}$$

In principle, we could instead compute a CV_3 bootstrap variance matrix, but using (29) makes the bootstrap computations much faster. Transforming the bootstrap score vectors in the way proposed in MacKinnon et al. (2023b) (see below) achieves much the same effect as using CV_3, but at far less computational cost.

As usual, several different bootstrap P values can be computed. For cross-sectional models estimated by least squares, where bias is generally not a problem, the symmetric bootstrap P value is usually appropriate. It is computed as

$$\hat{P}^*_{\mathrm{s}}(t_a) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\Big(|t_a^{*b}| > |t_a|\Big), \tag{31}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. We reject the null hypothesis for a test at level α whenever $\hat{P}^*_{\mathrm{s}}(t_a) < \alpha$. An alternative is the equal-tail bootstrap P value

$$\hat{P}^*_{\mathrm{et}}(t_a) = \frac{2}{B}\min\left(\sum_{b=1}^{B}\mathbb{I}(t_a^{*b} > t_a),\ \sum_{b=1}^{B}\Big(\mathbb{I}(t_a^{*b} \leq t_a)\Big)\right). \tag{32}$$

Because the estimated slope coefficients for binary response models tend to be biased away from zero (MacKinnon and Smith 1998), it might be preferable to use (32) instead of (31) for these models. However, we did not find any real difference between them in the experiments of Section 6.

The WCL bootstrap methods that we have just described are the analogs for binary response models of the classic wild cluster bootstrap methods for OLS regression, which are called WCR-C and WCU-C in MacKinnon et al. (2023b) to distinguish them from newer variants introduced in that paper. We therefore refer to the two WCL methods as the WCLR-C and WCLU-C bootstraps. As usual, the "R" and "U" here indicate whether the bootstrap DGP uses restricted or unrestricted estimates. The "-C" indicates that the score vectors are not transformed before generating the bootstrap samples.

Many of the computations for WCR-C/WCU-C and WCLR-C/WCLU-C are identical. For the former, everything depends on the score vector contributions, $\boldsymbol{X}_g^\top \ddot{\boldsymbol{u}}_g$, and the negative Hessian matrix contributions, $\boldsymbol{X}_g^\top \boldsymbol{X}_g$. For the latter, everything depends in exactly the same way on the $\ddot{\boldsymbol{s}}_g$ and the $\ddot{\boldsymbol{J}}_g$.

This insight shows that the WCLR and WCLU bootstraps can easily be modified to make them analogous to the WCR-S and WCU-S bootstraps proposed in MacKinnon et al. (2023b). The modification involves replacing the empirical scores $\ddot{\boldsymbol{s}}$ in (26) by transformed empirical scores based on the cluster jackknife. The "-S" in the names stands for "transformed score." The key equations, adapted to the present case, are

$$\acute{\boldsymbol{s}}_g = \hat{\boldsymbol{s}}_g - \hat{\boldsymbol{J}}_g \hat{\boldsymbol{b}}^{(g)}, \quad g = 1, \ldots, G, \tag{33}$$

for the unrestricted scores, and, assuming that the only restriction is $\beta_k = 0$,

$$\dot{\boldsymbol{s}}_g = \tilde{\boldsymbol{s}}_g - \tilde{\boldsymbol{J}}_{1g} \tilde{\boldsymbol{b}}_1^{(g)}, \quad g = 1, \ldots, G, \tag{34}$$

for the restricted scores. Equations (33) and (34) are, respectively, analogous to (38) and (37) in MacKinnon et al. (2023b). In (34), the matrix $\tilde{\boldsymbol{J}}_{1g}$ contains the first $k-1$ columns of $\tilde{\boldsymbol{J}}_g$, and the vector $\tilde{\boldsymbol{b}}_1^{(g)}$ contains the first $k-1$ elements of $\tilde{\boldsymbol{b}}^{(g)}$. When there are $r < k$ linear restrictions, (34) can be replaced by a more complicated equation analogous to (34) in MacKinnon et al. (2023b).

Using the transformed empirical scores from (33) or (34) yields what we will call the WCLU-S and WCLR-S bootstraps, respectively. The purpose of the transformations is to undo the distortions of the empirical scores caused by estimating $\boldsymbol{\beta}$, at least to the extent that it is feasible to do so. This should allow the bootstrap DGP to mimic the unknown true DGP more accurately. Simulation evidence in MacKinnon et al. (2023b) suggests that the WCR-S and WCU-S bootstraps can perform substantially better than the classic WCR-C and WCU-C bootstraps in many cases. This also seems to be true for WCLR-S and WCLU-S relative to WCLR-C and WCLU-C; see Section 6. In particular, confidence intervals based on WCLU-S perform very much better than ones based on WCLU-C.

All the methods proposed in this section are implemented in the `Stata` package `logitjack`; see Appendix B.

3.1 The Logit and Probit Models

For the probit model, the cumulative standard normal distribution function $\Phi(\cdot)$ and the standard normal density $\phi(\cdot)$ play the roles of the functions $F(\cdot)$ and $f(\cdot)$. In our simulations and empirical examples, however, we focus on the logit model. It seems to be more widely used and is computationally a bit simpler than the probit model. The logistic function is

$$\Lambda(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}, \tag{35}$$

and its first derivative is

$$\lambda(x) = \frac{e^x}{(1+e^x)^2} = \Lambda(x)\Lambda(-x). \tag{36}$$

The functions $\Lambda(\cdot)$ and $\lambda(\cdot)$ play the roles of $F(\cdot)$ and $f(\cdot)$ for the logit model,

For most binary response models, the variance matrix estimators in (10) and (15) are different. In the case of the logit model, however, they are numerically identical. From (8), it is easy to see that a typical diagonal element of $\hat{\boldsymbol{\Upsilon}}$ is simply $\hat{\Upsilon}_{gi} = \Lambda(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})\Lambda(-\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})$. But this is also what the absolute value of each of the factors that multiply $\boldsymbol{X}_{gi}^\top\boldsymbol{X}_{gi}$ in (11) and (12) simplify to when they are evaluated at $\hat{\boldsymbol{\beta}}$. Because the Hessian appears twice in the sandwich estimator (10), it is only the absolute value that matters. Thus, for the logit model, $\mathrm{CV}_{1\mathcal{I}}$ and $\mathrm{CV}_{1\mathrm{H}}$ coincide.

For the logit model, the score vectors defined in (4) simplify to

$$\boldsymbol{s}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \Big(y_{gi} - \Lambda(\boldsymbol{X}_{gi}\boldsymbol{\beta})\Big)\boldsymbol{X}_{gi}, \quad g = 1, \ldots, G, \tag{37}$$

because $\lambda(x) = \Lambda(x)\Lambda(-x)$. In addition, the g^{th} contribution to the information matrix is

$$\boldsymbol{J}_g(\boldsymbol{\beta}) = \sum_{i=1}^{N_g} \Lambda_{gi}(\boldsymbol{\beta})\Lambda_{gi}(-\boldsymbol{\beta})\boldsymbol{X}_{gi}^\top \boldsymbol{X}_{gi}, \tag{38}$$

which is a bit simpler than (22). In general, the logit model is easier and somewhat cheaper to estimate than the probit model.

3.2 The Linear Probability Model

It is not difficult to estimate a binary response model and then linearize it using (23), so as to obtain linear approximations to the delete-one-cluster estimates. However, an even simpler approach is to estimate a linear probability model (LPM) and use existing methods for inference in clustered least-squares regression models. The first step is to run the regression

$$y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\delta} + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{39}$$

where u_{gi} is a disturbance term to be discussed below. There is nothing to ensure that $0 \leq \boldsymbol{X}_{gi}\boldsymbol{\delta} \leq 1$ in (39). Nevertheless, when all the P_i are well away from both 0 and 1, and all of the regressors are dummy variables, least squares typically does yield estimated probabilities that lie in the [0,1] interval most of the time and are quite similar to the ones from a binary response model. Thus it is common, and often not very harmful, for investigators to estimate the LPM (39) instead of the binary response model (1).

When an LPM is appropriate, the number of clusters and (for treatment models) the number of treated clusters are both reasonably large, and there is not too much inter-cluster variation, we might expect inferences based on CV_3, or even CV_1, from (39) to be fairly reliable (MacKinnon et al. 2023a). When any of these conditions is not satisfied, it may be safer to use some variant of the restricted wild cluster, or WCR, bootstrap. When the Rademacher distribution is used, the bootstrap dependent variable can take on only two values, each with probability 1/2. If $\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}}$ denotes the gi^{th} fitted value from the LPM, evaluated at the restricted estimates, these are

$$y_{gi}^* = \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} + (y_{gi} - \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}}) = y_{gi} \quad \text{and} \quad y_{gi}^* = \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} - (y_{gi} - \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}}) = 2\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} - y_{gi}. \tag{40}$$

The first value here is just the actual value of y_{gi}, which is 0 or 1. But the second is either

$2\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}}$ or $2\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} - 1$. Unless $\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} = 1/2$, one of these numbers must always lie outside the [0,1] interval. Thus, the y^*_{gi} must look very different from the y_{gi}. However, they do have the correct expectation under the bootstrap DGP. If $\mathrm{E}^*(\cdot)$ denotes expectation under the bootstrap probability measure, that is, conditional on the sample, then

$$\mathrm{E}^*(y^*_{gi}) = \frac{1}{2}\mathrm{E}^*(y_{gi}) + \frac{1}{2}\Big(2\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} - \mathrm{E}^*(y_{gi})\Big) = \frac{1}{2}(\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} + \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}}) = \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}}.$$

Although the bootstrap regressand (40) for the LPM may seem rather strange, it leads to the WCR-C bootstrap score vector

$$\sum_{i=1}^{N_g}(y^*_{gi} - \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}})\boldsymbol{X}_{gi} = \begin{cases} \sum_{i=1}^{N_g}(y_{gi} - \boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}})\boldsymbol{X}_{gi} \text{ with prob. } 1/2, \\ \sum_{i=1}^{N_g}(\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}} - y_{gi})\boldsymbol{X}_{gi} \text{ with prob. } 1/2. \end{cases} \tag{41}$$

Similarly, from (37), the WCLR-C bootstrap score vector for the logit model is

$$\sum_{i=1}^{N_g}(y^*_{gi} - \tilde{\Lambda}_{gi})\boldsymbol{X}_{gi} = \begin{cases} \sum_{i=1}^{N_g}(y_{gi} - \tilde{\Lambda}_{gi})\boldsymbol{X}_{gi} \text{ with prob. } 1/2, \\ \sum_{i=1}^{N_g}(\tilde{\Lambda}_{gi} - y_{gi})\boldsymbol{X}_{gi} \text{ with prob. } 1/2. \end{cases} \tag{42}$$

The WCR-C bootstrap score vector (41) and the WCLR-C bootstrap score vector (42) look very similar. The only difference is that the former uses $\boldsymbol{X}_{gi}\tilde{\boldsymbol{\delta}}$ as the fitted value for observation gi, and the latter uses $\tilde{\Lambda}_{gi} = \Lambda(\boldsymbol{X}_{gi}\tilde{\boldsymbol{\beta}})$. This suggests that, when the LPM provides a reasonably good approximation to a logit model, inferences based on an LPM and either variant of the WCR bootstrap are likely to be quite similar to inferences based on a logit model and the corresponding variant of the WCLR bootstrap.

We would also expect inferences based on both variants of the WCU bootstrap to be similar to inferences based on the corresponding variants of the WCLU bootstrap, and inferences based on CV_3 for the LPM to be similar to inferences based on both CV_3 and $\mathrm{CV}_{3\mathrm{L}}$ for the logit model. We will investigate these conjectures in Section 6.

4 Cluster Fixed Effects

It is very common for models where cluster-robust inference is employed to include cluster fixed effects. This creates some important computational issues, which we discuss in this section. The probability that $y_{gi} = 1$ is now

$$F\Big(\boldsymbol{X}_{gi}\boldsymbol{\beta} + \sum_{h=1}^{G}\delta_h D^h_{gi}\Big), \tag{43}$$

where the D_{gi}^h are cluster fixed-effect dummies, with $D_{gi}^h = 1$ whenever $g = h$ and $D_{gi}^h = 0$ otherwise. There are $k + G$ parameters to estimate, but interest focuses on the vector $\boldsymbol{\beta}$, which has k elements that do not include a constant term.

Under standard regularity conditions, it should be possible to estimate (43) by maximum likelihood using the entire sample. But when cluster h is omitted, it will be impossible to identify δ_h, because $D_{gi}^h = 0$ for all $g \neq h$. For linear regression models, MacKinnon et al. (2023b) discusses how to compute cluster-jackknife variance matrices when there are cluster fixed effects. The cheapest and easiest method is often to partial out the fixed effects before running either the full-sample regression or any of the delete-one-cluster regressions. But this partialing-out method is simply not feasible for (43), or indeed for any model that is nonlinear in the fixed effects.

A second method, also discussed in MacKinnon et al. (2023b), is to use a generalized inverse. For a linear regression model, this sets the coefficient δ_h to 0 for the regression that omits cluster h, but $\hat{\boldsymbol{\beta}}^{(h)}$ is the same as it would be for the partialing-out method. Whether or not this method could be used in the case of CV_3 or $\text{CV}_{3\text{J}}$ is unclear. It would mean using a logit or probit estimation routine that employs a generalized inverse and relying on that routine to do the right thing whenever one coefficient is completely unidentified. This seems risky and heavily dependent on the software employed. However, it is possible to use the generalized-inverse method for the linearized variance matrix estimators, $\text{CV}_{3\text{L}}$ and $\text{CV}_{3\text{LJ}}$, since computing each of the delete-one-cluster estimates just involves a linear regression. This is also the case for the WCLR-S and WCLU-S bootstraps, where a generalized inverse can be used in (24) or the analogous equation for the restricted case.

