Research Article

Yong Cai, Ivan A. Canay*, Deborah Kim and Azeem M. Shaikh

On the Implementation of Approximate Randomization Tests in Linear Models with a Small Number of Clusters

https://doi.org/10.1515/jem-2021-0030
Received October 7, 2021; accepted March 14, 2022; published online April 14, 2022

Abstract: This paper provides a user's guide to the general theory of approximate randomization tests developed in Canay, Romano, and Shaikh (2017a. "Randomization Tests under an Approximate Symmetry Assumption." Econometrica 85 (3): 1013–30) when specialized to linear regressions with clustered data. An important feature of the methodology is that it applies to settings in which the number of clusters is small – even as small as five. We provide a step-by-step algorithmic description of how to implement the test and construct confidence intervals for the parameter of interest. In doing so, we additionally present three novel results concerning the methodology: we show that the method admits an equivalent implementation based on weighted scores; we show the test and confidence intervals are invariant to whether the test statistic is studentized or not; and we prove convexity of the confidence intervals for scalar parameters. We also articulate the main requirements underlying the test, emphasizing in particular common pitfalls that researchers may encounter. Finally, we illustrate the use of the methodology with two applications that further illuminate these points: one to a linear regression with clustered data based on Meng, Qian, and Yared (2015. "The Institutional Causes of china's Great Famine, 1959–1961." The Review of Economic Studies 82 (4): 1568–611) and a second to a linear regression with temporally dependent data based on Munyo and Rossi (2015. "First-day Criminal Recidivism." Journal of Public Economics 124: 81–90). The companion `R` and `Stata` packages facilitate the implementation of the methodology and the replication of the empirical exercises.

Keywords: randomization tests, linear regression, clustered data, time series

JEL Classification: C12, C14

1 Introduction

This paper provides a user's guide to the general theory of approximate randomization tests (ARTs) developed in Canay, Romano, and Shaikh (2017a) when specialized to linear regressions with clustered data. Here, clustered data refers to data that may be grouped so that there may be dependence within each cluster, but distinct clusters are approximately independent in a way to be made precise below. Such data is remarkably common, including not only data that are naturally grouped into clusters, such as villages or repeated observations over time on individual units, but also data with weak temporal dependence, in which pseudo-clusters may be formed using blocks of consecutive observations. An important feature of the methodology

***Corresponding author: Ivan A. Canay**, Department of Economics, Northwestern University, Evanston, USA, E-mail: iacanay@northwestern.edu
Yong Cai and Deborah Kim, Department of Economics, Northwestern University, Evanston, USA, E-mail: yongcai2023@u.northwestern.edu (Y. Cai), deborahkim@u.northwestern.edu (D. Kim)
Azeem M. Shaikh, Department of Economics, University of Chicago, Chicago, USA, E-mail: amshaikh@uchicago.edu

is that it applies to commonly encountered settings in which the number of clusters is small – even as small as five. In this respect, the proposed methodology contrasts sharply and meaningfully with many commonly employed methods for inference in such settings. We briefly elaborate on this point in our discussion of related literature below.

A principal goal of this paper is to make the general theory developed in Canay, Romano, and Shaikh (2017a) more accessible by providing a step-by-step algorithmic description of how to implement the test and construct confidence intervals for the quantity of interest in these types of settings. In order to do so, we develop three novel results concerning the methodology in Section 3. Our first result shows that what we view as the most natural implementation of the test, as described in Algorithm 2.1, is numerically equivalent to an alternative implementation based on weighted scores (see Algorithm 3.1). Our second result shows that when the parameter of interest is a scalar parameter, studentizing or not the t-statistic entering the test does not affect the results of the test or the associated confidence intervals. We therefore focus on the unstudentized statistic in Algorithm 2.1. Finally, our third result shows that the confidence sets for scalar parameters that are conceptually described by test inversion are indeed a closed interval of the real line. This further leads to a simple closed form expression for the lower and upper bound of the confidence intervals (see Algorithm 3.1). These results are new to this paper and play an important role in developing simple algorithms for the implementation of ARTs.

We additionally provide a discussion of the main requirements underlying the test in Section 4. These requirements essentially demand that the quantity of interest is suitably estimable cluster-by-cluster. As discussed further in Section 4, when this is not satisfied, a researcher need not conclude that it is not possible to exploit the results in Canay, Romano, and Shaikh (2017a). Instead, several remedies are possible, including clustering more coarsely or changing the specification to ensure that this requirement is satisfied. We provide two applications that further elucidate these points: one to a linear regression with clustered data based on Meng, Qian, and Yared (2015) and a second to a linear regression with temporally dependent data based on Munyo and Rossi (2015). The required software to replicate these empirical exercises and to aid researchers wishing to employ the methods elsewhere is provided in both `R` and `Stata`.[1]

The methodology described in this paper is part of a large and active literature on inference with clustered data. Following Bertrand, Duflo, and Mullainathan (2004), researchers are acutely aware of the need to adjust inferences appropriately to account for this sort of dependence. Many of the most commonly employed methods for doing so, however, are inadequate for the unusually common situation in which the number of clusters is small. Conventional wisdom suggests that the number of clusters is small when it is less than forty. For example, the method described in Liang and Zeger (1986), which has enjoyed considerable popularity due to its availability in software packages such as `Stata`, is widely acknowledged to perform poorly when this rule-of-thumb is not satisfied. Similarly, the cluster wild bootstrap described in Cameron, Gelbach, and Miller (2008) requires either a sufficiently large number of clusters or, as shown by Canay, Santos, and Shaikh (2021), stringent homogeneity across clusters, to perform reliably. As explained further in Section 4, the methods developed in Canay, Romano, and Shaikh (2017a) and described in this paper, require neither a large number of clusters nor such homogeneity across clusters. We note that the methods by Ibragimov and Müller (2010, 2016), which are closely related to the ones described here, also do not require such restrictions, but are generally less powerful and permit testing a less rich variety of hypotheses. See Canay, Romano, and Shaikh (2017a) for further discussion of these points as well as Conley, Gonçalves, and Hansen (2018) for an insightful and thorough review of the related literature more broadly.

The remainder of this paper is organized as follows. In Section 2, we first formalize the setting and establish some notation. We then describe the implementation of approximate randomization tests (ARTs) in an algorithmic fashion, including how to use these tests to construct confidence intervals for the quantity of interest. In Section 3 we present three results that play an important role in developing these algorithms. In Section 4, we articulate the main requirements underlying the tests and discuss remedies for cases where

1 The `Stata` and `R` packages `ARTs` can be downloaded from http://sites.northwestern.edu/iac879/software/.

these requirements are not satisfied. Our two empirical applications are contained in Section 5. Finally, we provide some concluding remarks in Section 6.

2 Review of ARTs in Regression Models

We start by reviewing the inference approach proposed by Canay, Romano, and Shaikh (2017a) in the context of a linear regression model with clustered data. In order to do so, we index clusters by $j \in J \equiv \{1, \dots, q\}$ and units in the jth cluster by $i \in I_{n,j} \equiv \{1, \dots, n_j\}$. We also denote by $n = \sum_{j=1}^{q} n_j$ the total number of observations. The observed data consists of an outcome of interest, $Y_{i,j}$, and a vector of covariates, $Z_{i,j} \in \mathbf{R}^{d_z}$, that are related through the equation

$$Y_{i,j} = Z_{i,j}'\beta + \epsilon_{i,j}, \tag{1}$$

where $\beta \in \mathbf{R}^{d_z}$ are unknown parameters and our requirements on $\epsilon_{i,j}$ are explained below in Section 4. Our goal is to test

$$H_0 : c'\beta = \lambda \quad \text{vs.} \quad H_1 : c'\beta \neq \lambda, \tag{2}$$

for given values of $c \in \mathbf{R}^{d_z}$ and $\lambda \in \mathbf{R}$, at level $\alpha \in (0, 1)$. An important special case of this framework is a test of the null hypothesis that a particular component of β equals a given value, i.e.

$$H_0 : \beta_\ell = \lambda \quad \text{vs.} \quad H_1 : \beta_\ell \neq \lambda,$$

for some $\ell \in \{1, \dots, d_z\}$, by simply setting c to be a standard unit vector with a one in the ℓth component and zeros otherwise. More generally, the approach we describe below extends immediately to the case where the hypothesis of interest involves multiple elements of β, in which case the test becomes

$$H_0 : R\beta = \Lambda \quad \text{vs.} \quad H_1 : R\beta \neq \Lambda, \tag{3}$$

for a given $p \times d_z$-dimensional matrix R and p-dimensional vector Λ, at level $\alpha \in (0, 1)$.