A third method, which is the only one we are aware of that will work for CV_3 and $\text{CV}_{3\text{J}}$, is to estimate $G+1$ different binary response models. The model for the full sample will have $k + G$ coefficients, but the model for each of the delete-one-cluster samples will have only $k+G-1$ coefficients, because the fixed-effect dummy for the deleted cluster must be omitted. Although this is conceptually straightforward, it may be challenging to program efficiently, because the set of fixed effects will be different for each of the $G + 1$ models. As with the other two methods, the $\hat{\boldsymbol{\beta}}^{(g)}$ are computed for all clusters, but the $\hat{\delta}_h^{(g)}$ are not, because they cannot be computed when $g = h$. This means that the cluster-jackknife variance matrices can be computed only for $\hat{\boldsymbol{\beta}}$. Because this method is tricky and time-consuming, employing any of the bootstrap methods or $\text{CV}_{3\text{L}}$ is easier (and often much faster) than obtaining CV_3.

5 Confidence Intervals

There are many ways to construct confidence intervals for binary response models. Some of these are computationally convenient, but others are inconvenient because the models are nonlinear. In this section, we briefly discuss several methods. The intervals that are easy to compute will be studied in Section 6.

The simplest approach to constructing a $100(1-\alpha)\%$ confidence interval, where α often equals either 0.05 or 0.01, is to employ a symmetric interval of the form

$$\left[\hat{\beta}_j - c_{1-\alpha/2}\mathrm{se}(\hat{\beta}_j), \;\; \hat{\beta}_j + c_{1-\alpha/2}\mathrm{se}(\hat{\beta}_j)\right], \tag{44}$$

where $\hat{\beta}_j$ is the maximum likelihood estimate of the coefficient of interest, and $c_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of some distribution. There are, in principle, many different confidence intervals based on (44). The critical value $c_{1-\alpha/2}$ might come from either the $\mathrm{N}(0,1)$ distribution or the $t(G-1)$ distribution, and the standard error might come from any of several different cluster-robust variance estimators or numerous different bootstrap distributions.

The standard normal distribution is the default for logit and probit models in `Stata`, but quantiles of the $t(G-1)$ distribution are usually employed when constructing intervals like (44) for linear regression models using CV_1 or CV_3 standard errors. The results in Section 6 suggest that this is always a better choice for binary response models too.

Instead of using CV_1 or CV_3 standard errors, we can use a bootstrap standard error based on B bootstrap estimates, $\hat{\beta}_j^{*b}$. This is simply

$$\mathrm{se}_{\mathrm{boot}}(\hat{\beta}_j) = \left(\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_j^{*b} - \bar{\beta}_j^{*}\right)^2\right)^{1/2}, \tag{45}$$

where $\bar{\beta}_j^{*}$ is the arithmetic mean of the $\hat{\beta}_j^{*b}$. Any bootstrap DGP that does not impose the null hypothesis can be used to generate the bootstrap samples. However, using the best-known such DGP, namely, the pairs cluster bootstrap, would be extremely expensive, because it would involve estimating a nonlinear model for each of B bootstrap samples. In contrast, the wild cluster linearized bootstrap methods proposed in Section 3 are very inexpensive when the computational tricks of Roodman et al. (2019) are employed. In principle, either WCLU-C or WCLU-S could be used, but the latter seem to work much better; see Section 6.

Instead of using a WCLU bootstrap to estimate a bootstrap standard error from (45), we could construct a studentized bootstrap interval of the form

$$\left[\hat{\beta}_j - c^{*}_{1-\alpha/2}\mathrm{se}_1(\hat{\beta}_j), \;\; \hat{\beta}_j - c^{*}_{\alpha/2}\mathrm{se}_1(\hat{\beta}_j)\right]. \tag{46}$$

Here $\text{se}_1(\hat{\beta}_j)$ is the CV_1 standard error of $\hat{\beta}_j$, and $c^*_{\alpha/2}$ and $c^*_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of the bootstrap t-statistics. For example, if $B = 999$ and $\alpha = 0.05$, these would be numbers 25 and 975 in the list of bootstrap t-statistics sorted from smallest to largest. It may seem odd to use the CV_1 standard error in (46), because we have argued in MacKinnon et al. (2023b) that the CV_3 standard error is more reliable. But it is essential to use the same standard error in (46) as in the WCLU bootstrap itself. The advantages of using cluster-jackknife standard errors apply to the WCLU-S bootstrap through the transformation of the scores. This suggests that intervals based on WCLU-S should outperform ones based on WCLU-C.

In theory, the studentized bootstrap interval (46) may perform better than the interval (44) using bootstrap standard errors, for the same bootstrap DGP, because the former is based on a test statistic that is asymptotically pivotal and allows the t-statistic to have an asymmetric distribution. In contrast, the latter is not based on an asymptotically pivotal quantity and imposes symmetry on the distribution. We shall investigate this conjecture, and others, in Section 6.

For an unrestricted bootstrap DGP, the same set of bootstrap samples can used to form confidence intervals for, and test hypotheses about, any coefficient or set of coefficients. In contrast, for a restricted bootstrap DGP, a different set of bootstrap samples is needed every time we calculate a bootstrap P value. This means that, to obtain a WCLR confidence interval, the binary response model has to be estimated many times subject to the restriction that β_j equals each candidate value for the limits of the interval; see MacKinnon (2023, Section 3.4). When we attempted to implement this method, we sometimes encountered numerical problems in the logit routine. This made it infeasible to perform simulations with a large number of replications. We therefore decided not to include WCLR-based intervals in our simulations, and, at present, we cannot recommend them in most cases.

6 Simulation Evidence

We have performed a large number of simulation experiments for quite a few different tests, all for the logit model. How well the tests perform inevitably depends on many features of the model and DGP. Nevertheless, several interesting regularities emerge from our experiments. In particular, the classic CV_1-based t-test is prone to over-reject, often severely, and it almost always does so to a greater extent than the jackknife and bootstrap tests proposed in Sections 2 and 3. In many circumstances, but not all, we find that CV_3 t-tests and WCLR-S bootstrap tests are particularly reliable. In Section 7, we provide some advice about how to proceed when alternative tests yield differing inferences.

In order to investigate the finite-sample properties of cluster-robust t-tests, we need to generate samples with intra-cluster correlation. In principle, this could be done in many different ways. The one that we use is particularly easy to implement, since it just requires a uniform random number generator. First, we specify a parameter ϕ between 0 and 1. Then we generate G independent random variates $v_g \sim \mathrm{U}(0,1)$, N independent random variates $e_{gi} \sim \mathrm{U}(0,1)$, and up to N more independent random variates $v_{gi} \sim \mathrm{U}(0,1)$. For all $g = 1, \ldots, G$ and $i = 1, \ldots, N_g$, we then compute

$$u_{gi} = v_g \text{ if } e_{gi} \leq \phi, \text{ and } u_{gi} = v_{gi} \text{ if } e_{gi} > \phi. \tag{47}$$

$$y_{gi} = 0 \text{ if } F(\boldsymbol{X}_{gi}) \leq u_{gi}, \text{ and } y_{gi} = 1 \text{ if } F(\boldsymbol{X}_{gi}) > u_{gi}. \tag{48}$$

Thus, with probability ϕ, the random variate u_{gi} is equal to v_g, and, with probability $1 - \phi$, it is equal to v_{gi}. At one extreme, when $\phi = 0$, all of the u_{gi} are independent. At the other extreme, when $\phi = 1$, they all take the same value u_g. The value of the binary variate y_{gi} is then equal to 0 with probability $1 - F(\boldsymbol{X}_{gi}\boldsymbol{\beta})$ and to 1 with probability $F(\boldsymbol{X}_{gi}\boldsymbol{\beta})$, as usual, but these events are not independent across observations within each cluster unless $\phi = 0$.

Most of our experiments deal with tests of a restriction on one parameter in a logit model. The function $F(\boldsymbol{X}_{gi}\boldsymbol{\beta})$ is given by

$$\Lambda\Big(\beta_1 + \sum_{j=2}^{k-1} \beta_j X_{gij} + \beta_k T_{gi}\Big), \tag{49}$$

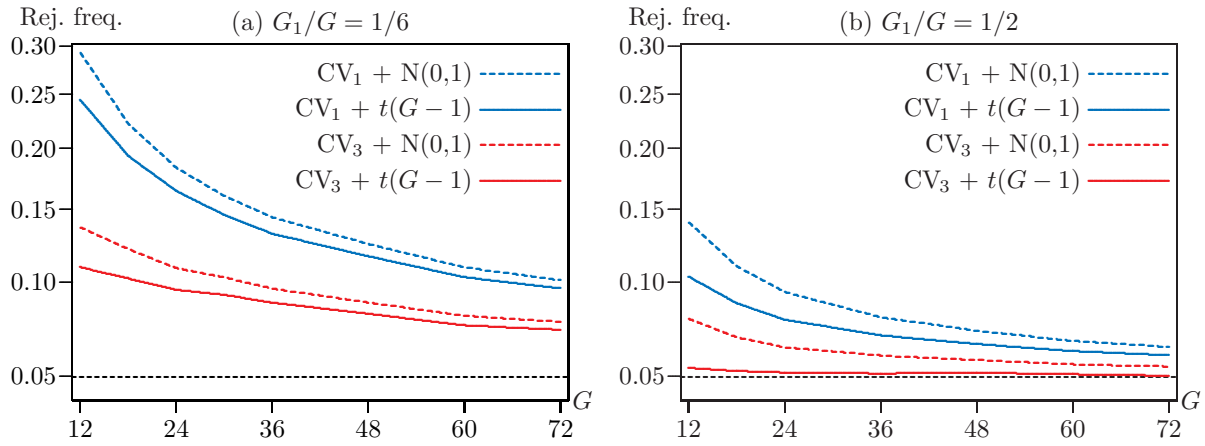
where the X_{gij} are binary random variables. For each j and for each g, a probability ω_g between 0.25 and 0.75 is chosen at random for each replication. Then, with probability ω_g, we set $X_{gij} = 1$ for all $i = 1, \ldots, N_g$, and otherwise we set $X_{gij} = 0$. This design is intended to mimic the situation, often encountered in treatment regressions, where all of the regressors are dummies. It allows these variables to vary moderately across clusters. In most experiments, $\beta_j = 1$ for $1 < j < k$. The model would fit better (worse) if these coefficients were larger (smaller). The treatment regressor T_{gi} equals 1 for G_1 randomly chosen clusters and 0 for the remaining $G_0 = G - G_1$ clusters, with $\beta_k = 0$ in most experiments. The unconditional expectation of y_{gi} is π, which depends on the β_j and the distribution of the X_{gij}. When we vary it, we do so by changing β_1, the constant term.

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

$$N_g = \left\lfloor N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right\rfloor, \quad g = 1, \ldots, G-1, \tag{50}$$

where $\lfloor x \rfloor$ means the integer part of x. The value of N_G is then set to $N - \sum_{g=1}^{G-1} N_g$. This

Figure 1: Rejection frequencies for tests at the 0.05 level as functions of G



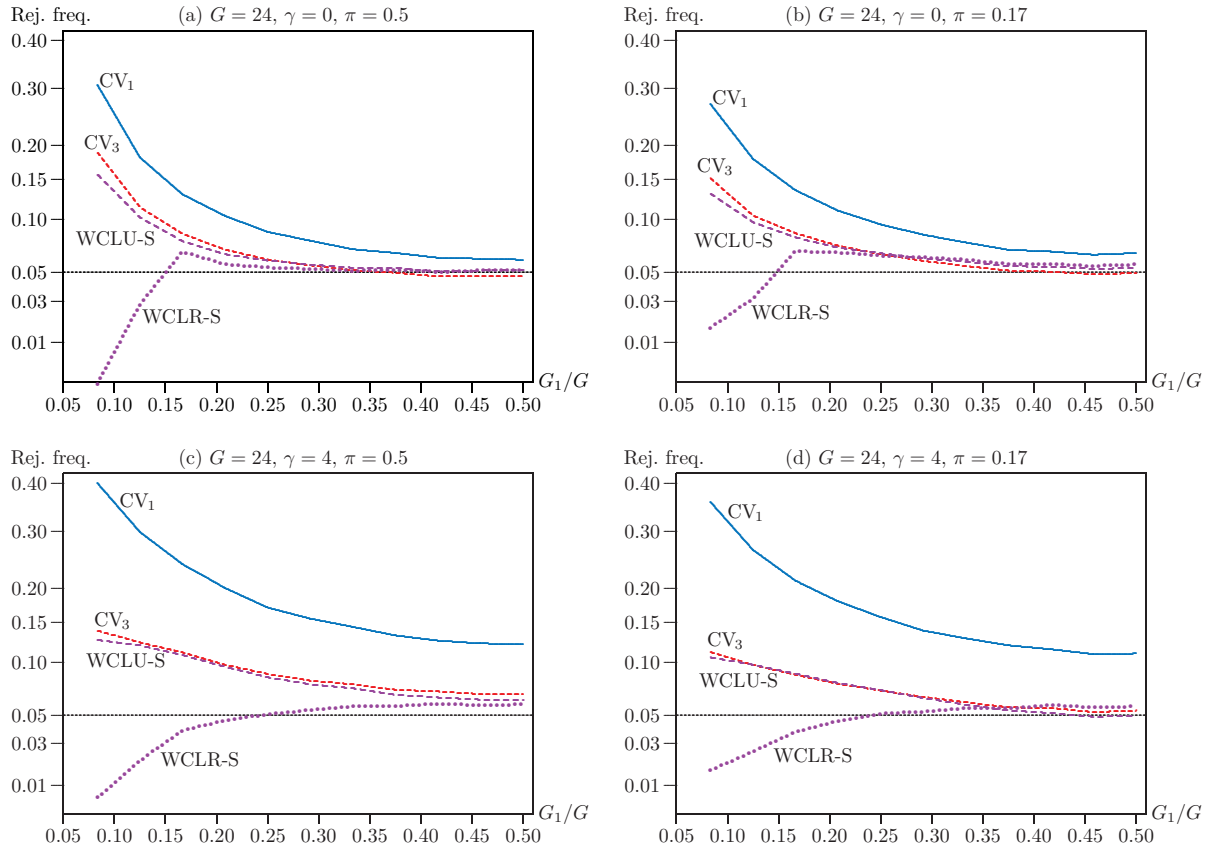
Notes: These experiments use 100,000 replications, with $G = 12, 18, 24, 30, 36, 48, 60, 72$, and $N = 500G$. The value of G_1 is $G/6$ in Panel (a) and $G/2$ in Panel (b). There are 7 regressors, one of which is a treatment dummy that is assigned at random, plus a constant term. The value of ϕ is 0.1. The extent to which cluster sizes vary is determined by the parameter γ in (50), which equals 2. The unconditional expectation of y_{gi} is $\pi = 0.316$. CV_1 and CV_3 denote cluster-robust t-statistics based on (10) and (17), respectively.

procedure has been used in MacKinnon and Webb (2017), Djogbenou et al. (2019), and several other papers. The key parameter here is γ, which determines how uneven the cluster sizes are. When $\gamma = 0$ and N/G is an integer, (50) implies that $N_g = N/G$ for all g. For $\gamma > 0$, cluster sizes vary more and more as γ increases. The largest value that we use is 4. In that case, when $G = 24$ and $N = 12000$, the largest cluster (1889 observations) is about 47 times as large as the smallest (40 observations). In contrast, when $\gamma = 2$, the largest cluster (1120 observations) is just under seven times as large as the smallest (163 observations).

In the first set of experiments, we let N vary from 6,000 to 36,000, with $G = N/500$ and $\gamma = 2$. We focus on tests of $\beta_k = 0$ in the logit model with $\text{E}(y_{gi})$ given by (49). These are based on either CV_1 or CV_3 standard errors and either the $\text{N}(0,1)$ or the $t(G-1)$ distribution. Cluster sizes vary moderately, with $\gamma = 2$. The amount of intra-cluster correlation is also fairly modest ($\phi = 0.1$). We believe this is realistic, especially for models with cluster fixed effects. For reasons of computational cost, however, our model does not have them.

Figure 1 shows rejection frequencies as functions of G for four t-tests. The vertical axis has been subjected to a square root transformation in order to handle the wide range of rejection frequencies that are observed in Panel (a). The results in this figure are striking. When only one-sixth of the clusters are treated, all the tests over-reject substantially, even with 72 clusters. However, when half of the clusters are treated, the most reliable test over-rejects only very slightly, even with just 12 clusters. This test uses CV_3 standard errors and $t(G-1)$ critical values. It is the most reliable test in all cases, whereas the test that uses

Figure 2: Rejection frequencies for tests at the 0.05 level as functions of G_1/G



Notes: These experiments use 100,000 replications, with $N = 12{,}000$, $G = 24$, and G_1 varying from 2 to 12. There are 7 regressors, one of which is a treatment dummy that is assigned at random. The value of ϕ is 0.1. The extent to which cluster sizes vary is determined by the parameter γ in (50), and π is the unconditional expectation of y_{gi}. CV_1 and CV_3 denote cluster-robust t-tests based on the $t(23)$ distribution. WCLR-S and WCLU-S denote tests based on symmetric bootstrap P values for the transformed score versions of the wild cluster linearized bootstrap using $B = 399$ bootstrap samples.

CV_1 standard errors and $\text{N}(0,1)$ critical values is always the least reliable.

We include results for $\text{N}(0,1)$ critical values because, as of Version 18, `Stata` reports P values and confidence intervals based on the $\text{N}(0,1)$ distribution for logit models, even though it reports ones based on the $t(G-1)$ distribution for linear regression models. Using standard normal critical values necessarily yields higher rejection frequencies than using $t(G-1)$ critical values, and the over-rejection caused by inappropriately using the latter is not at all negligible, especially for smaller values of G. In all the remaining experiments, we use $t(G-1)$ critical values for the asymptotic tests.

It is evident from Figure 1 that the number of treated clusters matters greatly. The second set of experiments focuses on this issue. In all cases, $G = 24$, $N = 12{,}000$, and

$k = 8$. The number of treated clusters varies between 2 and 12. The smallest value of G_1 is 2 because methods based on the cluster jackknife (including the WCLR/WCLU-S bootstraps) cannot handle the case where $G_1 = 1$, since the coefficient β_k is not identified when the single treated cluster is omitted. The largest value is $G/2 = 12$ because, with clusters treated at random, the results for $G_1 = G_1'$ and $G_1 = G - G_1'$, with $G_1' \leq G/2$, must be identical.

Figure 2 shows rejection frequencies at the 0.05 level for four of the most interesting tests as functions of G_1/G for $2 \leq G_1 \leq 12$. In Panels (a) and (b), $\gamma = 0$, so that all clusters are the same size. In Panels (c) and (d), $\gamma = 4$, so that clusters vary greatly in size. As usual, all tests perform poorly when G_1/G is small. Most of the tests over-reject in that case, but WCLR-S under-rejects. WCR bootstrap tests for linear regression models are well known to behave in the same way; see MacKinnon and Webb (2017, 2018) for an explanation. In all cases, tests based on CV_1 standard errors and the $t(23)$ distribution over-reject more than the other tests. Tests based on CV_3 standard errors always perform much better.

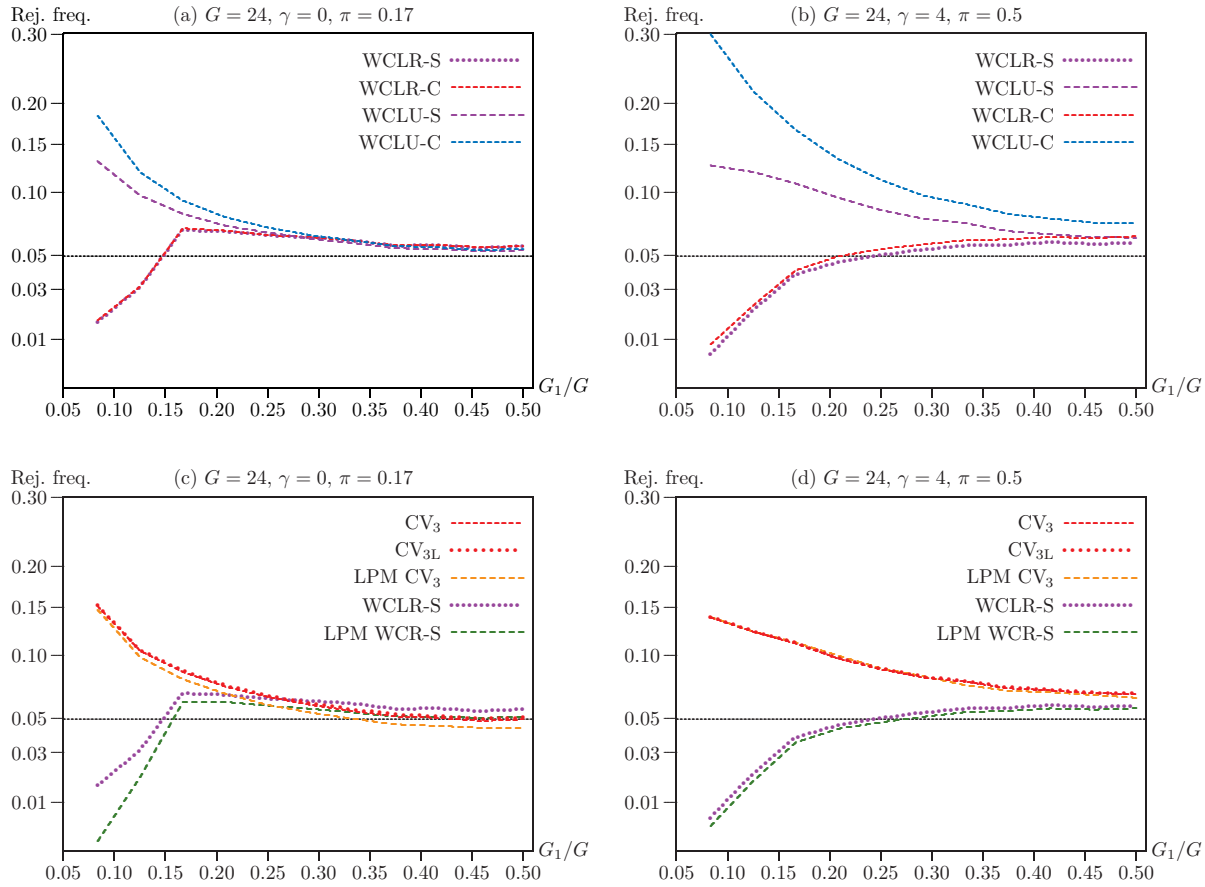
The bootstrap tests generally perform very well for larger values of G_1/G. In most cases, the WCLU-S bootstrap tests reject more often than the WCLR-S bootstrap tests. This is always true for small values of G_1/G, but it is not true in Panels (b) and (d) for large values. The WCLU-S bootstrap tests perform very similarly to the CV_3 t-tests.

Figure 3 shows rejection frequencies for some additional tests. Panels (a) and (c) are for the same case as Panel (b) in Figure 2, and Panels (b) and (d) are for the same case as Panel (c) in Figure 2. The top two panels compare WCLR/WCLU-C bootstrap tests with the WCLR/WCLU-S ones already shown in Figure 2. For the WCLR bootstraps, there are almost no differences between the two versions in Panel (a) and only small differences in Panel (b). For the WCLU bootstraps, however, the differences are quite substantial, especially in Panel (b). In that case, WCLR-S and WCLR-C are the clear winners for larger values of G_1/G, but WCLU-S also performs quite well when G_1/G is sufficiently large.

Panels (c) and (d) of Figure 3 compare the CV_3 t-test and the WCLR-S bootstrap with three other procedures. The first of these is the t-test based on $\text{CV}_{3\text{L}}$, which employs the linearized cluster jackknife variance matrix in (25). It performs almost identically to the t-test based on CV_3 in both panels. We include this test in later figures as well, because it is computationally attractive. It often performs very much like the CV_3 t-test, but not always.

The other two tests are for the linear probability model, or LPM. In Panel (d), where $\pi = 0.5$, the LPM CV_3 t-test yields results that are visually indistinguishable from those for the two t-tests of the logit model. This is not surprising, since the LPM tends to perform much like the logit model when the average value of the dependent variable is close to one-half, at least for models like (49) where all the regressors are dummy variables. The WCR-S test is almost indistinguishable from the WCLR-S test. This is also not surprising in view of

Figure 3: Rejection frequencies for tests at the 0.05 level as functions of G_1/G



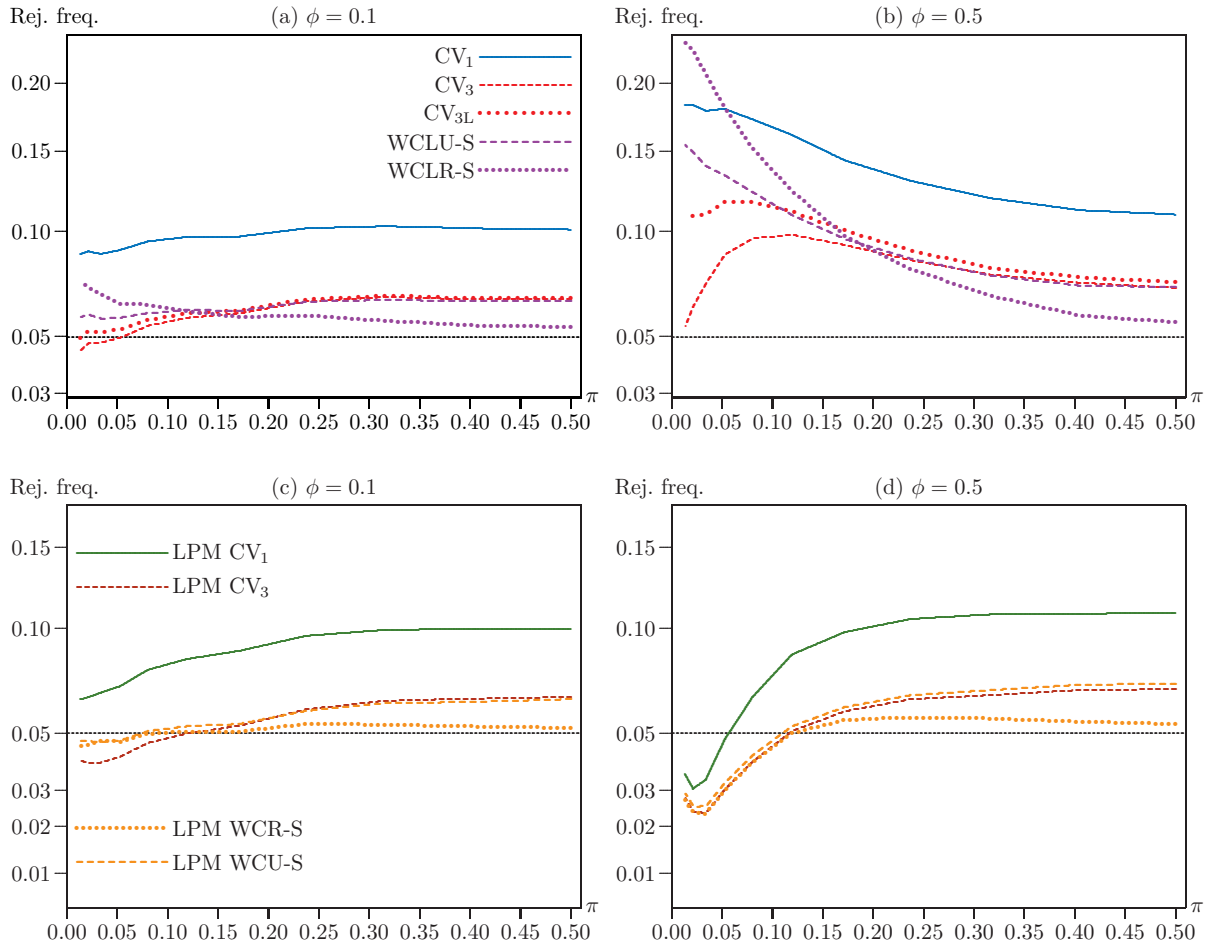
Notes: See notes to Figure 2. Panels (a) and (c) are for the same case as Panel (b) in that figure, and Panels (b) and (d) are for the same case as Panel (c). WCLR-S and WCLU-S denote tests based on symmetric P values with $B = 399$ bootstrap samples that use transformed empirical scores based on (34) and (33), respectively. WCLR-C and WCLU-C are similar, but they employ the classic versions of the wild cluster bootstrap that do not transform the empirical scores. CV_{3L} denotes a t-test based on the linearized cluster jackknife variance matrix in (25). Two tests are based on the linear probability model. "LPM CV_3" is a t-test using the CV_3 variance matrix and the $t(23)$ distribution, and "LPM WCR-S" is the restricted wild cluster bootstrap test that uses transformed empirical scores.