ARTs were developed more generally in Canay, Romano, and Shaikh (2017a) and admit a variety of different applications that go beyond the linear model considered here. For example, the method accommodates non-linear models, non-linear hypotheses, or even applications that go beyond inference with a small number of clusters (e.g. Canay and Kamat (2018) develop a variation that applies to inference in the regression discontinuity design). Here, we abstract away from the generality of the method and focus on the steps needed to use ARTs to test the null hypothesis in (2) in the context of the model in (1).

2.1 How to Implement ARTs

The most straightforward way to test the hypotheses in (2) via ARTs is by following the steps described in Algorithm 2.1 below.

Algorithm 2.1. (ARTs via within-cluster estimates). This implementation of ARTs involves the following steps:

Step 1: For each cluster $j \in J$, run an ordinary least squares regression of $Y_{i,j}$ on $Z_{i,j}$ using the n_j observations in cluster j. Denote the corresponding estimators of β by

$$\{\hat{\beta}_{n,j} : j \in J\}.$$

Step 2: For each $j \in J$, define the random variables

$$S_{n,j} \equiv \sqrt{n_j}\left(c'\hat{\beta}_{n,j} - \lambda\right), \tag{4}$$

and then construct the test statistic

$$T_n = \left| \frac{1}{q} \sum_{j=1}^{q} S_{n,j} \right|. \tag{5}$$

Step 3: Let $\mathbf{G} = \{1, -1\}^q$, so $g = (g_1, \ldots, g_q) \in \mathbf{G}$ is simply a q-dimensional vector with elements g_j being either 1 or -1. For any element $g \in \mathbf{G}$, define

$$T_n(g) = \left| \frac{1}{q} \sum_{j=1}^{q} g_j S_{n,j} \right|. \tag{6}$$

Step 4: Compute the $1 - \alpha$ quantile of $\{T_n(g) : g \in \mathbf{G}\}$ as

$$\hat{c}_n(1-\alpha) \equiv \inf \left\{ u \in \mathbf{R} : \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{T_n(g) \leq u\} \geq 1 - \alpha \right\}. \tag{7}$$

Step 5: Compute the test as

$$\phi_n \equiv I\{T_n > \hat{c}_n(1-\alpha)\}, \tag{8}$$

where T_n is as in (5) and $\hat{c}_n(1-\alpha)$ is as in (7). The associated p-value is

$$\hat{p}_n \equiv \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{T_n(g) \geq T_n\}, \tag{9}$$

where $T_n(g)$ is as in (6).

Algorithm 2.1 involves five steps that are easy to implement from a computational standpoint, but some of the steps deserve some clarification. Step 1 involves q within-cluster regressions that lead to q estimates of β. This essentially demands that the parameter β is identified cluster-by-cluster, and may fail to hold if some of the variables in the vector $Z_{i,j}$ are constant within cluster. We discuss possible remedies for this problem in Section 4 and illustrate their use in one of the applications in Section 5. An important feature of the method is that from Step 2 onwards, the original data is no longer needed as all the calculations only involve the q estimators of the parameter β obtained in Step 1.

Step 2 defines a type of unstudentized t-statistic that is appropriate for the null hypothesis in (2). We discuss the connection to its studentized version in Section 3.2 below. If the null hypothesis of interest is the one in (3), then a Wald-type test statistic could be used instead, i.e.

$$T_n^{\text{wald}} \equiv q \left(\frac{1}{q} \sum_{j=1}^{q} S_{n,j} \right)' \Sigma_S^{-1} \left(\frac{1}{q} \sum_{j=1}^{q} S_{n,j} \right), \tag{10}$$

where

$$S_{n,j} \equiv \sqrt{n}(R\hat{\beta}_{n,j} - \Lambda) \quad \text{and} \quad \Sigma_S \equiv \frac{1}{q} \sum_{j=1}^{q} S_{n,j} S_{n,j}'.$$

Step 3 does not require one to recompute the estimates of β. It rather uses the q estimates from Step 1 and applies sign changes to the q-dimensional vector $\{S_{n,j} : j \in J\}$. Since the cardinality of $\mathbf{G}$ is $|\mathbf{G}| = 2^q$, it exceeds 2000 when $q > 10$ and in such cases it may be convenient to use a stochastic approximation. This may be done while still controlling the rejection probability under the null hypothesis (see Canay, Romano, and Shaikh 2017a, Remark 2.2). Formally, in this case we let

$$\hat{\mathbf{G}} \equiv \{g^1, \ldots, g^B\}, \tag{11}$$

where $g^1 = \iota \equiv (1, \ldots, 1)$ is the identity vector and $g^b = \left(g_1^b, \ldots, g_q^b\right)$, for $b = 2, \ldots, B$, are i.i.d. Rademacher random variables; i.e. each g_j^b equals ± 1 with equal probability. To retain validity of the test regardless of

the value of B, we require that $g^1 = \iota$. We note, however, that the power of the test may still depend on B. For this reason, we implement Algorithm 2.1 with $\hat{\mathbf{G}}$ replacing $\mathbf{G}$ everywhere and set $B = 1000$ (or any other reasonably large number chosen by the analyst).

Step 4 requires computing the $1-\alpha$ quantile of $\{T_n(g) : g \in \mathbf{G}\}$, which can be typically obtained by sorting the values of $\{T_n(g) : g \in \mathbf{G}\}$ and then taking the $\lceil |\mathbf{G}|(1-\alpha)\rceil^{\text{th}}$ highest element in the ordered list. Thus, if we denote the ordered values of $\{T_n(g) : g \in \mathbf{G}\}$ by

$$T_n^{(1)} \leq T_n^{(2)} \leq \ldots \leq T_n^{(B)},$$

then we may define $\hat{c}_n(1-\alpha)$ in (7) as $\hat{c}_n(1-\alpha) = T^{(\lceil |\mathbf{G}|(1-\alpha)\rceil)}$. This representation suggests that the test may have trivial power for very low values of q. For example, when $\alpha = 10\%$, this problem arises if $q \leq 4$. For $q = 5$ the test already has non-trivial power and is only slightly conservative under the null. Similarly, when $\alpha = 5\%$ the test has non-trivial power for any $q \geq 6$.

Step 5 is straightforward and it provides both the test ϕ_n and the p-value $\hat{p}_n$. Each of these correspond to the non-randomized version of ARTs as opposed to their randomized counterparts (see Remark 2.4 in Canay, Romano, and Shaikh 2017a) since practitioners often prefer tests that do not involve exogenous randomness. In any case, the differences between the randomized and non-randomized versions of the test have been found to be minimal in simulations (see, e.g. Canay, Romano, and Shaikh 2017a).

2.2 How to Compute Confidence Intervals

We now discuss how to compute confidence intervals for the parameter $c'\beta$ by developing a novel algorithm that exploits the properties derived in Section 3.3. As before, a particularly important case is when c selects the ℓth component of β and then the confidence set is simply a confidence interval for β_ℓ. Conceptually we can simply form the confidence set by collecting all values of $c'\beta$ that cannot be rejected by our test at level α. That is, for the test ϕ_n in (8) we define

$$C_n = \left\{\lambda \in \mathbf{R} : \phi_n = 0 \text{ when testing } H_0 : c'\beta = \lambda\right\}. \tag{12}$$

In an asymptotic framework where $n \to \infty$ while q remains fixed, Canay, Romano, and Shaikh (2017a) show that ϕ_n is asymptotically level α under H_0. It follows from that result that, by construction, C_n covers $c'\beta$ with probability at least equal to $1-\alpha$ asymptotically. In Section 3.3 we show that C_n is indeed a closed interval in $\mathbf{R}$ and so it takes the form

$$C_n = [\lambda_l, \lambda_u], \tag{13}$$

where λ_l is the *smallest* value of λ that cannot be rejected by ϕ_n and λ_u is the *largest* value of λ that cannot be rejected by ϕ_n. The analysis in Section 3.3 also reveals that λ_l and λ_u admit simple closed-form representations that we exploit to develop Algorithm 2.2 below.

Algorithm 2.2. (ART-based confidence intervals for $c'\beta$). For $\{\hat{\beta}_{n,j} : j \in J\}$ as defined in Step 1 of Algorithm 2.1, the construction of the confidence interval involves the following steps:

Step 1: For every $g \in \mathbf{G}$, compute the following objects,

$$a(g) \equiv \frac{1}{q}\sum_{j=1}^{q}\sqrt{n}j g_j, \quad b(g) \equiv \frac{1}{q}\sum_{j=1}^{q}\sqrt{n}j g_j c'\hat{\beta}_{n,j}, \quad \text{and} \quad \lambda_0 \equiv \frac{b(\iota)}{a(\iota)}, \tag{14}$$

where $\iota = (1, \ldots, 1) \in \mathbf{G}$ is the vector with all ones.