equations (41) and (42). To our knowledge, however, there is not at present any asymptotic theory to justify using the wild cluster bootstrap for linear probability models.

In Panel (c), where $\pi = 0.17$, the LPM CV_3 t-test always rejects less often than the other t-tests. For values of G_1/G greater than about 0.35, it actually under-rejects slightly. The WCR-S bootstrap test for the LPM rejects less often than the WCLR-S bootstrap test for the logit model, albeit to a minor extent for larger values of G_1/G. The strong performance of this test may, in part, be a consequence of the fact that all the regressors are binary.

The finite-sample properties of estimators and test statistics in binary response models

Figure 4: Rejection frequencies for tests at the 0.05 level as functions of π



Notes: See notes to Figures 2 and 3, where the notation used in Panels (a) and (b) is explained. In all panels, $N = 12{,}000$, $G = 24$, $G_1 = 7$, and $\gamma = 2$. "LPM CV_1" and "LPM CV_3" denote t-tests for the linear probability model using the $t(23)$ distribution. "LPM WCR-S" and "LPM WCU-S" denote symmetric bootstrap tests for the linear probability model based on the transformed-score versions of the wild cluster restricted and unrestricted bootstraps, respectively.

often depend on how close the average value of the dependent variable is to one-half. Therefore, in Figure 4, G_1 is fixed at 7, and the horizontal axis shows the value of π, the unconditional expectation of y_{gi}, which is varied by changing the value of β_1 in (49).

In the two left-hand panels, $\phi = 0.1$, as before. In the two right-hand panels, $\phi = 0.5$, which implies a great deal of intra-cluster correlation. The top two panels report rejection frequencies for tests of the logit model, and the bottom two panels report rejection frequencies for tests of the linear probability model. The horizontal axis stops at 0.5 because, in this model, the results for π must be the same as those for $1 - \pi$. There is no room in the figures to show results for other values of G_1. The t-tests tend to reject less often as G_1 increases,

but the pattern is more complicated for the bootstrap tests.

In Panel (a), the WCLR-S bootstrap test over-rejects less often than the other tests for $\pi > 0.15$, but for very small values of π it over-rejects more often than all but the CV_1 t-test. This pattern is much more pronounced in Panel (b), where the WCLR-S bootstrap test over-rejects severely when π is small. It is actually the worst test for $\pi \leq 0.05$. Evidently, the combination of high intra-cluster correlation and a dependent variable that is mostly 0s can lead to poor performance for all the tests, especially the bootstrap tests.

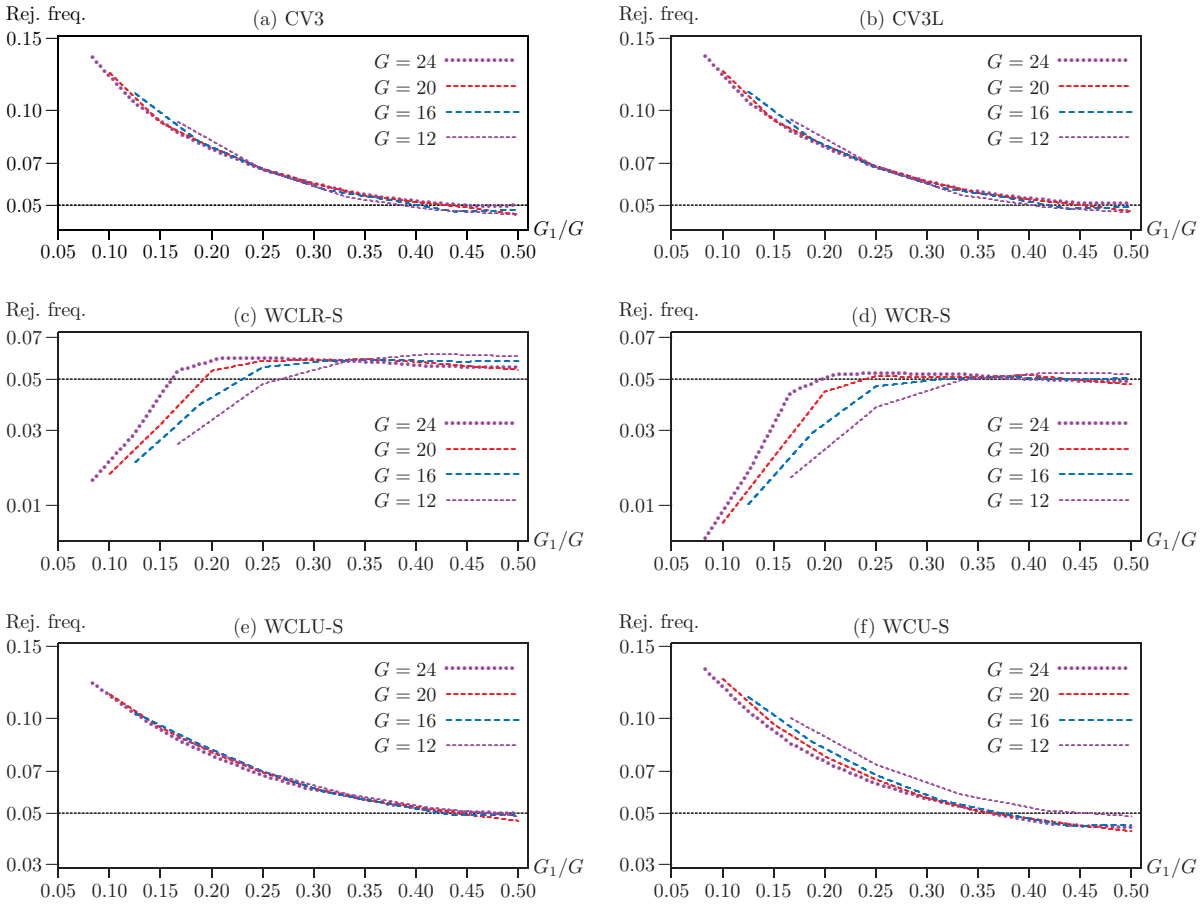
In Panel (a), the $\text{CV}_{3\text{L}}$ t-test is indistinguishable from the CV_3 t-test when π is large, but it rejects a little more often when π is small. The differences between the two tests have the same pattern in Panel (b), but they are much greater. These results suggest that the linearization (23) may not work well when π is close to 0 or 1 and there is more than a small amount of intra-cluster correlation.

Panels (c) and (d) of Figure 4 focus on tests for the linear probability model. The CV_1 t-test for the LPM in Panel (c) performs much better than the same test for the logit model in Panel (a), and the CV_3 t-test performs somewhat better for some but not all values of π. Both tests perform better for the LPM in Panel (d) than for the logit model in Panel (b), although both of them, especially CV_3, now under-reject for small values of π.

In Panel (c), the WCR-S bootstrap test performs remarkably well for all values of π. It also performs quite well in Panel (d), over-rejecting less than any of the other tests for larger values of π. Thus, the WCR-S bootstrap test seems to be the best one overall in this set of experiments, even though the data were not generated by the linear probability model. Figure 4 does not show results for the WCR-C bootstrap, but they are very similar to the ones for WCR-S.

Up to this point, all our experiments have involved 24 clusters with an average of 500 observations per cluster. In previous work (MacKinnon et al. 2023b), we have found that varying the average number of observations per cluster has almost no impact once that number is moderately large. It is more interesting to vary the number of clusters. In Figure 5, we plot rejection frequencies for six tests as functions of G_1/G for $G = 12$, 16, 20, and 24, still with $N/G = 500$. For the CV_3 and $\text{CV}_{3\text{L}}$ t-tests in Panels (a) and (b), and the WCLU-S bootstrap tests in Panel (e), the relationships between rejection frequencies and the fraction of treated clusters are almost identical for all four sample sizes. They are also quite similar for the WCU-S bootstrap tests in Panel (f). However, they are noticeably different for the WCLR-S and WCR-S tests in Panels (c) and (d). As G declines, G_1/G has to be larger for the under-rejection associated with low values of G_1/G to go away. That is because it is primarily the number of treated clusters that matters for restricted wild cluster bootstrap tests, not the fraction of treated clusters; see MacKinnon and Webb (2017, 2018).

Figure 5: Rejection frequencies for tests at the 0.05 level for different sample sizes



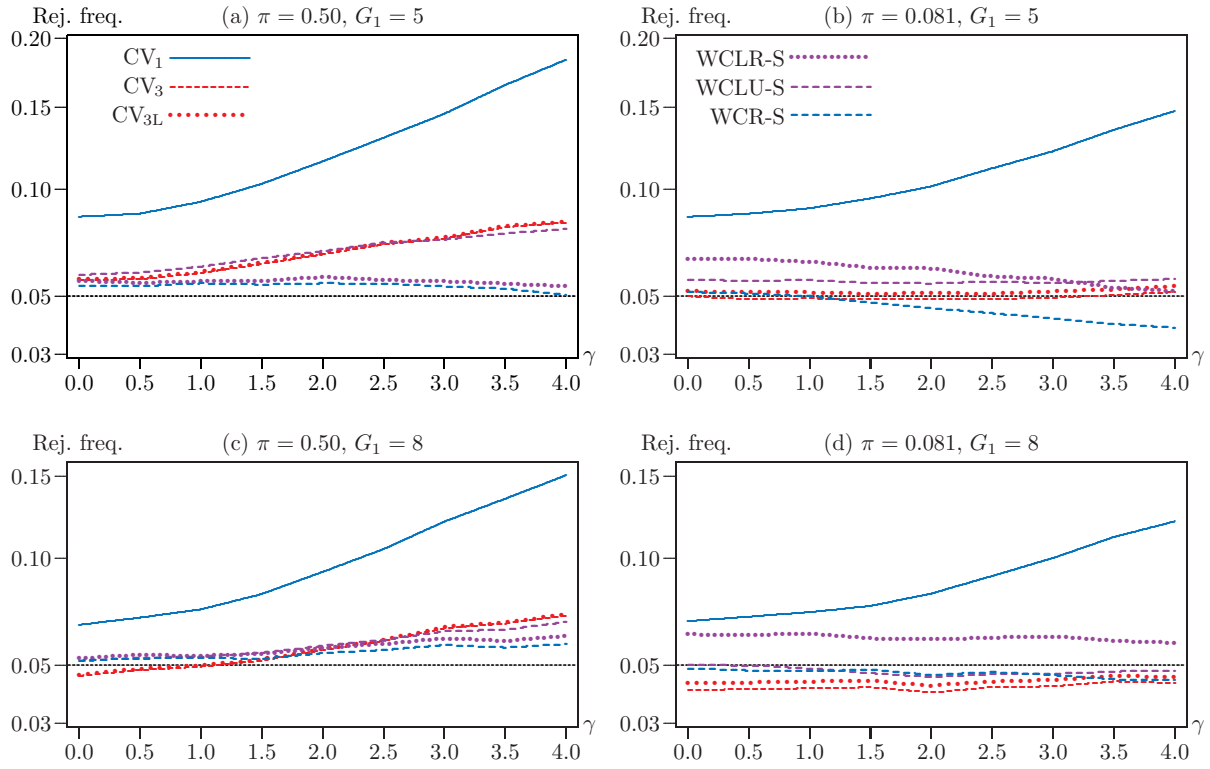
Notes: See notes to Figures 2 and 3. All panels are based on the same experiments, with $\gamma = 2$, $\phi = 0.1$, and $\pi = 0.17$. There are $G = 12$, 16, 20, or 24 clusters, with $N = 500G$ observations. Values of G_1 range from 2 to $G/2$, because the cluster jackknife cannot be computed for $G_1 = 1$.

It is hard to choose the best test in these experiments. The WCLR-S and WCR-S tests work well for a broader range of values of G_1/G than the other tests, although they under-reject when G_1/G is small. On the other hand, the two t-tests and the two unrestricted bootstrap tests all over-reject severely for small values of G_1/G, but they perform well, or at least acceptably, for large values.

The results in Figure 5 suggest that, except for small values of G_1/G, the results for $G = 16$ and $G = 24$ tend to be very similar. Since the computations for the former case are substantially less costly than for the latter, we use $G = 16$ in the next two experiments.

Figure 6 deals with the effects of cluster size variability, with γ varying between 0 (all cluster sizes equal 500) and 4 (cluster sizes vary greatly) on the horizontal axis. In Panels (a) and (c), the expectation of the dependent variable is $\pi = 0.50$, and in Panels (b) and (d) it is $\pi = 0.081$. In the top two panels, $G_1 = 5$, so that just under one-third of the clusters are

Figure 6: Rejection frequencies for tests at the 0.05 level as functions of γ



Notes: For notation, see the notes to Figures 2 and 3. There are 8,000 observations and 16 clusters, with $\phi = 0.1$. All experiments use 100,000 replications, and bootstrap tests use $B = 399$.

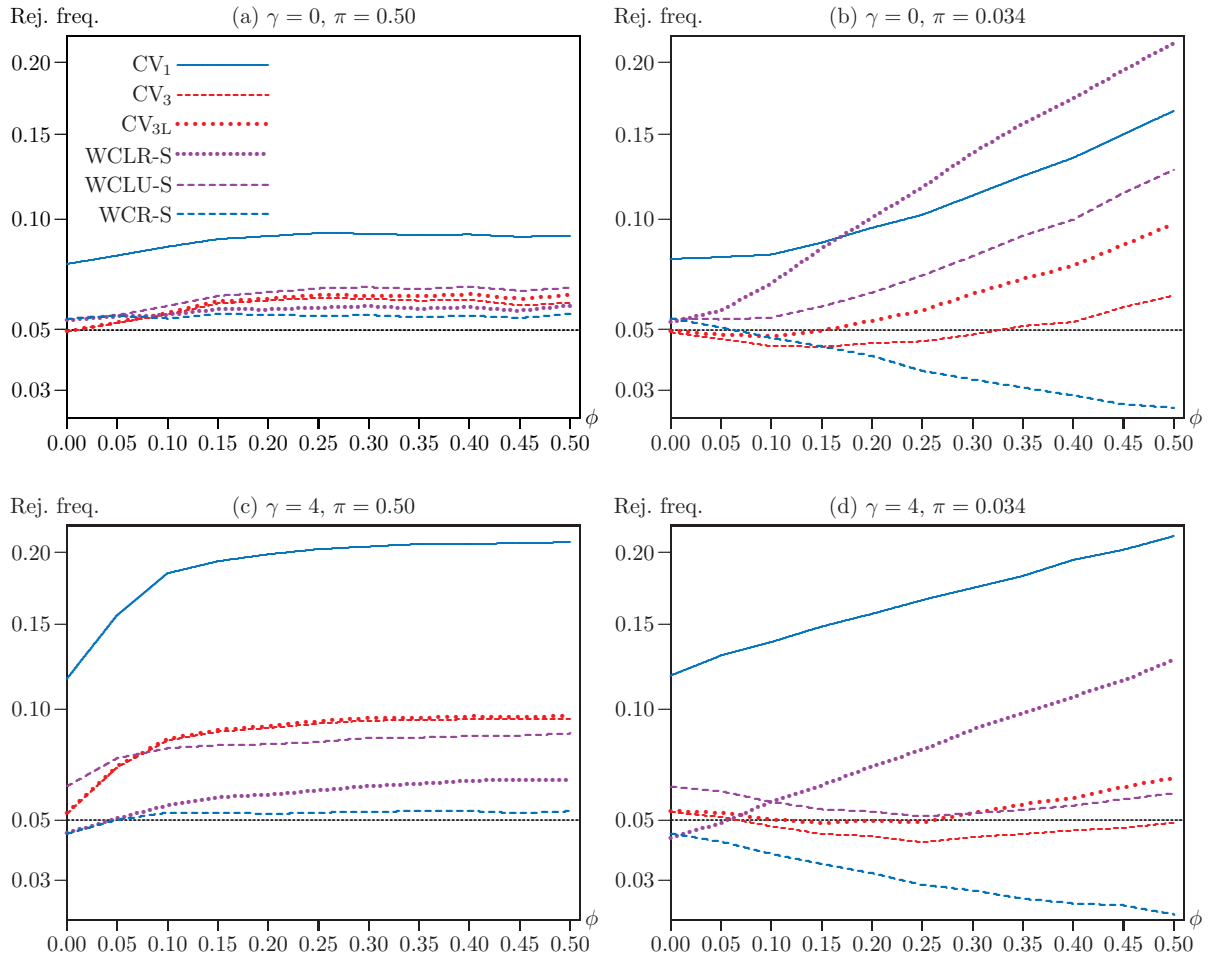
treated. In the bottom two panels, $G_1 = 8$, so that exactly half the clusters are treated.

In all cases, the CV_1 t-test rejects substantially more often than any of the other tests, and it does so to a greater extent as γ increases. In Panels (a) and (c), the CV_3 and CV_{3L} t-tests and the WCLU-S bootstrap test perform all but identically. In Panels (b) and (d), however, the WCLU-S bootstrap test rejects a bit more often than the CV_{3L} t-test which in turn rejects a bit more often than the CV_3 t-test.

The WCR-S and WCLR-S bootstrap tests perform very similarly in Panels (a) and (c), with the former rejecting a little bit less frequently than the latter. They are usually the best tests here. In Panels (b) and (d), however, the WCR-S test rejects noticeably less often than the WCLR-S test, and the latter does not perform particularly well.

Figure 7 deals with the effects of intra-cluster correlation, with ϕ varying between 0.00 and 0.50 on the horizontal axes. The top two panels set $\gamma = 0$ (every cluster has 500 observations), and the bottom two panels set $\gamma = 4$ (cluster sizes vary greatly). For the two panels on the left, the expected value of the dependent variable is 0.5, and for the two panels on the right it is 0.034. Thus Panel (a) is a case where asymptotic theory might be expected

Figure 7: Rejection frequencies for tests at the 0.05 level as functions of ϕ



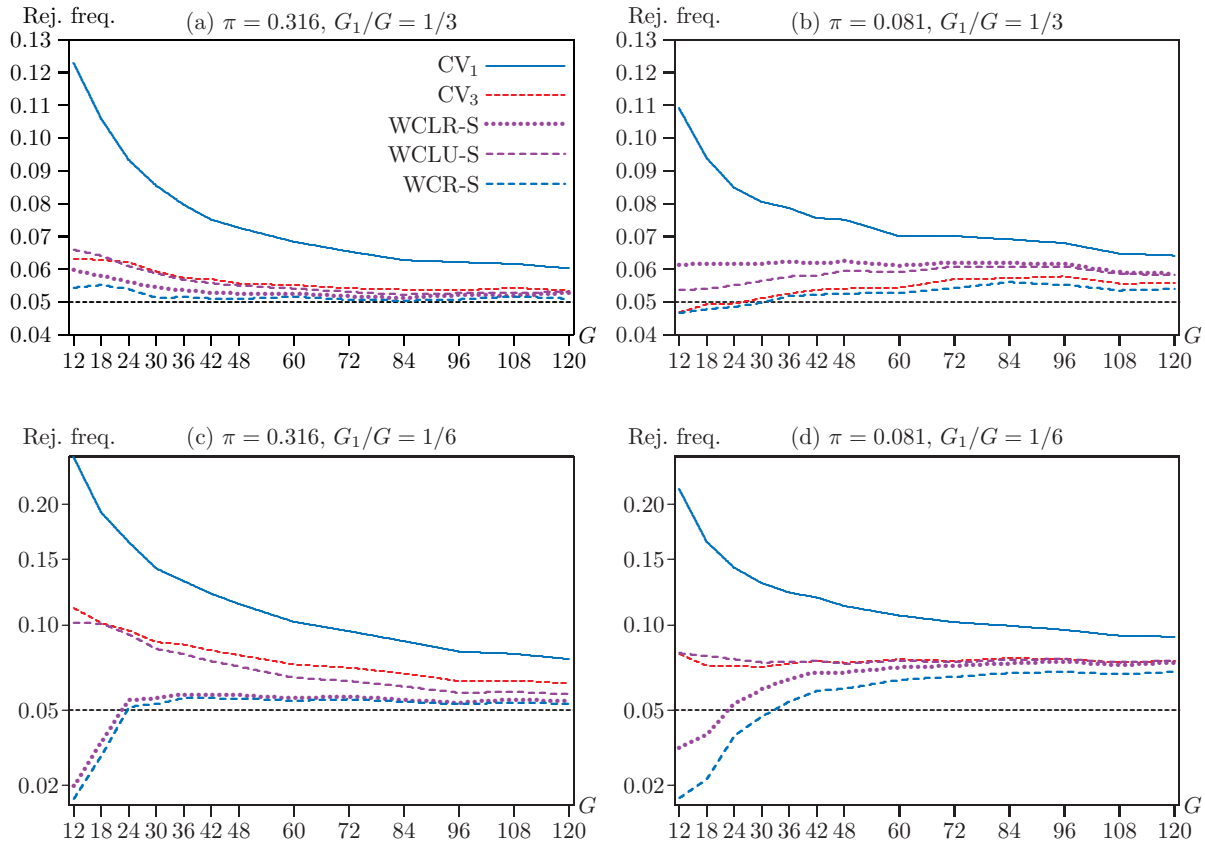
Notes: For notation, see the notes to Figures 2 and 3. There are 8,000 observations and 16 clusters, of which 5 are treated. All experiments use 100,000 replications, and bootstrap tests use $B = 399$.

to perform well, and Panel (d) is a case where it might be expected to perform poorly.

In Panel (a), which is the best case, all tests except the CV_1 t-test perform extremely well for small values of ϕ. They would have performed even better if G_1, which equals 5 in all these experiments, had been 6, 7, or 8. The performance of all tests deteriorates as ϕ increases, but the two restricted bootstrap tests always reject less than 6.0% of the time.

In Panel (b), where $\pi = 0.034$, all tests perform well when ϕ is very small, but their performance then changes dramatically as ϕ increases. Most noticeably, the WCLR-S bootstrap over-rejects more severely than any of the other tests for $\phi > 0.15$; recall Panel (b) of Figure 4. In contrast, the WCR-S bootstrap under-rejects quite substantially for larger values of ϕ. Surprisingly, the most reliable test is the CV_3 t-test, which actually under-rejects slightly for some values of ϕ and never rejects more than 6.9% of the time. The CV_{3L} t-test

Figure 8: Rejection frequencies for tests at the 0.05 level as functions of G



Notes: For notation, see the notes to Figures 2 and 3. There are between 12 and 120 clusters, with an average of 500 observations per cluster, $\gamma = 2$, and $\phi = 0.1$. All experiments use 100,000 replications, and bootstrap tests use $B = 399$.

is somewhat less reliable than the CV$_3$ t-test, and the WCLU-S bootstrap test is even worse.

In Panel (c), where cluster sizes vary greatly but $\pi = 0.5$, the relative performance of all the tests is similar to what we see in Panel (a), but their absolute performance is worse. The WCR-S bootstrap test is the best performer, never rejecting more than 5.6% of the time, followed by the WCLR-S bootstrap. The CV$_1$ t-test is the worst performer by far. The remaining three tests are hard to distinguish from each other, but they reject noticeably more often than they did in Panel (a).

We might expect Panel (d) to be the worst case, since it combines highly variable cluster sizes with a small value of $\pi = \mathrm{E}(y_{gi})$. The WCLR-S bootstrap test does indeed over-reject severely for large values of ϕ, although it is nothing like as bad as the CV$_1$ t-test. In contrast, WCR-S under-rejects even more than it did in Panel (b). However, the CV$_3$ t-test and the WCLU-S bootstraps perform fairly well and in a remarkably similar way.

Figure 8 shows what happens as the number of clusters increases from 12 to 120. There

are four cases. The fraction of treated clusters is either 1/3, in Panels (a) and (b), or 1/6, in Panels (c) and (d). The expected value of the dependent variable is moderate ($\pi = 0.316$) in Panels (a) and (b) but quite small ($\pi = 0.081$) in Panels (c) and (d). Note that the vertical axis is not transformed in Panels (a) and (b), because there was no need, but it is subjected to a square root transformation in Panels (c) and (d).

In Panel (a), which is moderate in both dimensions of the DGP, every method improves steadily (albeit sometimes slowly) as G increases. CV_3 t-tests and the three bootstrap tests always perform much better than CV_1 t-tests. The best method, by a very small margin for larger values of G, is the WCR-S bootstrap. In Panel (b), π is more extreme, but the fraction of treated clusters is still 1/3. The best-performing test is once again the WCR-S bootstrap, but by a slightly larger margin than in Panel (a). The second best performer is now the CV_3 t-test.