Step 2: For every $g \in \mathbf{G}$ define

$$\lambda_l(g) \equiv \begin{cases} \frac{b(\imath)}{a(\imath)} \frac{|a(\imath)|}{|a(\imath)| + |a(g)|} + \frac{b(g)}{a(g)} \frac{|a(g)|}{|a(\imath)| + |a(g)|} & \text{if } \frac{b(g)}{a(g)} \le \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\imath)}{a(\imath)} \frac{|a(\imath)|}{|a(\imath)| - |a(g)|} - \frac{b(g)}{a(g)} \frac{|a(g)|}{|a(\imath)| - |a(g)|} & \text{if } \frac{b(g)}{a(g)} > \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\imath)}{a(\imath)} - \frac{|b(g)|}{a(\imath)} & \text{if } |a(g)| = 0 \\ -\infty & \text{if } g = \pm\imath \end{cases}. \tag{15}$$

and

$$\lambda_u(g) \equiv \begin{cases} \frac{b(\imath)}{a(\imath)} \frac{|a(\imath)|}{|a(\imath)| + |a(g)|} + \frac{b(g)}{a(g)} \frac{|a(g)|}{|a(\imath)| + |a(g)|} & \text{if } \frac{b(g)}{a(g)} \ge \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\imath)}{a(\imath)} \frac{|a(\imath)|}{|a(\imath)| - |a(g)|} - \frac{b(g)}{a(g)} \frac{|a(g)|}{|a(\imath)| - |a(g)|} & \text{if } \frac{b(g)}{a(g)} < \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\imath)}{a(\imath)} + \frac{|b(g)|}{a(\imath)} & \text{if } |a(g)| = 0 \\ +\infty & \text{if } g = \pm\imath \end{cases}. \tag{16}$$

Step 3: Compute the lower bound λ_l in the confidence interval (13) as the α quantile of $\{\lambda_l(g) : g \in \mathbf{G}\}$, i.e.

$$\lambda_l \equiv \inf\left\{ u \in \mathbf{R} : \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{\lambda_l(g) \le u\} \ge \alpha \right\}. \tag{17}$$

Compute the upper bound λ_u in the confidence interval (13) as the negative of the α quantile of $\{-\lambda_u(g) : g \in \mathbf{G}\}$, i.e.

$$\lambda_u \equiv -\inf\left\{ u \in \mathbf{R} : \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I\{-\lambda_u(g) \le u\} \ge \alpha \right\}. \tag{18}$$

Report the confidence interval C_n as in (13).

Algorithm 2.2 requires three steps that are straightforward to compute and that exploit the results in Section 3.3. We refer the reader to that section for the details on why λ_l and λ_u admit the expressions in (17) and (18), respectively.

3 Three Results on Implementation of ARTs

Before we review the main requirement underlying ARTs, we present three properties related to the implementation of ARTs that we believe practitioners should be aware of and that are novel to this paper. The first property establishes a connection between the implementation of ARTs as described in Algorithm 2.1 and an alternative implementation based on weighted scores. The second property establishes the numerical equivalence of ARTs for the null in (2) when the test statistics in (5) is replaced by its studentized version. The third and final result shows that ARTs confidence set for $c'\beta$ is indeed a closed interval in $\mathbf{R}$ and provides a representation for the upper and lower bounds of the interval that lead to Algorithm 2.2.

3.1 Equivalence with Weighted Scores

It turns out that ARTs can be implemented by an algorithm that does not involve estimating the parameter β within each cluster. This alternative algorithm involves replacing Steps 1 and 2 in Algorithm 2.1 by the two alternative steps described in Algorithm 3.1 below, while keeping Steps 3 to 5 unaffected.

Algorithm 3.1. (ARTs via within-cluster weighted scores). This implementation of ARTs involves the following steps:

Step 1′: Run a full-sample least squares regression of $Y_{i,j}$ on $Z_{i,j}$ subject to the restriction imposed by the null hypothesis, i.e. $c'\beta = \lambda$. Denote by $\hat{\epsilon}^{\mathrm{r}}_{i,j}$ the restricted residuals from this regression and by $\hat{\beta}^{\mathrm{r}}_n$ the restricted LS estimator of β.

Step 2′: For each cluster $j \in J$, define

$$S_{n,j} \equiv c'\hat{\Omega}^{-1}_{n,j}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}\hat{\epsilon}^{\mathrm{r}}_{i,j}, \tag{19}$$

where

$$\hat{\Omega}_{n,j} \equiv \frac{1}{n_j}\sum_{i\in I_{n,j}} Z_{i,j}Z'_{i,j} \tag{20}$$

is a $d_z \times d_z$ matrix that is assumed to be full rank with inverse $\hat{\Omega}^{-1}_{n,j}$.

Steps 3–5: Same as in Algorithm 2.1.

Note that Steps 3–5 remain unchanged given the alternative definition of $S_{n,j}$ in Step 2′. When it comes to Steps 1 and 2, there are two differences worth discussing. The first difference is that Step 1′ requires a single full-sample restricted least squares estimator of β as opposed to the q cluster-by-cluster estimators in Step 1 of Algorithm 2.1. The second difference is that Step 2′ is based on within-cluster weighted scores as opposed to the centered within-cluster estimates of β in Step 2 of Algorithm 2.1. Interestingly, these two implementations are numerically equivalent and so implementing ARTs via Algorithm 2.1 or Algorithm 3.1 leads to identical results. To see this formally, it is enough to show that $S_{n,j}$ as defined in (4) and (19) are the same using the following argument. For each $j \in J$,

$$\begin{aligned}
S_{n,j} &\equiv c'\hat{\Omega}^{-1}_{n,j}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}\hat{\epsilon}^{\mathrm{r}}_{i,j} \\
&= c'\hat{\Omega}^{-1}_{n,j}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}\left(Y_{i,j} - Z'_{i,j}\hat{\beta}^{\mathrm{r}}_n\right) \\
&= c'\hat{\Omega}^{-1}_{n,j}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}Y_{i,j} - c'\hat{\Omega}^{-1}_{n,j}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}Z'_{i,j}\hat{\beta}^{\mathrm{r}}_n \\
&= \sqrt{n_j}\left(c'\hat{\beta}_{n,j} - c'\beta\right) - \sqrt{n_j}\left(c'\hat{\beta}^{\mathrm{r}}_n - c'\beta\right) \\
&= \sqrt{n_j}\left(c'\hat{\beta}_{n,j} - \lambda\right),
\end{aligned}$$

where the fourth equality follows by adding and subtracting $\sqrt{n_j}c'\beta$ and the last equality holds because $c'\hat{\beta}^{\mathrm{r}}_n = c'\beta = \lambda$ under the null hypothesis in (2). It thus follows that $S_{n,j}$ in (4) and in (19) are identical and so ARTs can be alternatively implemented via Algorithm 2.1 or 3.1. The following lemma summarizes our discussion above:

Lemma 3.1. *Let $\hat{\Omega}_{n,j}$ in (20) be full rank for each $j \in J$. Denote by C_n a confidence interval for $c'\beta$ computed using Algorithm 2.1 and by C'_n a confidence interval for $c'\beta$ computed using Algorithm 3.1. Then $C_n = C'_n$.*

3.2 Equivalence with Studentized Version of the *t*-statistic

The ART defined in (8) of Algorithm 2.1 is based on the unstudentized test statistic T_n defined in (5). It may perhaps appear more desirable to instead consider the studentized version of this test statistic as studentization commonly improves performance in a variety of other settings. Here, we prove that this is not

the case for ARTs when the null hypothesis is the one in (2) and that both versions of the test statistic lead to numerically identical results.

To see this, start by defining the studentized version of the test statistic in (5) as $T_n^s \equiv T_n^s(\iota)$, where for each $g \in \mathbf{G}$,

$$T_n^s(g) \equiv \sqrt{q}\frac{\left|\frac{1}{q}\sum_{j=1}^q g_j S_{n,j}\right|}{\hat{\sigma}_s(g)} \quad \text{and} \quad \hat{\sigma}_s(g) \equiv \sqrt{\frac{1}{q}\sum_{j=1}^q \left(g_j S_{n,j} - \frac{1}{q}\sum_{j=1}^q g_j S_{n,j}\right)^2}. \tag{21}$$

Then note that

$$\hat{\sigma}_s^2(g) = \frac{1}{q}\sum_{j=1}^q g_j^2 S_{n,j}^2 - \left(\frac{1}{q}\sum_{j=1}^q g_j S_{n,j}\right)^2 = V_n - T_n^2(g),$$

where $V_n \equiv \frac{1}{q}\sum_{j=1}^q g_j^2 S_{n,j}^2$ does not depend on g as $g_j^2 = 1$ for all $j \in J$. It follows that we can write the studentized test statistic as

$$T_n^s(g) = \sqrt{q}\frac{T_n(g)}{\sqrt{V_n - T_n^2(g)}}.$$

Since the function $x \mapsto \frac{x}{\sqrt{1-x^2}}$ is strictly increasing for $x \in [0,1)$, it follows that $T_n^s(g)$ is a strictly monotonic transformation of $T_n(g)$ for each $g \in \mathbf{G}$. We conclude that $I\{T_n(g) \geq T_n(\iota)\} = I\left\{T_n^s(g) \geq T_n^s(\iota)\right\}$ for all $g \in \mathbf{G}$ and so the ART based on $T_n(g)$ and $T_n^s(g)$ are identical. This discussion is summarized in the following lemma:

Lemma 3.2. *Let $\hat{\Omega}_{n,j}$ in (20) be full rank for each $j \in J$. Denote by C_n a confidence interval for $c'\beta$ computed using Algorithm 2.1 and by C_n' a confidence interval for $c'\beta$ computed using Algorithm 2.1 with T_n^s in place of T_n and $T_n^s(g)$ in place of $T_n(g)$. Here, $T_n^s(g)$ is given by (21) and T_n^s is understood to be $T_n^s(\iota)$, where ι is the identity transformation. Then, $C_n = C_n'$.*

3.3 Convexity of the Confidence Intervals

The ART-based confidence intervals for $c'\beta$ defined in (12) can be computed by test inversion. From a computational standpoint, however, computing confidence sets by test inversion may be cumbersome and the resulting set may not even be an interval. That is, it may not be closed and convex. In this section we prove that this is not a concern for ART-based confidence intervals for $c'\beta$ and so such confidence intervals could be easily computed by a standard bisection algorithm. In fact, our results go even further. We derive closed form expressions for the lower and upper bounds of the confidence interval that imply that computing ART-based confidence intervals for $c'\beta$ is straightforward from a computational standpoint. In order to derive these results, we slightly change our notation to make explicit the dependence on λ of each of the elements entering the test in (8). To this end, let

$$T_n(g,\lambda) \equiv \left|\frac{1}{q}\sum_{j=1}^q g_j S_{n,j}(\lambda)\right| \quad \text{where} \quad S_{n,j}(\lambda) = \sqrt{n_j}\left(c'\hat{\beta}_{n,j} - \lambda\right),$$

and note that $T_n = T_n(\iota, \lambda)$. Using this notation, we can re-write the confidence interval in (12) as

$$C_n = \left\{\lambda \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} \geq \alpha\right\},$$

which is simply the values of λ for which the p-value of the test, as defined in (9), is not below α. In order to show that this confidence set is a closed interval, we claim that the p-value

$$\hat{p}_n(\lambda) = \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} \tag{22}$$

is equal to 1 for $\lambda_0 \equiv b(\iota)/a(\iota)$, monotonically increasing for any $\lambda < \lambda_0$, and monotonically decreasing for any $\lambda > \lambda_0$. The next lemma formalizes this result.

Lemma 3.3. *Let $\hat{\Omega}_{n,j}$ in (20) be full rank for each $j \in J$. Let $a(g)$, $b(g)$, and λ_0 be defined as in (14). The p-value in (22) equals*

$$\hat{p}_n(\lambda) = \begin{cases} \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{\lambda \geq \lambda_l(g)\} & \text{for } \lambda < \lambda_0 \\ 1 & \text{for } \lambda = \lambda_0, \\ \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{\lambda \leq \lambda_u(g)\} & \text{for } \lambda > \lambda_0 \end{cases} \tag{23}$$

where $\{\lambda_u(g) : g \in \boldsymbol{G}\}$ and $\{\lambda_l(g) : g \in \boldsymbol{G}\}$ are defined in Algorithm 2.2.

Proof. It is useful to re-write $T_n(g,\lambda)$ in terms of $a(g)$ and $b(g)$. To this end, note that

$$\begin{aligned} T_n(g,\lambda) &\equiv \left|\frac{1}{q}\sum_{j=1}^{q} g_j S_{n,j}(\lambda)\right| = \left|\frac{1}{q}\sum_{j=1}^{q} g_j\sqrt{n_j}c'\hat{\beta}_{n,j} - \lambda\frac{1}{q}\sum_{j=1}^{q} g_j\sqrt{n_j}\right| \\ &= |b(g) - \lambda a(g)|. \end{aligned} \tag{24}$$

Given $g \in \mathbf{G}$ and $a(g) \neq 0$, $T_n(g,\lambda)$ is a "V-shaped" function of λ taking the value 0 at $\frac{b(g)}{a(g)}$ and with slope $-|a(g)|$ for all $\lambda < \frac{b(g)}{a(g)}$ and slope $|a(g)| \leq a(\iota)$ for all $\lambda > \frac{b(g)}{a(g)}$. Figure 1 illustrates this for three values of g.

First, note that $I\{T_n(g,\lambda_0) \geq T_n(\iota,\lambda_0)\} = I\{T_n(g,\lambda_0) \geq 0\} = 1$ for all $g \in \mathbf{G}$ and so it follows immediately that $\hat{p}_n(\lambda_0) = 1$.

Second, restrict attention to the set $\Lambda^+ \equiv \{\lambda \in \mathbf{R} : \lambda > \lambda_0\}$ where $T_n(\iota,\lambda)$ is linearly increasing. In order to prove that $\hat{p}_n(\lambda)$ takes the form in (23) we prove that $I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = I\{\lambda \leq \lambda_u(g)\}$ for each $g \in \mathbf{G}$ by dividing the argument into three cases.

Case 1: Consider $g \in \mathbf{G}$ such that $a(g) \neq 0$ and $|a(g)| \neq a(\iota)$. Since $|a(g)| < a(\iota)$, it follows that $T_n(g,\lambda)$ and $T_n(\iota,\lambda)$ intersect only once on Λ^+ and this holds regardless of whether $\frac{b(g)}{a(g)} < \lambda_0$ or $\frac{b(g)}{a(g)} \geq \lambda_0$ (see Figure 1 for a graphical illustration of each of these cases). Denote the intersection point by $\lambda_u(g)$ and note that $T_n(g,\lambda) \geq T_n(\iota,\lambda)$ for all $\lambda_0 < \lambda \leq \lambda_u(g)$ and $T_n(g,\lambda) < T_n(\iota,\lambda)$ for all $\lambda > \lambda_u(g)$. Conclude that on Λ^+,

$$I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = I\{\lambda \leq \lambda_u(g)\}. \tag{25}$$

Simple algebra shows that the intersection point $\lambda_u(g)$ takes the form in (16).

Case 2: Consider $g \in \mathbf{G}$ such that $a(g) = 0$. Note that $b(\iota) - \lambda a(\iota) < 0$ for $\lambda \in \Lambda^+$. It thus follows that for $\lambda \in \Lambda^+$,

$$I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = I\{|b(g)| \geq |b(\iota) - \lambda a(\iota)|\} = I\left\{\lambda \leq \frac{b(\iota)}{a(\iota)} + \frac{|b(g)|}{a(\iota)}\right\},$$

and so (25) holds in this case with $\lambda_u(g) = \frac{b(\iota)}{a(\iota)} + \frac{|b(g)|}{a(\iota)}$, as defined in (16).

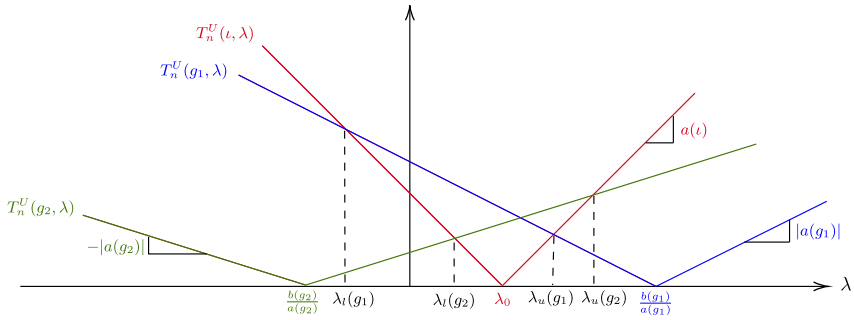

Figure 1: $T_n(g,\lambda)$ as functions of λ for $g \in \{\iota, g_1, g_2\}$.

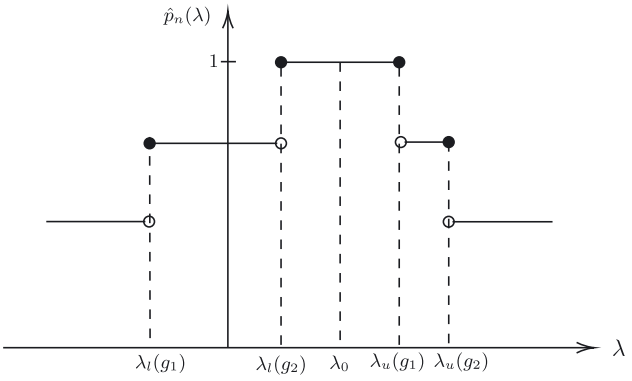

Figure 2: $\hat{p}_n(\lambda)$ as a function of λ.