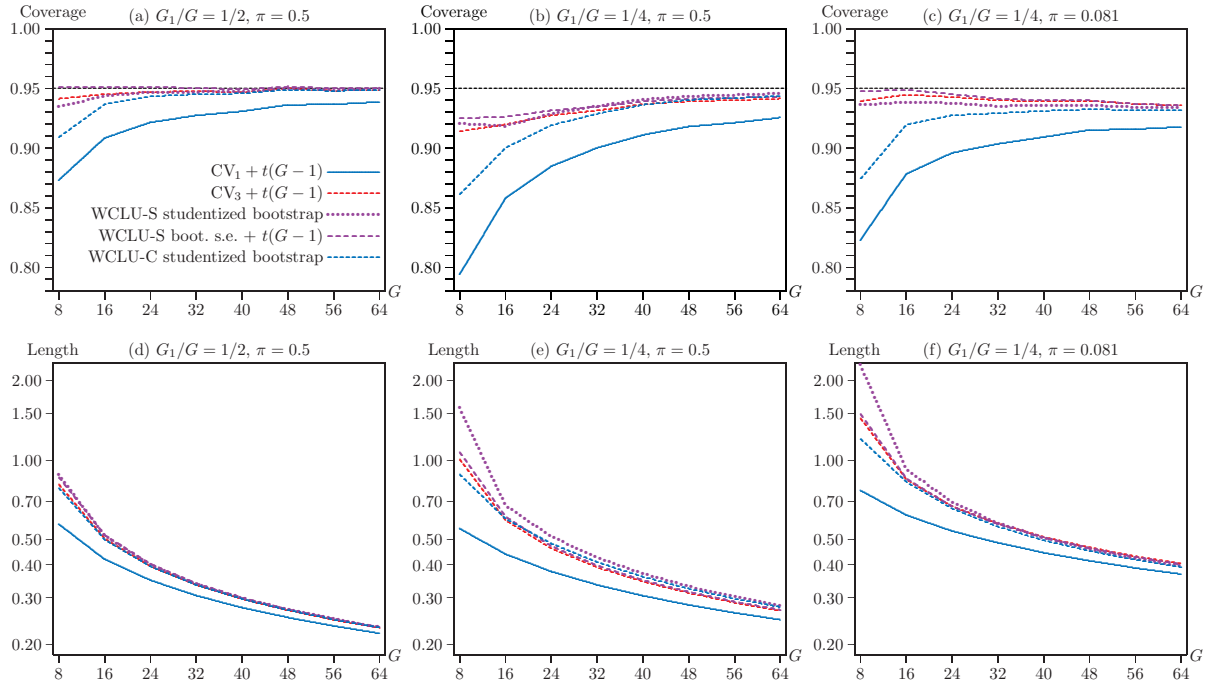
In Panels (c) and (d), where the fraction of treated clusters is only 1/6, all the methods perform considerably worse. Because $G_1 = 2$ for $G = 12$ and 3 for $G = 18$, the WCLR-S and WCR-S bootstraps under-reject in these two cases. The latter also under-rejects for $G = 24$ in Panel (d). For larger values of G, WCR-S is the best-performing test in both of the lower panels, with WCLR-S the second best.

In Panels (b) and (d), somewhat surprisingly, most of the tests do not improve much as G increases. Indeed, in Panel (d), most of them reject slightly more often for $G = 120$ than for smaller values of G. In additional experiments, not reported, we find that this phenomenon occurs whenever π is small. In some cases, rejection frequencies do not start to fall until G is rather large, perhaps 256 or more. This occurs for existing t-tests based on CV_1 as well as for the new tests we are proposing. It seems that, when π is small, the asymptotic theory on which all the tests are based may not provide a good approximation unless G is very large.

Finally, in Figure 9, we study the performance of several confidence intervals. For the reasons discussed in Section 5, they are all based on t-statistics or unrestricted bootstrap methods. The number of clusters varies from 8 to 64, and the number of observations from 4000 to 32,000. The top three panels show coverage, and the bottom three panels show average length. For the left-most panels, where $G_1/G = 1/2$ and $\pi = 0.5$, asymptotic theory should perform relatively well. For the middle panels, G_1/G is reduced from 1/2 to 1/4, and, for the right-most panels, π is in addition reduced from 0.5 to 0.081. Thus we would expect all methods to perform less well as we move from left to right.

The conventional interval based on CV_1 standard errors and $t(G-1)$ critical values always has the worst coverage. It would have performed considerably worse, especially for small values of G, if $\text{N}(0,1)$ critical values had been used, as `Stata` does. In contrast, the interval that uses CV_3 standard errors under-covers very slightly in Panels (b) and (c). Using

Figure 9: Performance of several confidence intervals as function of G



Notes: The top panels show the coverage of five different 95% confidence intervals. The bottom panels show the average lengths of the same intervals. Two of the intervals use standard errors from either CV_1 or CV_3 together with critical values from the $t(G-1)$ distribution, two are studentized bootstrap intervals based on the WCLU-C and WCLU-S bootstraps, and one uses WCLU-S bootstrap standard errors together with $t(G-1)$ critical values. There are between 8 and 64 clusters, with an average of 500 observations per cluster, $\gamma = 2$, and $\phi = 0.1$. All experiments use 100,000 replications, and the bootstrap intervals use $B = 999$.

$\mathrm{CV}_{3\mathrm{L}}$ instead of CV_3 standard errors does not affect the results to a discernible extent. The WCLU-S studentized bootstrap interval performs slightly better than the CV3 interval in all three panels. However, the WCLU-C studentized bootstrap interval under-covers noticeably for smaller values of G. In Panel (c), the three most reliable intervals actually under-cover slightly more as G increases. This is similar to what we saw in Panels (b) and (d) of Figure 8.

The last interval reported in Figure 9 computes a WCLU-S bootstrap standard error using (45) and then constructs a t-statistic, which is combined with $t(G-1)$ critical values. Surprisingly, this is often the best-performing interval, although several intervals perform nearly identically. Of course, it is not at all clear that $t(G-1)$ critical values should be used here. Interestingly, whereas bootstrap standard errors from WCLU-S seem to be quite reliable, ones from WCLU-C are always too small, leading to intervals (not shown) that cover substantially less than the WCLU-C studentized bootstrap intervals based on the same bootstrap samples. This should not have been a surprise. Using bootstrap samples based on transformed residuals is evidently more important for estimating standard errors, which

are not asymptotically pivotal, than for estimating the critical values of t-statistics, which are; see Hall (1992).

Panels (d), (e), and (f) report the average lengths of all five intervals, although it is often difficult to make out all five lines. The CV_1 intervals are noticeably shorter than the others, which is to be expected given their under-coverage. In Panels (e) and (f), the studentized WCLU-S interval is substantially longer than any of the others for smaller values of G, even though it does not have the best coverage. The interval that uses WCLU-S bootstrap standard errors is always shorter, on average, than the studentized bootstrap one, even though it has better coverage. This is probably because these standard errors are less variable than the CV_3 ones, especially for small G. In Panel (d), all intervals except the CV_1 interval are just about the same length, on average.

Taken together, our simulation results suggest that the finite-sample performance of cluster-robust tests and confidence intervals for logit models can vary greatly. Nevertheless, it seems fairly safe to draw the following conclusions:

- Conventional t-tests based on the CV_1 variance matrix and the $t(G-1)$ distribution generally over-reject, often severely, and the corresponding confidence intervals often under-cover seriously. This method cannot safely be relied upon.
- Cluster jackknife, or CV_3, t-tests always appear to be more reliable than conventional CV_1 t-tests. However, they can under-reject moderately in a few cases, and they can over-reject significantly in others, especially when the fraction of treated clusters is small, the average value of the dependent variable differs greatly from one-half, or the amount of intra-cluster correlation is large.
- Linearized cluster jackknife, or $\text{CV}_{3\text{L}}$, standard errors are much cheaper to compute than CV_3 ones. They are usually very similar, but not always.
- The WCLR-S bootstrap often performs well, but it can over-reject substantially when there is a lot of intra-cluster correlation. When its performance can be distinguished from that of the WCLR-C bootstrap, it almost always rejects less frequently.
- All methods can be somewhat unreliable when the binary outcomes are unbalanced, with most equal to either 0 or 1. This can happen even when G is quite large.
- Methods based on the linear probability model, notably the WCR-S bootstrap, can perform very well indeed. In many cases, the WCR-S and WCLR-S bootstraps yield similar results. However, they can differ greatly when there is a lot of intra-cluster correlation and the binary outcomes are unbalanced.
- The restricted bootstrap methods usually (but not always) perform better than the

unrestricted ones. However, the WCLU-S bootstrap can sometimes outperform both WCLR ones, and it often performs much better than the WCLU-C bootstrap.

- Because confidence intervals based on the WCLR bootstraps are difficult to compute, we did not study them and cannot recommend them.
- Except in the unbalanced case, confidence intervals based on CV_3 or $\text{CV}_{3\text{L}}$ standard errors and the $t(G-1)$ distribution usually perform quite well. So do intervals based on the WCLU-S bootstrap.
- If bootstrap standard errors are desired, they should always be based on the WCLU-S bootstrap. Surprisingly, it appears that confidence intervals based on these standard errors may be shorter and have slightly better coverage than studentized WCLU-S bootstrap intervals.

Since all these conclusions are based on simulation experiments, they should be interpreted with caution. They suggest that, for any empirical application, it is always informative to report the mean of the outcome variable, the number of clusters, the number of treated clusters (if the regressor of interest is a treatment dummy), and at least one measure of cluster size variability (MacKinnon et al. 2023c). All of those things affect finite-sample properties in ways that we have discussed. It may also be desirable to perform placebo regression experiments, although this may require quite bit of effort; see Bertrand, Duflo, and Mullainathan (2004), MacKinnon et al. (2023a, Section 3.5), and the next section.

7 Empirical Examples

In this section, we illustrate the tests and confidence intervals that we have discussed using two empirical examples. The first has a relatively small sample ($N = 1861$) with a moderate number of clusters ($G = 34$) and treatment at the cluster level. The second has a much larger sample ($N = 127{,}518$) with a small number of clusters ($G = 10$), a continuous explanatory variable, and cluster fixed effects.

7.1 Cash Incentives

Angrist and Lavy (2009) studies the impact of a randomized cash incentive on the outcome of a high-stakes examination. A significant sum of money was offered to "low-achieving" students in some Israeli high schools for passing the exams required to earn their high school matriculation certificate, or Bagrut. This certificate is a prerequisite for enrolling in university in Israel. Treatment was assigned randomly at the school level.

We focus on the estimates for 1861 female students who were enrolled in $G = 34$ schools in the 2001 panel of the study. These are reported in Table 2, columns 5 and 6, of the original paper. Students were offered the cash awards in $G_1 = 16$ of the schools. In addition to the treatment dummy, the equation includes nine other explanatory variables, some of which (notably, measures of past performance on examinations) have considerable explanatory power. Because treatment was at the school level, school fixed effects cannot be included.

Angrist and Lavy (2009) reports estimates for both the LPM and logit model. Our results for the former agree with the ones in the paper to the number of digits reported. Our results for the latter do not quite agree, however, because the paper reports marginal effects rather than coefficient estimates. However, the t-statistic that is implicitly reported is within the range of the ones that we obtain.

Angrist and Lavy (2009) reports CV_2 standard errors for the LPM and similar ones for the logit model. These are almost certainly more reliable than CV_1 standard errors. However, because the number of clusters is quite small, cluster sizes vary considerably (from 12 to 146), and there is quite a bit of variation in partial leverage across clusters (see notes to Table 1), CV_3 standard errors are likely to be more reliable than ones based on either CV_1 or CV_2 (MacKinnon et al. 2023b).

Table 1 reports several results for a large number of methods. Of course, we do not recommend reporting this many numbers in practice. The sixth column shows P values calculated in many different ways, and the next two columns show the lower and upper limits of 95% confidence intervals. For the LPM, all P values are less than 0.05, and all confidence intervals exclude zero. For the logit model, every P value is larger than the corresponding one for the LPM, three of them exceed 0.05, and the three confidence intervals to which the latter correspond include zero. Overall, there seems to be modest evidence against the null hypothesis, but the evidence is much less convincing than we might suppose if we simply looked at the results for either CV_1 or CV_2 and $\text{CV}_{2\text{L}}$ standard errors.

The final column of Table 1 contains rejection frequencies for a placebo regression experiment, where for each replication we add one additional regressor to the original model and test the hypothesis that the coefficient on it equals zero. The placebo regressor equals 1 for 16 randomly chosen schools and 0 for the remaining 18 schools. There are ${}_{34}\text{C}_{16} = 2,203,961,430$ ways to choose the placebo regressor. We did this 400,000 times and recorded the fraction of rejections at the 0.05 level.