Case 3: Consider $g \in \mathbf{G}$ such that $|a(g)| = a(\iota)$ and so $g = \pm\iota$. If $g = \iota$, $I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = 1$ for all $\lambda \in \mathbf{R}$. We conclude that (25) holds with $\lambda_u(g) = \infty$. If $g = -\iota$, then we have that $a(-\iota) = -a(\iota)$ and $b(-\iota) = -b(\iota)$ so that $\frac{b(-\iota)}{a(-\iota)} = \lambda_0$ and again $I\{T_n(-\iota,\lambda) \geq T_n(\iota,\lambda)\} = 1$ for all $\lambda \in \mathbf{R}$. We conclude that (25) holds with $\lambda_u(g) = \infty$, as defined in (16). This completes the proof of (23) for the case $\lambda \in \Lambda^+$.

Finally, the construction for $\lambda \in \Lambda^- \equiv \{\lambda \in \mathbf{R} : \lambda < \lambda_0\}$ parallels the one for $\lambda \in \Lambda^+$ so we omit the arguments here. Putting all the cases together, (23) follows and this completes the proof. □

Figure 2 illustrates the p-value in (23) as a function of λ for the groups in Figure 1. Since $\hat{p}_n(\lambda)$ is right continuous and increasing for $\lambda < \lambda_0$, we can define λ_l as the smallest value of λ for which $\hat{p}_n(\lambda) \geq \alpha$. Such value exists and is unique. Similar, since $\hat{p}_n(\lambda)$ is left continuous and decreasing for $\lambda > \lambda_0$, we can define λ_u as the largest value of λ for which $\hat{p}_n(\lambda) \geq \alpha$. Such value exists and is again unique. This argument leads to the representation of C_n in (13), showing that ART-based confidence intervals for $c'\beta$ are indeed intervals in $\mathbf{R}$. Furthermore, note that (23) implies that the smallest value of λ for which $\hat{p}_n(\lambda) \geq \alpha$ can be defined as

$$\inf\left\{\lambda \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{\lambda \geq \lambda_l(g)\} \geq \alpha\right\},$$

which is just the definition of the α quantile of $\lambda_l(g)$, as defined in Algorithm 2.2. A similar result holds for λ_u and so C_n can be computed in closed form by Algorithm 2.2.

4 What We Need for ARTs to Work

The main requirement underlying ARTs is Assumption 3.1 in Canay, Romano, and Shaikh (2017a). This assumption guarantees that the test delivers rejection probabilities under the null hypothesis that are close to the nominal level α in an asymptotic framework where $n \to \infty$ and q remains fixed. In the context of the linear model in (1), this translates into the following two conditions summarized in Assumption 4.1 below.

Assumption 4.1. Let $\{\hat{\beta}_{n,j} : j \in J\}$ be the cluster-by-cluster estimators of β defined in Algorithm 2.1. Assume that:

(a) $\{\hat{\beta}_{n,j} : j \in J\}$ jointly converge in distribution at some (possibly unknown) rate; i.e.

$$\begin{pmatrix} a_{n,1}(\hat{\beta}_{n,1} - \beta) \\ \vdots \\ a_{n,q}(\hat{\beta}_{n,q} - \beta) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} S_1 \\ \vdots \\ S_q \end{pmatrix} \tag{26}$$

for a sequences $a_{n,j} \to \infty$ and random variables $(S_1, \ldots, S_q)'$.

(b) The limiting random variables $(S_1, \ldots, S_q)'$ are invariant to sign changes, i.e.

$$(g_1 S_1, \ldots, g_q S_q) \stackrel{d}{=} (S_1, \ldots, S_q), \tag{27}$$

for any g in $\mathbf{G}$, where $\mathbf{G}$ is defined in Step 4 of Algorithm 2.1.

Condition (26) holds, for example, when $Z_{i,j}$ and $\epsilon_{i,j}$ are uncorrelated and the analyst assumes some form of weak dependence within clusters that permits the application of an appropriate central limit theorem. In such a case, (26) typically holds with $a_{n,j} = \sqrt{n_j}$ and each S_j being a mean-zero normal random variable. In fact, under the commonly used assumption of independent clusters, it also follows that $S_j \perp\!\!\!\perp S_{j'}$ for any $j \neq j'$. In this case the normally distributed random variables may not be identically distributed but are indeed independent. Condition (27), in turn, requires each S_j to be symmetrically distributed around zero and independent of each other. This is immediately satisfied when each S_j is a mean-zero normal random variable and clusters are independent. Importantly, these assumptions allow for the normally distributed random variables to have different variances across clusters; a type of heterogeneity not allowed by the cluster wild bootstrap approach popularized by Cameron, Gelbach, and Miller (2008) and later studied formally by Canay, Santos, and Shaikh (2021).

Remark 4.1. The asymptotic normality in (26) arises frequently in applications, but is not necessary for the validity of ARTs. All that is required is that the estimators $\{a_{n,j}(\hat{\beta}_{n,j} - \beta)\colon j \in J\}$ have a limiting distribution that is the product of q distributions that are symmetric about zero. This may even hold in cases where the estimators have infinite variances or are inconsistent. See Canay, Romano, and Shaikh (2017a, Remark 4.5) for additional discussion on this point. □

Remark 4.2. It is worthwhile to contrast the requirements of Assumption 4.1 with those of "classical" methods, such as those described in Liang and Zeger (1986). These latter methods permit arbitrary dependence within each cluster, but require the size of the clusters to be small and the number of clusters to be large. As described above, Assumption 4.1(a), on the other hand, permits the number of clusters to be small, but requires the size of the clusters to be large and weak dependence within each cluster. We emphasize, however, that these restrictions are commonly employed in establishing the validity of other methods in settings with a small number of clusters, including, for example, the t-test approach by Ibragimov and Müller (2010) and the wild bootstrap Canay, Santos, and Shaikh (2021). □

Remark 4.3. We focus our exposition on the case where $Z_{i,j}$ is exogenous but we emphasize that the conditions in (26) and (27) typically hold in instrumental variable (IV) models. Accommodating IV to ARTs then only requires modifying Step 1 in Algorithm 2.1 so that the least squares regression is replaced with the appropriate IV regression. Steps 2–6 remain unaffected. □

An implicit requirement behind ARTs that deserves further comments lies in Step 1 of Algorithm 2.1, which requires that the analyst runs cluster-by-cluster regressions. This step implicitly assumes that the parameter β is identified within each cluster. In practice, this means that the matrix $\hat{\Omega}_{n,j}$ in (20) must be invertible for each $j \in J$ and hence the same requirement applies to Algorithm 3.1. This restriction may be substantially important in some applications and so here we discuss common ways in which the problem may manifest and two alternative remedies.

One case in which running least squares cluster-by-cluster is not feasible is when the coefficient of interest is associated with a variable that only varies across clusters. For example, consider the model in (1)

and partition $Z_{i,j}$ into a constant term, a scalar variable that only varies across clusters, $Z_j^{(1)}$, and another variable that varies across and within clusters, $Z_{i,j}^{(2)}$. That is,

$$Y_{i,j} = \beta_0 + Z_j^{(1)}\beta_1 + Z_{i,j}^{(2)}\beta_2 + \epsilon_{i,j}, \tag{28}$$

where the analysts' interest lies in the coefficient β_1, i.e. $c'\beta = \beta_1$. Clearly, the regression in Step 1 of Algorithm 2.1 would not separately identify β_0 and β_1 as $Z_j^{(1)}$ is perfectly colinear with the constant term. The matrix $\hat{\Omega}_{n,j}$ in (20) is simply singular. This situation arises, for example, in the empirical application considered by Canay, Romano, and Shaikh (2017b) where $j \in J$ indexes schools and the variable of interest is a treatment indicator at the school level. A natural remedy in a situation like this is clustering more coarsely (e.g. by combining clusters) to obtain variation within the re-defined clusters. This is possible for ARTs since the validity of the method does not rely on having a large number of clusters and thus it can afford to work with coarser clustering. In fact, in certain settings combining clusters may be quite natural. For example, Canay, Romano, and Shaikh (2017b) re-defined clusters as "pairs" of schools (as opposed to just schools) given that the treatment assignment mechanism of the experiment was a matched pairs design and so the pairs used at the randomization stage represented natural groupings. In other settings where it is less clear how to group clusters, any grouping that satisfies the requisite identification condition leads to a valid test, but it may be further desirable to combine such tests to limit concerns about "data snooping" across groupings. To this end, results in DiCiccio, DiCiccio, and Romano (2020) on combining tests may be relevant.