As can be seen from the last column of Table 1, several methods actually under-reject, and no method over-rejects much more than 9% of the time. The methods that come very close to 0.05 are the WCR-S and WCR-C bootstraps for the LPM, and the WCLR-C, WCLU-S, and WCLU-C bootstraps for the logit model. Interestingly, t-tests based on CV_3 and $\text{CV}_{3\text{L}}$

Table 1: Effects of Cash Incentives on Passing the Bagrut

Model	Method	Coef.	Std. error	t stat.	P value	CI lower	CI upper	Placebo
LPM	CV_1	0.1047	0.0444	2.3572	0.0245	0.0143	0.1952	0.0866
LPM	CV_2	0.1047	0.0466	2.2483	0.0314	0.0100	0.1995	0.0681
LPM	CV_3	0.1047	0.0506	2.0695	0.0464	0.0018	0.2077	0.0454
LPM	WCR-C	0.1047		2.3572	0.0393	0.0055	0.2033	0.0530
LPM	WCR-S	0.1047		2.3572	0.0418	0.0042	0.2041	0.0497
LPM	WCU-C	0.1047		2.3572	0.0381	0.0064	0.2031	0.0603
LPM	WCU-C*	0.1047	0.0437	2.3982	0.0223	0.0159	0.1936	0.0918
LPM	WCU-S	0.1047		2.3572	0.0401	0.0053	0.2042	0.0555
LPM	WCU-S*	0.1047	0.0513	2.0400	0.0494	0.0003	0.2092	0.0430
Logit	CV_1	0.7164	0.3149	2.2746	0.0296	0.0756	1.3571	0.0794
Logit	CV_{2L}	0.7164	0.3303	2.1687	0.0374	0.0443	1.3884	0.0607
Logit	CV_3	0.7164	0.3609	1.9850	0.0555	−0.0179	1.4506	0.0373
Logit	CV_{3L}	0.7164	0.3592	1.9941	0.0545	−0.0145	1.4472	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	0.0151	1.4175	0.0529
Logit	WCLU-C*	0.7164	0.3095	2.3142	0.0264	0.0866	1.3461	0.0846
Logit	WCLU-S	0.7164		2.2476	0.0487	0.0042	1.4280	0.0476
Logit	WCLU-S*	0.7164	0.3645	1.9655	0.0578	−0.0251	1.4579	0.0364

Notes: There are 1861 observations and 34 clusters. The mean of the dependent variable is 0.287. The coefficient of variation of partial leverage across clusters is 0.9655. Two measures of the effective number of clusters are $G^*(0) = 24.3$ and $G^*(1) = 14.3$; see Carter et al. (2017) and MacKinnon et al. (2023c). Methods based directly on t-statistics use the $t(33)$ distribution. Bootstrap methods use the Rademacher distribution and 9,999,999 bootstrap samples so as to minimize dependence on random numbers. Methods with an asterisk employ bootstrap standard errors computed using (45) and t-statistics based on them. Methods for which no standard error is shown use symmetric bootstrap P values based on (31) and studentized bootstrap confidence intervals based on (46). Entries in the rightmost column are rejection frequencies for placebo regressions based on 400,000 replications with $B = 999$.

both under-reject somewhat. Reassuringly, the methods that over-reject most significantly are the ones that yield the smallest P values for the actual dataset. These P values should evidently not be trusted. Based on all these results, we conclude that the true P value for the hypothesis under test is probably very close to 0.05.

7.2 Tuition Fees

There is an extensive literature about the effects of college or university tuition on educational attainment. Many studies have examined the relationship between tuition and the likelihood of attending college or attaining a degree; see, for example, Heller (1999).

We examine the effects of tuition on university attendance in Canada in recent years. Specifically, we use data from the public-use version of the Labour Force Survey (LFS), combined with data on average university tuition in each province. The LFS surveys individuals once per month, and individuals are included in the survey for six months. There is much less variation in tuition fees across schools in Canada than in the United States, because (for the most part) the provinces regulate them. The tuition data come from Statistics Canada "Canadian and international tuition fees by level of study" Table 37-10-0045-01.

We use the LFS data from 2009–2019 for males aged 20 and 21 who reside in one of the ten provinces. The public-use version of the LFS does not give us the exact age of respondents, so we treat them all as being the same age. We restrict the sample to the standard Canadian university academic calendar and therefore omit responses from May through August. We estimate the following logistic regression at the individual level:

$$\Pr(\text{Student}_{ipt} = 1) = \Lambda\Big(\alpha + \beta\,\text{Tuition}_{pt} + \text{YEAR}_t + \text{PROV}_p + \boldsymbol{X}_{ipt}\boldsymbol{\gamma}\Big), \tag{51}$$

where the outcome variable Student_{ipt} equals 1 if person i in province p in year t is listed as either a part-time or full-time student. The regressor of interest is Tuition_{pt}, which is the average domestic tuition in province p in year t expressed in thousands of Canadian dollars. Because there are year fixed effects, we do not bother to convert these into constant dollars.

The row vector $\boldsymbol{X}_{ipt}$ contains two binary variables. One of these equals 1 when a person lives in any of the nine largest cities in Canada. We cannot use dummies for different large cities because each of them is located in only one province. This would make it impossible to estimate, say, the coefficient on Montreal when a jackknife sample clustering at the provincial level omits the province of Quebec. The other dummy variable in $\boldsymbol{X}_{ipt}$ indicates whether someone is a citizen/permanent resident or not. The LFS includes both permanent residents and citizens, who pay domestic tuition fees, and non-permanent residents, who pay international tuition fees. In order to minimize the number of individuals who have to pay international tuition fees, our sample excludes immigrants who have been in Canada for less than ten years. We cluster by province, because our measure of tuition is constant at the province-year level and highly persistent across years.

We initially estimated the logit model (51) and the corresponding LPM for men, women, and both together. However, we only report results for men, because they are the only ones for which the tuition variable appears to be significant using CV_1 standard errors.[1] Since our

[1]The sample of women contained 120,309 observations. The tuition coefficient was −0.0739 in the logit model, not much more than half the value of −0.1302 shown in Table 2. The CV_1 standard error was slightly larger (0.0529 instead of 0.0469), and the corresponding t-statistic was therefore much smaller (−1.3965 instead of −2.7745).

Table 2: Effects of Tuition on University Attendance

Model	Method	Coef.	Std. error	t stat.	P value	CI lower	CI upper	Placebo
LPM	CV_1	−0.0296	0.0106	−2.7899	0.0211	−0.0537	−0.0056	0.1332
LPM	CV_3	−0.0296	0.0184	−1.6120	0.1414	−0.0712	0.0120	0.0601
LPM	WCR-C	−0.0296		−2.7899	0.1414	−0.0480	0.0167	0.0658
LPM	WCR-S	−0.0296		−2.7899	0.1534	−0.0480	0.0154	0.0548
LPM	WCU-C	−0.0296		−2.7899	0.0232	−0.0543	−0.0050	0.1018
LPM	WCU-C*	−0.0296	0.0101	−2.9405	0.0165	−0.0524	−0.0068	0.1502
LPM	WCU-S	−0.0296		−2.7899	0.1018	−0.0651	0.0059	0.0747
LPM	WCU-S*	−0.0296	0.0194	1.5290	0.1606	−0.0735	0.0142	0.0508
Logit	CV_1	−0.1302	0.0469	−2.7745	0.0216	−0.2364	−0.0240	0.1298
Logit	CV_3	−0.1302	0.0799	−1.6301	0.1375	−0.3109	0.0505	0.0574
Logit	CV_{3L}	−0.1302	0.0800	−1.6280	0.1380	−0.3112	0.0507	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	−0.2362	−0.0243	0.0993
Logit	WCLU-C*	−0.1302	0.0445	−2.9244	0.0169	−0.2310	−0.0029	0.1464
Logit	WCLU-S	−0.1302		−2.7745	0.0912	−0.2634	0.0165	0.0724
Logit	WCLU-S*	−0.1302	0.0843	−1.5442	0.1569	−0.3210	0.0605	0.0485

Notes: There are 127,518 observations and 10 clusters. The mean of the dependent variable is 0.4208. The coefficient of variation of partial leverage across clusters is 1.2113, and $G^*(0) = 4.575$. Methods based directly on t-statistics use the $t(9)$ distribution. Bootstrap methods use the six-point distribution of Webb (2023) and 9,999,999 bootstrap samples so as to minimize dependence on random numbers. Methods with an asterisk employ bootstrap standard errors computed using (45) and t-statistics based on them. Methods for which no standard error is shown use symmetric bootstrap P values based on (31) and studentized bootstrap confidence intervals based on (46). Entries in the rightmost column are rejection frequencies for placebo regressions based on 400,000 replications with $B = 999$.

objective is to illustrate the consequences of using different methods of inference, we focus on the case where different methods yield different inferences. There are 127,518 observations and just ten clusters. The cluster sample sizes vary from 3,402 (P.E.I.) to 37,109 (Ontario). Thus they vary by a factor of about eleven. Note that the LFS sample sizes vary much less than actual provincial populations. For example, as of 2019-Q4, the population of Ontario was about 93 times the population of P.E.I.

Table 2 is similar to Table 1. It reports several quantities for a large number of methods. One striking feature is how much P values and confidence intervals vary across methods. Six P values are less than 0.03. These are the ones for the CV_1 t-statistics for both the LPM and logit models, for the WCU-C and WCLU-C bootstraps, and for t-statistics based on bootstrap standard errors using those two bootstrap methods. At the other extreme, all the restricted wild bootstrap methods yield P values greater than 0.135. So do t-statistics

based on both WCU-S and WCLU-S bootstrap standard errors.

With only 10 clusters that vary quite a bit in size, and substantial variation in the partial leverages, it is possible that no method is very reliable. We attempt to get a sense of which methods work best by performing a placebo regression experiment, where a placebo regressor is added to the original model. We generate artificial tuition series by using an AR(1) model, which is simulated separately for each province. The only parameter that seems to matter is the autoregressive coefficient. Reported results are for the random walk case, where this parameter equals 1. For smaller values of this parameter, rejection frequencies tended to be a little higher.

The rightmost column of Table 2 shows rejection frequencies for the coefficient on the placebo regressor based on 400,000 replications. Because of the fairly large sample size, these experiments were much more expensive than the comparable experiments in Section 7.1. Computing the CV_3 variance matrix for the logit model is by far the most costly part of the process, because it requires G additional logit estimations. In fact, calculating CV_3 takes about 70% of all the computer time for the placebo regression experiments of this section. Estimating the LPM and the original logit model and performing all the bootstrap computations, with $B = 999$, for both models takes only about 30% of the time. Remarkably, the cost of calculating $\text{CV}_{3\text{L}}$, which yields results almost identical to CV_3 here, is only about 1/41 of the cost of calculating the latter.

There is evidently a strong, inverse relationship between the placebo rejection frequencies and the reported P values. That was also the case for the example of Section 7.1. All the methods with P values less than 0.05 over-reject approximately 10–15% of the time. Conversely, the methods that perform reasonably well all yield P values greater than 0.13. The methods that perform particularly well include the WCR-S and WCLR-S bootstraps, along with t-tests based on WCU-S and WCLU-S bootstrap standard errors. The worst methods for both models are the ones that use t-tests based on either CV_1 standard errors or WCU-C and WCLU-C bootstrap standard errors. Interestingly, methods for the logit model and the LPM that are similar (e.g. WCLR-S and WCR-S) tend to perform almost the same in the placebo regressions.

We conclude that, in sharp contrast to what conventional methods of inference suggest, there seems to be very limited evidence that average tuition fees affected university attendance by men in Canada during the 2009–2019 period.

8 Concluding Remarks

In this paper, we propose several new procedures for inference in binary response models with clustered disturbances, focusing on logit models. The default settings in `R` and `Stata` use CV_1 standard errors combined with critical values from the $\text{N}(0,1)$ distribution, and our simulations show that the resulting tests can over-reject severely. Conceptually the simplest of the new procedures is to employ t-tests, or Wald tests, based on the cluster jackknife (CV_3) variance matrix, which apparently has not been studied previously in the context of binary response models, although `Stata` has been able to compute it for many years.

We also propose several new procedures based on a linear approximation to the original nonlinear model, which can be used for a wide variety of nonlinear models in addition to binary response models. The simplest procedures involve tests based on the $\text{CV}_{3\text{L}}$ variance matrix, which is just a cluster jackknife matrix for the linear approximation evaluated at the unrestricted estimates. Computing $\text{CV}_{3\text{L}}$ can be orders of magnitude cheaper than computing CV_3 when there are large numbers of observations and/or clusters. In many cases, including both of our empirical examples, the two variance matrices yield almost identical results, but they can yield noticeably different ones when the linear approximation does not work well. The same linear approximation can also be used to compute the $\text{CV}_{2\text{L}}$ variance matrix, which is analogous to the CV_2 matrix for linear regression models; see Appendix A.

The other new tests that we propose are variations of the wild cluster bootstrap. They all start with the same linear approximation as $\text{CV}_{3\text{L}}$. Conditional on it, they are computationally almost identical to corresponding variants of the wild cluster bootstrap for linear regression models. We study four bootstrap tests. Two of these, denoted WCLR, evaluate the linear approximation at restricted estimates, and the other two, denoted WCLU, evaluate it at unrestricted estimates. For each of them, the classic (or "-C") version generates bootstrap samples directly from the cluster-level empirical scores, and the score (or "-S") version generates them from empirical scores that have been transformed so as to undo some of the distortions caused by the estimation process, as proposed in MacKinnon et al. (2023b).

The WCLR/WCLU-S bootstraps employ the usual CV_1 variance matrix, not the cluster-jackknife one. It would be very much more expensive to employ the latter, and simulation results for linear models in MacKinnon et al. (2023b) suggest that, in most cases, doing so would not lead to better finite-sample properties.

Extensive simulation experiments, in Section 6, suggest that the new procedures work better, often very much better, than the conventional approach that uses CV_1 t-tests. However, which of them works best seems to vary from case to case. CV_3 and $\text{CV}_{3\text{L}}$ t-tests are always more reliable than CV_1 t-tests. In a few cases, they are actually more reliable than

some or all of the bootstrap tests. In many cases, the WCLR-S bootstrap works very well. There are a few cases in which it can perform poorly, however. This tends to happen when the fraction of 1s in the sample is very small or very large, and when there is a lot of intra-cluster correlation. In most cases, the WCR-S bootstrap for the linear probability model rejects less frequently than the WCLR-S bootstrap. The difference is often tiny, but it can sometimes be substantial, especially when the latter over-rejects noticeably.

For confidence intervals, WCLU bootstrap methods are much more convenient than WCLR ones, because there is no need to estimate the restricted logit model multiple times. The choice between WCLU-C and WCLU-S is very important, because intervals based on the latter seem to provide much better coverage with small numbers of clusters. Perhaps surprisingly, confidence intervals that combine WCLU-S standard errors with $t(G-1)$ critical values seem to work at least as well as studentized bootstrap intervals.

Appendix A: The $\mathrm{CV}_{2\mathrm{L}}$ Variance Matrix

The $\mathrm{CV}_{2\mathrm{L}}$ variance matrix can readily be computed by combining the linearization proposed in Section 3 with the procedure for calculating CV_2 given in MacKinnon et al. (2023b), which is based on an ingenious algorithm proposed in Niccodemi et al. (2020). First, form the $k \times k$ matrices

$$\boldsymbol{A}_g = (\hat{\boldsymbol{J}}^\top \hat{\boldsymbol{J}})^{-1/2} \hat{\boldsymbol{J}}_g^\top \hat{\boldsymbol{J}}_g (\hat{\boldsymbol{J}}^\top \hat{\boldsymbol{J}})^{-1/2}, \quad g = 1, \ldots, G, \tag{A.1}$$

where $\boldsymbol{J}_g(\boldsymbol{\beta})$ was defined in (22), and

$$\hat{\boldsymbol{J}} = \sum_{g=1}^{G} \hat{\boldsymbol{J}}_g = \sum_{g=1}^{G} \boldsymbol{J}_g(\hat{\boldsymbol{\beta}}) = \boldsymbol{X} \hat{\boldsymbol{\Upsilon}} \boldsymbol{X} \tag{A.2}$$

is the empirical information matrix. Then calculate the rescaled score vectors

$$\grave{\boldsymbol{s}}_g = (\hat{\boldsymbol{J}}^\top \hat{\boldsymbol{J}})^{1/2} (\mathbf{I}_k - \boldsymbol{A}_g)^{-1/2} (\hat{\boldsymbol{J}}^\top \hat{\boldsymbol{J}})^{-1/2} \hat{\boldsymbol{s}}_g, \quad g = 1, \ldots, G, \tag{A.3}$$

where $\hat{\boldsymbol{s}}_g = \boldsymbol{s}_g(\hat{\boldsymbol{\beta}})$, and $\boldsymbol{s}_g(\boldsymbol{\beta})$ was defined in (4). The variance matrix we want is then

$$\mathrm{CV}_{2\mathrm{L}}: \qquad \hat{\boldsymbol{V}}_{2\mathrm{L}}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{J}}^\top \hat{\boldsymbol{J}})^{-1} \Big(\sum_{g=1}^{G} \grave{\boldsymbol{s}}_g \grave{\boldsymbol{s}}_g^\top \Big) (\hat{\boldsymbol{J}}^\top \hat{\boldsymbol{J}})^{-1}. \tag{A.4}$$

$\mathrm{CV}_{2\mathrm{L}}$ looks very similar to $\mathrm{CV}_{1\mathcal{I}}$ given in (15). It just omits the leading scalar factor and replaces the $\hat{\boldsymbol{s}}_g$ by the $\grave{\boldsymbol{s}}_g$ given in (A.3).

Appendix B: The `logitjack` Package

We have developed a `Stata` package called `logitjack` that computes the CV_{3L} and (optionally) CV_3 variance matrices and performs the WCLR-C, WCLR-S, WCLU-C, and WCLU-S bootstraps. The latest version may be obtained from https://github.com/mattdwebb/logitjack. The data and programs used in the paper may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/logitjack/.

B.1 Syntax

The syntax for `logitjack` is

```
logitjack varlist, cluster(varname) [fevar(varlist) bootstrap nonull
reps(#) jackknife sample(string)]
```

Here `varlist` contains a list of variables. The first one is the dependent variable, the second is the regressor for which standard errors and P values are to be calculated, and the remaining ones are all the other continuous and binary regressors. Categorical variables to be treated as fixed effects should be listed using the `fevar` option.

`cluster(varname)` is mandatory, where `varname` is the name of the variable by which the observations are clustered. For every observation, it should equal one of G positive integers.

`fevar(varlist)`. Categorical variables to be included in the model as fixed effects should be listed here. They are handled equivalently to `i.varlist` in a logit model. Since this option uses a generalized inverse, CV_3 can be calculated even when some of the omit-one-cluster subsamples are singular. This always happens with cluster-level fixed effects. In contrast, the `Stata` command `jackknife: logit y x i.clustervar, cluster(clustervar)` is unable to estimate CV_3. It drops every subsample because each contains a different fixed effect which is not estimable.

`bootstrap` requests that bootstrap P values be computed. The default number of bootstraps is 999. This can be changed using the `reps`(#) option. The weight distribution used depends on the number of clusters. When there are 13 or more clusters, Rademacher weights are used. When there are 12 or fewer clusters, Webb (2023) weights are used. This option requests restricted versions of the wild cluster bootstrap. The `nonull` option instead requests unrestricted versions.

`nonull` specifies that the bootstrap DGP should be unrestricted. When it is specified, the package displays bootstrap standard errors, confidence intervals, and P values, based on both the WCLU-C and WCLU-S bootstraps. This option has the same effect whether it is used alone or in addition to the `bootstrap` option.

`reps`(#) allows the number of bootstrap replications to be specified. When it is not invoked, the `bootstrap` and `nonull` options both default to 999 replications. If this option is invoked in isolation, then restricted versions of the bootstrap are calculated, as if `boot` had been specified without `nonull`.

`jackknife` requests calculation of the CV_3 standard error. This is an option because CV_3 is relatively expensive. The CV_1 and CV_{3L} standard errors are always calculated. This option is useful when CV_3 is desired but the inclusion of cluster-level fixed effects causes issues for `Stata`'s `jackknife` prefix.

`sample(string)` limits the sample. Use the text you would enter after an "if" in a regression command. For instance, `sample(female==1)` is equivalent to "`if female==1`."

B.2 Illustration

In the remainder of this appendix, we illustrate the use of `logitjack` with an example that employs the `webuse` dataset `nlswork`. The objective is to predict whether a person is a college graduate. The variable of interest is a dummy variable indicating that the person is from a southern state. There is clustering by industry, with just twelve industries.

The first commands load and clean the dataset.

```
webuse nlswork, clear
gen age2 = age*age
drop if race==3
drop if inlist(ind,41,54)
gen white = race==1
```

For comparison purposes, the native Stata logit estimate is obtained from the command

```
logit collgrad south msp white union ln_wage age age2 i.ind, cluster(ind)
```

It yields the results

```
Logistic regression                                     Number of obs = 18,919
Wald chi2(7)  =      .
Prob > chi2   =      .
Log pseudolikelihood = -6873.2595                       Pseudo R2     = 0.2622
```

```
(Std. err. adjusted for 12 clusters in ind_code)
------------------------------------------------------------------------------
             |               Robust
    collgrad | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       south |   .3468109   .1905475     1.82   0.069    -.0266554    .7202773
```

The simplest `logitjack` command for this model is

```
logitjack collgrad south msp white union ln_wage, cluster(ind) fevar(ind)
```

The resulting output is:

```
Jackknife cluster statistics for binary response models.
Estimates for south when clustered by ind_code.
There are 18919 observations within 12 ind_code clusters.
Logistic Regression Output

  s.e. |      Coeff   Sd. Err.   t-stat  P value    CI-lower    CI-upper
-------+------------------------------------------------------------------
   CV1 |   0.346811   0.190638   1.8192   0.0962   -0.072781    0.766403
  CV3L |   0.346811   0.303466   1.1428   0.2774   -0.321113    1.014735
------------------------------------------------------------------------------

Cluster Variability

 Statistic |       Ng    Lin beta no g
-----------+--------------------------
       min |    38.00         0.050280
        q1 |   153.50         0.333767
    median |   987.00         0.356937
      mean |  1576.58         0.336269
        q3 |  2318.00         0.376996
       max |  6247.00         0.433176
-----------+--------------------------
   coefvar |     1.19         0.282305
```

Adding the jackknife option adds an additional row to the first table and an additional column to the second.

```
logitjack collgrad south msp white union ln_wage, cluster(ind) fevar(ind) jack
```

```
Logistic Regression Output

  s.e. |      Coeff   Sd. Err.   t-stat  P value    CI-lower    CI-upper
-------+------------------------------------------------------------------
   CV1 |   0.346811   0.190638   1.8192   0.0962   -0.072781    0.766403
   CV3 |   0.346811   0.295580   1.1733   0.2654   -0.303757    0.997379
  CV3L |   0.346811   0.303466   1.1428   0.2774   -0.321113    1.014735
-------------------------------------------------------------------------

Cluster Variability

 Statistic |       Ng      Lin beta no g   beta no g
-----------+-----------------------------------------
       min |    38.00           0.050280    0.059133
        q1 |   153.50           0.333767    0.333777
    median |   987.00           0.356937    0.356958
      mean |  1576.58           0.336269    0.337106
        q3 |  2318.00           0.376996    0.377489
       max |  6247.00           0.433176    0.432746
-----------+-----------------------------------------
   coefvar |     1.19           0.282305    0.274484
```

The next command calculates restricted wild bootstrap P values with the default number of replications.

```
logitjack collgrad south msp white union ln_wage, cluster(ind) fevar(ind) boot

Restricted Bootstrapped Linearized Regression Output

      WCLR |      Coeff   Sd. Err.   t-stat  P value
-----------+-----------------------------------------
   CLASSIC |   0.346811   0.190638   1.8192   0.4565
     SCORE |   0.346811   0.190638   1.8192   0.3774
-------------------------------------------------------
P-values calculated with 999 replications and Webb weights.
```

The following command is essentially the same as the last one, but it specifies an alternate number of replications.

```
logitjack collgrad south msp white union ln_wage, cluster(ind)///
fevar(ind) reps(1999)

Restricted Bootstrapped Linearized Regression Output

      WCLR |      Coeff   Sd. Err.   t-stat  P value
-----------+------------------------------------------
   CLASSIC |   0.346811   0.190638   1.8192   0.4777
     SCORE |   0.346811   0.190638   1.8192   0.4122
-------------------------------------------------------
P-values calculated with 1999 replications and Webb weights.
```

The next command estimates unrestricted wild bootstrap P values and confidence intervals with the default number of replications.

```
logitjack collgrad south msp white union ln_wage, cluster(ind) ///
fevar(ind) nonull

Unrestricted Bootstrapped Linearized Regression Output

      WCLU |      Coeff   Sd. Err.   t-stat  P value
-----------+------------------------------------------
   CLASSIC |   0.346811   0.190638   1.8192   0.3323
     SCORE |   0.346811   0.190638   1.8192   0.3854
-------------------------------------------------------
P-values calculated with 999 replications and Webb weights.

Unrestricted Bootstrapped Confidence Intervals

          WCLU |      Coeff    std.er.      WCLU CI-low       WCLU CI-up
---------------+-----------------------------------------------------------
CLASSIC-CV1-se |   0.346811   0.190638          -0.4316           1.1428
 CLASSIC-WB-se |   0.346811   0.183550          -0.0572           0.7508
---------------+-----------------------------------------------------------
  SCORE-CV1-se |   0.346811   0.190638          -0.5141           1.2153
   SCORE-WB-se |   0.346811   0.316932          -0.3508           1.0444
-------------------------------------------------------------------------
```

In this example, the default P value from native `Stata`, using the $\mathrm{N}(0,1)$ distribution, is 0.069. Because G is only 12 and cluster sizes vary greatly, this is much too small. Using any of the procedures described in this paper changes inferences noticeably. For instance, the $\mathrm{CV}_{3\mathrm{L}}$ and CV_3 P values are both over 0.25, and the bootstrap P values are all above 0.30.

References

Angrist, J. and V. Lavy (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review 99*, 1384–1414.

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

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*, 249–275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*, 137–151.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–427.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*, 317–372.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017). Asymptotic behavior of a t test robust to cluster heterogeneity. *Review of Economics and Statistics 99*, 698–709.

Conley, T. G., S. Gonçalves, and C. B. Hansen (2018). Inference with dependent data in accounting and finance applications. *Journal of Accounting Research 56*, 1139–1203.

Davidson, R. and J. G. MacKinnon (1984). A new form of the information matrix test. *Journal of Econometrics 25*, 241–262.

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*. New York: Oxford University Press.

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

Esarey, J. and A. Menger (2019). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods 7*, 541–559.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hansen, B. E. (2023). Jackknife standard errors for clustered regression. Working paper, University of Wisconsin.

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of*

Econometrics 210, 268–290.

Heller, D. E. (1999). The effects of tuition and state financial aid on public college enrollment. *Review of Higher Education 23*, 65–89.

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

MacKinnon, J. G. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics 52*, 851–881.

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

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

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

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

MacKinnon, J. G. and A. A. Smith (1998). Approximate bias correction in econometrics. *Journal of Econometrics 85*, 205–230.

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

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

MacKinnon, J. G. and M. D. Webb (2020). Clustering methods for statistical inference. In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*. Cham, Switzerland: Springer.

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

Niccodemi, G., R. Alessie, V. Angelini, J. Mierau, and T. Wansbeek (2020). Refining clustered standard errors with few clusters. Working Paper 2020002-EEF, University of Groningen.

Roodman, D., J. G. MacKinnon, M. Ø. Nielsen, and M. D. Webb (2019). Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal 19*, 4–60.

Webb, M. D. (2023). Reworking wild bootstrap-based inference for clustered errors. *Canadian Journal of Economics 56*, 839–858.