Remark 4.4. A quick inspection of (28) may lead the analyst to believe there is a workaround that does not involve combining clusters if one instead uses some estimator of β_0 from a full sample regression. For example, the full sample least squares estimator $\hat{\beta}_{n,0}$ from the regression in (1). Then, assuming for simplicity that $Z_j^{(1)} \neq 0$ for all $j \in J$, one may consider modifying Step 1 in Algorithm 2.1 by running a regression of $Y_{i,j}$ on an intercept and $Z_{i,j}^{(2)}$ (not including $Z_j^{(1)}$) and then redefining $\{\hat{\beta}_{n,j} : j \in J\}$ as the difference between the within cluster intercept estimates, $\hat{\beta}_{j,0}$ and the full sample estimate $\hat{\beta}_{n,0}$, i.e. $\hat{\beta}_{n,j} = \hat{\beta}_{j,0} - \hat{\beta}_{n,0}$. Such strategies unfortunately introduce dependence between the q estimators of β (as they all depend on $\hat{\beta}_{n,0}$) and thus end up violating one of the two main conditions needed for ARTs to be asymptotically valid; mainly condition (27). □

Another case where the lack of identification within cluster may manifest is when the variable of interest actually varies within clusters but the model specification involves other variables that are collinear with some other variable (including the variable of interest or the constant term) within clusters. For example, consider the model in (1) where instead of individuals indexed by $i \in I_{n,j}$, units within cluster are indexed over time $t \in T$. Partition $Z_{j,t}$ into the variable of interest, $Z_{j,t}^{(1)}$, and time fixed effects δ_t. That is,

$$Y_{j,t} = Z_{j,t}^{(1)}\beta_1 + \sum_{\tilde{t}\in T} I\{\tilde{t} = t\}\delta_{\tilde{t}} + \epsilon_{j,t}. \tag{29}$$

It then follows that, within each cluster $j \in J$, the time fixed effect δ_t absorbs all the variation in $Z_{j,t}^{(1)}$ and so β_1 is not identified. In cases like this the analyst could again combine clusters to obtain variation within the re-defined clusters. An alternative remedy is to change the specification by, for example, replacing the time fixed effect with a cluster-specific time trend. Such specification is more restrictive than the time fixed effect in the sense that it imposes a linear trend but, at the same time, is more general as it allows for heterogeneity across clusters in the linear trend. We illustrate this approach in the application we consider in Section 5.1.

The need to identify β within each cluster is in our view the main limitation of ARTs, but a limitation that needs to be dealt with in certain settings. One may then wonder why not simply use some other inference method that is valid when the number of clusters is small and that does not rely on estimating β cluster-by-cluster. Perhaps the most popular approach in that category is the cluster wild bootstrap popularized by Cameron, Gelbach, and Miller (2008) and recently studied formally by Canay, Santos, and Shaikh (2021). While

not having to estimate β within each cluster represents an advantage over ARTs, this additional flexibility comes at a cost in terms of the degree of heterogeneity that the model can deal with. In particular, the results in Canay, Santos, and Shaikh (2021) show that the cluster wild bootstrap is expected to work well in settings with a small number of clusters *as long as* the clusters are "homogeneous", in a sense made precise in Canay, Santos, and Shaikh (2021). Intuitively, it is required that the variance covariance matrix $\hat{\Omega}_{n,j}$ defined in (20) is the same across clusters (up to scalar multiplication). Such stringent homogeneity condition is not required for ARTs to work well, as the method allows clusters to be arbitrarily heterogeneous as long as $\hat{\Omega}_{n,j}$ is invertible for $j \in J$.

Remark 4.5. For ease of exposition, we have written the requirement in (26) in terms of the differences $\hat{\beta}_{n,j} - \beta$, but it is possible to replace it with the differences $c'\hat{\beta}_{n,j} - c'\beta$ (or $R\hat{\beta}_{n,j} - R\beta$, depending on the null hypotheses of interest). In most cases, re-writing the condition in this way is not useful, but it is in cases where $c'\beta$ is identified within each cluster while β is not. For example, consider the model in (28) when the coefficient of interest is β_2 as opposed to β_1, i.e. $c'\beta = \beta_2$. In that case the entire term $\beta_0 + Z_j^{(1)}\beta_1$ may be absorbed into a cluster-specific intercept without affecting the identification and estimation of $c'\beta = \beta_2$ within each cluster. □

5 Empirical Applications

In this section we apply ARTs as described in Algorithm 2.1 and ART-based confidence intervals as described in Algorithm 2.2 in the context of two distinct empirical applications. The `R` and `Stata` packages and codes required to replicate the results in this section are available as part of the online supplemental material.

5.1 Meng, Qian, and Yared (2015)

Meng, Qian, and Yared (2015, MQY) argue that China's Great Famine, from 1959 to 1961, was the result of an inflexible food procurement policy by the central government. To make this point, they show that food production and mortality become positively correlated during the time of famine, when this coefficient is otherwise negative or not significantly different from 0 in normal times.

MQY consider the following regression,

$$Y_{j,t+1} = Z_{j,t}^{(1)}\beta_1 + Z_{j,t}^{(2)}\beta_2 + \delta_t + \epsilon_{j,t}$$

where j indexes provinces (ranging from 1 to 19) and t indexes years (ranging from 1953 to 1982). Here,

$$\begin{aligned} Y_{j,t+1} &= \log(\text{number of deaths in province } j \text{ during year } t+1) \\ Z_{j,t}^{(1)} &= \log(\text{predicted grain production in province } j \text{ during year } t) \\ &\quad \times I\{t \text{ is a famine year}\} \\ Z_{j,t}^{(2)} &= \log(\text{predicted grain production in province } j \text{ during year } t) \\ \delta_t &= \text{time fixed effects.} \end{aligned}$$

In this application the level of clustering is a province, and so in order to apply ARTs as described in Section 2.1, one needs to estimate $\beta = (\beta_1, \beta_2)'$ and δ_t province-by-province. This illustrates one of the situations where including time fixed effects province-by-province is infeasible for the implementation of ARTs, given that the only source of remaining variation within a province is indeed time. The second identification problem described in Section 4 then arises. As we discussed in that section, one way to deal with this issue

consists of replacing the time fixed effects with a cluster-specific time trend, i.e. in Step 1 of Algorithm 2.1 estimate

$$Y_{j,t+1} = Z_{j,t}^{(1)}\beta_1 + Z_{j,t}^{(2)}\beta_2 + \gamma_j t + \epsilon_{j,t}. \tag{30}$$

We will refer to this as Analysis #1. In addition, we also consider the following alternative specifications studied by MQY:

- Analysis #2: Repeating Analysis #1 using only data between 1953 and 1965.
- Analysis #3: Repeating Analysis #1 using four additional autonomous provinces.
- Analysis #4: Repeating Analysis #2 using four additional autonomous provinces.
- Analysis #5: Repeating Analysis #1 using actual rather than constructed grain production.
- Analysis #6: Repeating Analysis #2 using actual rather than constructed grain production.

As with Analysis #1, the above analyses differ from their MQY counterparts only in that a linear time trend $\gamma_j t$ replaces time fixed effects δ_t. Table 1 summarizes the number of clusters and the number of observations for each of these analyses. We caution, however, that in this application, in addition to the number of clusters being small, the number of observations within each cluster may also be small. See Remark 4.2 for further discussion in relation to Assumption 4.1(a).

Meng, Qian, and Yared (2015) consider the following two null hypotheses of interest,

$$H_0^{(1)}: \beta_1 = 0 \quad \text{and} \quad H_0^{(2)}: \beta_1 + \beta_2 = 0. \tag{31}$$

In Table 2 we replicate the main table in Meng, Qian, and Yared (2015) using cluster robust standard errors (CCE) and also include the results associated with ARTs for both $H_0^{(1)}$ and $H_0^{(2)}$ in (31). For $H_0^{(1)}$ we report p-values and 95% confidence intervals, while for $H_0^{(2)}$ we just report p-values following MQY. The authors note in footnote 33 that using the cluster wild bootstrap led to similar results as those presented in their main table so we do not include cluster wild bootstrap results here either.

We comment on the following main features of Table 2:

1. For the null hypothesis $H_0^{(1)}$ associated with the parameter β_1, the ART p-values are of comparable magnitude to traditional CCE p-values. Similarly, ART-based confidence intervals are of roughly the same length as those obtained based on CCE although the ART-based confidence intervals do not contain the LS estimates. This is because ART-based confidence intervals are centered around the mean of the province-by-province estimates, which may not necessarily be equal to the full sample LS estimate of β_1.
2. For the null hypothesis $H_0^{(2)}$ associated with the parameter $\beta_1 + \beta_2$, the ART p-value is sometimes higher and sometimes lower than the CCE p-value depending on the specification. Given the relatively small number of clusters in this application, the ART p-values are likely to be more reliable than those associated with CCE as CCE is known to perform poorly when the number of clusters is not sufficiently large.

5.2 Munyo and Rossi (2015)

Munyo and Rossi (2015) study criminal recidivism of former prisoners by looking at the relationship between the number of inmates released from incarceration on a given day and the number of offenses committed on

Table 1: Cluster information. 'Min. Size', 'Med. Size', 'Max. Size' denote the minimum, the median, and the maximum size of clusters.

Analysis	# of Clusters	Min. Size	Med. Size	Max. Size	Mean
#1, #5	19	29	30	30	29.95
#2, #6	19	12	13	13	12.95
#3	23	29	30	30	29.96
#4	23	12	13	13	12.96

Table 2: Results for Analyses #1–6, comparable to those in Table 2 of Meng, Qian, and Yared (2015). 'LS Estimate' denotes the full sample OLS estimate for β_1. CCE refers to cluster-robust standard errors. ART p-values are obtained using Algorithm 2.1. ART-based 95% confidence intervals are obtained using Algorithm 2.2.

	#1	#2	#3	#4	#5	#6
LS estimate: β_1	0.063	0.057	0.071	0.067	0.064	0.058
CCE: Province						
se	0.007	0.007	0.007	0.008	0.007	0.007
p-value	0.000	0.000	0.000	0.000	0.000	0.000
95% CI	[0.050, 0.077]	[0.043, 0.071]	[0.057, 0.086]	[0.051, 0.083]	[0.051, 0.078]	[0.044, 0.071]
ART						
p-value	0.000	0.002	0.000	0.000	0.000	0.000
95% CI	[0.032, 0.055]	[0.018, 0.047]	[0.038, 0.066]	[0.028, 0.067]	[0.032, 0.058]	[0.029, 0.050]
$\beta_1 + \beta_2 = 0$						
CCE p-value	0.050	0.009	0.059	0.005	0.266	0.363
ART p-value	0.098	0.571	0.096	0.487	0.080	0.001
Observations	569	246	689	298	569	246
Short sample	No	Yes	No	Yes	No	Yes
Auto. Region	No	No	Yes	Yes	No	No
Pred. Grain prod.	Yes	Yes	Yes	Yes	No	No

the same day. They claim that the liquidity constraints that inmates face on the day of release increase the likelihood of recidivism on the same day. Using data of 2631 days between January 1st 2004 and March 15 2011 collected from the criminal incidents reports in Montevideo in Uruguay, they estimate the following linear model by least squares

$$Y_t = Z_t'\beta + \epsilon_t$$

where t indexes days and

$$Y_t = \text{the total number of offenses on day } t$$

$$Z_t = \text{the total number of inmates released, temperature, rainfall, hours of sunshine on day } t, \text{ a dummy for holidays, a dummy for December 31st and a yearly trend.}$$

We refer to this as Analysis #1. Munyo and Rossi (2015) additionally consider the following four analyses:

- Analysis #2: Z_t includes a daily trend in place of a yearly trend.
- Analysis #3: Z_t includes a monthly trend in place of a yearly trend.
- Analysis #4: Z_t includes an intra-month daily trend, month- and year-level fixed effects and their interactions in place of a yearly trend.
- Analysis #5: Z_t includes month- and year-level fixed effects and their interactions in place of a yearly trend.

Analysis #5 is their preferred specification. Munyo and Rossi (2015) report the results of these analyses in Table 2 in their paper. They report least squares estimates of β with Newey-West heteroskedasticity-autocorrelation-consistent (HAC) standard errors. In addition, they report ART p-values as described in Algorithm 2.1 for the null hypothesis that $H_0 : c'\beta = 0$ as in (2), where c selects the coefficient on the total number of inmates released on day t.

In this application the level of clustering is not naturally determined by the data, but pseudo-clusters may be formed using blocks of consecutive observations under the assumption of weak temporal dependence. In order to apply ARTs as described in Algorithm 2.1 we then form q pseudo-clusters by dividing the data into

q consecutive blocks of size $b_n = \lfloor n/q \rfloor$ where $n = 2631$ is the number of total observations. More concretely, we define the jth pseudo-cluster as

$$X_j^{(n)} = \left\{ (Y_t, Z_t')' : t = (j-1)b_n + 1, \dots, jb_n \right\} \quad \text{where} \quad j = 1, \dots, q-1,$$

and let the last qth pseudo-cluster contain all the remaining $n - b_n(q-1)$ observations. Note that in this application the number of pseudo-clusters q is a tuning parameter that the analyst must specify. Munyo and Rossi (2015) set $q = 10$. We repeat their analyses with alternative values of q and investigate how sensitive the results are to this choice. The relevant cluster information is given in Table 3.

Table 4 shows LS estimates of β, p-values for the hypothesis in (2), and 95% confidence intervals for each analysis. Following Munyo and Rossi (2015), we report results based on HAC standard errors. The table also shows ART p-values as described in Algorithm 2.1 and ART-based 95% confidence intervals as described in Algorithm 2.2 for $q = 8$, $q = 10$, and $q = 16$.

We summarize the main findings of the results in Table 4 as follows:

1. The choice of q is important for the results of ARTs but currently there is no theory developed to choose this tuning parameter according to some data dependent criteria. The smaller q is, the more observations are available within each cluster. Having more observations per cluster is important for one of the requirements behind ARTs, mainly (26). A small value of q, however, tends to affect the power of ARTs despite not really affecting the control of the rejection probability under the null hypothesis. This feature

Table 3: Pseudo-cluster size for different values of q.

# of Clusters (q)	Cluster Size
8	328
10	263
16	164

Table 4: Results for Analyses #1–5, comparable to those in Table 2 of Munyo and Rossi (2015). 'LS Estimate' denotes the full sample LS estimate of β. HAC refers to the heteroskedasticity and autocorrelation consistent standard error. ART p-values are obtained using Algorithm 2.1. ART-based 95% confidence intervals are obtained using Algorithm 2.2.

Specification	#1	#2	#3	#4	#5
LS estimate	0.225	0.260	0.259	0.225	0.234
HAC					
se	0.124	0.123	0.123	0.096	0.096
p-value	0.068	0.034	0.034	0.019	0.015
95% CI	[−0.017, 0.468]	[0.02, 0.5]	[0.019, 0.5]	[0.038, 0.413]	[0.046, 0.421]
ART: $q = 8$					
p-value	0.008	0.023	0.023	0.102	0.102
95% CI	[0.124, 0.429]	[0.035, 0.391]	[0.035, 0.391]	[−0.07, 0.397]	[−0.067, 0.418]
ART: $q = 10$					
p-value	0.002	0.014	0.014	0.063	0.053
95% CI	[0.141, 0.603]	[0.068, 0.446]	[0.068, 0.458]	[−0.023, 0.431]	[−0.003, 0.452]
ART: $q = 16$					
p-value	0.002	0.006	0.006	0.027	0.010
95% CI	[0.131, 0.444]	[0.097, 0.369]	[0.087, 0.371]	[0.02, 0.324]	[0.056, 0.367]
Observations	2631	2631	2631	2631	2631
Time trend	Year	Day	Month	Intra-month day	None
Time fixed effect	No	No	No	Yes	Yes
Controls	No	No	No	No	No

can be seen in Table 4, where ARTs p-values are decreasing in q across different specifications. In this application, where there are still over a hundred observations when $q = 16$, a larger value of q like $q = 10$ or $q = 16$ may be preferable to smaller values, like $q = 8$, based on power considerations. Note, however, that except in Analyses #4–5, where the choice of q determines whether the null hypothesis is rejected at a given significance level, the results for Analyses #1–3 are in all agreement at a 5% level.

2. Overall, the test results based on standard t-test with HAC standard errors are consistent to those of ARTs when $q = 16$. Both methods reject the null hypothesis $H_0 : c'\beta = 0$ at a 10% nominal level across different specifications. The results support the authors' argument that the release of inmates from incarceration increase the chance of re-offenses on the day of release.

5.3 Computational Gains of the New Algorithm

Tables 5 and 6 report four alternative ways to compute ART-based confidence intervals in the two empirical applications we consider in this paper; Meng, Qian, and Yared (2015) and Munyo and Rossi (2015). The first alternative is to compute the confidence intervals by a simple grid search algorithm. The second alternative involves a bi-section algorithm. We implement both of these methods using a studentized and an unstudentized test statistic to illustrate the result in Section 3.2. The last alternative is to simply use Algorithm 2.2, as reported in Sections 5.1 and 5.2. In each case, we also report computational times to illustrate the computational advantages of the algorithm we propose in this paper. The `R` and `Stata` codes required to replicate the results in this section are available as part of the online supplemental material.

Starting from Table 5, we see that grid search take a significant amount of time to compute. Our convexity result (Lemma 3.3) facilitates the use of the bisection method, cutting implementation time by a factor of over 50. Moving from the bisection method to Algorithm 2.2 further leads to a speed up of at least 2 times. A similar pattern emerges in Table 6. Furthermore, comparing specification with $q = 8$ that with $q = 16$, the speed advantage of our method becomes far starker. For $q = 16$, grid search takes almost 100 times as long as the bisection method. The bisection method, meanwhile, takes close to 10 times as long as Algorithm 2.2.

Table 5: Computational gains of Algorithm 2.2 relative to grid search and bisection algorithms in the applications of Section 5.1. The top row for each specification is the confidence interval. The bottom row is time in seconds. For the bisection search, our tolerance is set to the absolute value of the LS estimate, divided by 1000. For comparability, we set the step-size of the grid search to the same value.

	Grid Search		Bisection		ART
	Stud.	Unstud.	Stud.	Unstud.	
#1	[0.032, 0.055]	[0.032, 0.055]	[0.032, 0.055]	[0.032, 0.055]	[0.032, 0.055]
	19.65	6.66	0.31	0.11	0.06
#2	[0.018, 0.047]	[0.018, 0.047]	[0.018, 0.047]	[0.018, 0.047]	[0.018, 0.047]
	46.63	16.43	0.35	0.12	0.02
#3	[0.038, 0.066]	[0.038, 0.066]	[0.038, 0.066]	[0.038, 0.066]	[0.038, 0.066]
	24.50	8.67	0.30	0.09	0.03
#4	[0.028, 0.067]	[0.028, 0.067]	[0.028, 0.067]	[0.028, 0.067]	[0.028, 0.067]
	62.52	21.56	0.34	0.11	0.02
#5	[0.032, 0.058]	[0.032, 0.058]	[0.032, 0.058]	[0.032, 0.058]	[0.032, 0.058]
	19.47	6.76	0.27	0.11	0.01
#6	[0.029, 0.050]	[0.029, 0.050]	[0.029, 0.050]	[0.029, 0.050]	[0.029, 0.050]
	23.32	8.26	0.30	0.11	0.01

Table 6: Computational gains of Algorithm 2.2 relative to grid search and bisection algorithms in the applications of Section 5.2. The top row for each specification is the confidence interval. The bottom row is time in seconds. For the bisection search, our tolerance is set to the absolute value of the LS estimate, divided by 1000. For comparability, we set the step-size of the grid search to the same value.

		Grid Search		Bisection		ART
		Stud.	Unstud.	Stud.	Unstud.	
$q = 8$	#1	[0.124, 0.429]	[0.124, 0.429]	[0.124, 0.429]	[0.124, 0.429]	[0.124, 0.429]
		2.62	0.92	0.11	0.02	0.06
	#2	[0.035, 0.391]	[0.035, 0.391]	[0.036, 0.391]	[0.036, 0.391]	[0.035, 0.391]
		3.85	1.40	0.08	0.03	0.00
	#3	[0.035, 0.391]	[0.035, 0.391]	[0.035, 0.390]	[0.035, 0.390]	[0.035, 0.390]
		4.09	1.50	0.08	0.03	0.00
	#4	[−0.070, 0.397]	[−0.070, 0.397]	[−0.070, 0.397]	[−0.070, 0.397]	[−0.070, 0.397]
		9.29	3.37	0.11	0.03	0.00
	#5	[−0.067, 0.418]	[−0.067, 0.418]	[−0.067, 0.418]	[−0.067, 0.418]	[−0.067, 0.418]
		9.20	3.18	0.07	0.04	0.01
$q = 10$	#1	[0.141, 0.603]	[0.141, 0.603]	[0.141, 0.603]	[0.141, 0.603]	[0.141, 0.603]
		30.19	10.41	0.33	0.11	0.01
	#2	[0.068, 0.446]	[0.068, 0.446]	[0.068, 0.446]	[0.068, 0.446]	[0.069, 0.445]
		26.61	9.34	0.33	0.13	0.00
	#3	[0.067, 0.458]	[0.067, 0.458]	[0.068, 0.458]	[0.068, 0.458]	[0.068, 0.458]
		28.37	9.78	0.33	0.11	0.02
	#4	[−0.024, 0.431]	[−0.024, 0.431]	[−0.024, 0.430]	[−0.024, 0.430]	[−0.023, 0.430]
		32.75	11.47	0.32	0.11	0.02
	#5	[−0.003, 0.452]	[−0.003, 0.452]	[−0.003, 0.452]	[−0.003, 0.452]	[−0.003, 0.451]
		31.67	11.02	0.34	0.11	0.02
$q = 16$	#1	[0.124, 0.447]	[0.124, 0.447]	[0.124, 0.447]	[0.124, 0.447]	[0.124, 0.447]
		373.86	127.07	3.19	1.11	0.13
	#2	[0.098, 0.364]	[0.098, 0.364]	[0.098, 0.364]	[0.098, 0.364]	[0.097, 0.364]
		451.67	153.20	3.16	1.11	0.14
	#3	[0.088, 0.368]	[0.088, 0.368]	[0.088, 0.368]	[0.088, 0.368]	[0.087, 0.368]
		415.67	142.55	3.25	1.10	0.16
	#4	[0.014, 0.325]	[0.014, 0.325]	[0.015, 0.325]	[0.015, 0.325]	[0.014, 0.325]
		703.13	248.68	3.43	1.14	0.11
	#5	[0.048, 0.365]	[0.048, 0.365]	[0.048, 0.365]	[0.048, 0.365]	[0.047, 0.365]
		572.54	193.69	3.09	1.10	0.14

6 Concluding Remarks

The goal of this paper is to make the general theory developed in Canay, Romano, and Shaikh (2017a) more accessible by providing a step-by-step algorithmic description of how to implement the test and construct confidence intervals in linear regression models with clustered data, as well as clarifying the main requirements and limitations of the approach. The main two takeaways are the following. First, ARTs-based confidence intervals for scalar parameters in linear regression models can be characterized in closed form and thus are straightforward to implement in practice. Algorithms 2.1 and 2.2 provide a clear explanation of how to apply ARTs in linear models, and the companion `Stata` and `R` packages available as part of the supplemental material are intended to facilitate doing so. Second, our discussion on the main requirements behind ARTs hopefully show that understanding the trade-offs between ARTs and other popular alternatives for inference with a small number of clusters, like the cluster wild bootstrap, is fundamental for practitioners to choose a method that aligns well with the features of their application. In particular, while ARTs essentially demand that the parameter of interest is suitably estimable cluster-by-cluster without imposing restrictions on the

degree of heterogeneity across clusters, the cluster wild bootstrap requires the clusters to be sufficiently homogeneous (see Canay, Santos, and Shaikh (2021)) without demanding identification of the parameter of interest cluster-by-cluster.

Acknowledgment: We would like to thank Matthew Thomas for excellent research assistance developing the `R` and `Stata` packages for this paper. The research of the fourth author is supported by NSF Grant SES-1530661.

References

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249–75.

Cameron, A. C., J. B. Gelbach, and D. L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90 (3): 414–27.

Canay, I. A., and V. Kamat. 2018. "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design." *The Review of Economic Studies* 85 (3): 1577–608.

Canay, I. A., J. P. Romano, and A. M. Shaikh. 2017a. "Randomization Tests under an Approximate Symmetry Assumption." *Econometrica* 85 (3): 1013–30.

Canay, I. A., J. P. Romano, and A. M. Shaikh. 2017b. "Supplement to 'Randomization Tests under an Approximate Symmetry Assumption'." *Econometrica Supplemental Material* 85: 1–14.

Canay, I. A., A. Santos, and A. M. Shaikh. 2021. "The Wild Bootstrap with a 'Small' Number of 'Large' Clusters." *The Review of Economics and Statistics* 103: 346–63.

Conley, T., S. Gonçalves, and C. Hansen. 2018. "Inference with Dependent Data in Accounting and Finance Applications." *Journal of Accounting Research* 56 (4): 1139–203.

DiCiccio, C. J., T. J. DiCiccio, and J. P. Romano. 2020. "Exact Tests via Multiple Data Splitting." *Statistics & Probability Letters* 166: 108865.

Ibragimov, R., and U. K. Müller. 2010. "*t*-Statistic Based Correlation and Heterogeneity Robust Inference." *Journal of Business & Economic Statistics* 28 (4): 453–68.

Ibragimov, R., and U. K. Müller. 2016. "Inference with Few Heterogeneous Clusters." *The Review of Economics and Statistics* 98 (1): 83–96.

Liang, K.-Y., and S. L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

Meng, X., N. Qian, and P. Yared. 2015. "The Institutional Causes of China's Great Famine, 1959–1961." *The Review of Economic Studies* 82 (4): 1568–611.

Munyo, I., and M. A. Rossi. 2015. "First-Day Criminal Recidivism." *Journal of Public Economics* 124: 81–90.