Cluster-Robust Inference when Treatment Effects Vary Across Clusters: A Design-Based Approach

Robin Danko

Department of Economics, Michigan State University

September 2025

Abstract

Clustered standard errors are common tools used by empirical researchers in the social sciences to obtain valid inference. Clusters can be defined by geography or by social strata such as gender. Typically, cluster-robust errors adjust for correlations induced by sampling the outcome from a data-generating process with correlated cluster-level components. The data-generating process is typically assumed to either contain a small, finite, and fixed number of clusters or an infinite number of clusters. In this paper, I show that modeling clusters as finite features of a population the number of conventional cluster-robust standard errors can be severely inflated when the number of clusters is large enough for valid asymptotic approximation. I propose new standard errors that correct this bias.

KEYWORDS: Finite population, causal estimands, clustering

Robin Danko: rdanko@msu.edu

I am grateful for the support from my committee Antonio Galvao, Timothy Vogelsang, and Jeffrey Wooldridge. I am also grateful for comments from Hugo Freeman and Kyoo il Kim.

1 Introduction

In the traditional approach to inference in the social sciences, observations are modeled as sampled from an infinitely large population. Moreover, when considering clustering, it is often supposed that as the sample size grows, more and more clusters will be sampled as well. However, it is not clear where those new clusters come from. It seems as though they are features of a population that contains an unlimited number of clusters; as one draws more samples from that population, one draws more clusters in expectation. This is most similar to sampling with replacement: one can flip a coin over and over to sample new, independent realizations. This is a reasonable approximation in some cases, such as in panel data settings where individuals are sampled repeatedly over time, each individual constitutes their own cluster, and only a small fraction of individuals are sampled. However, in cases where clustering occurs at the level of U.S. states, it seems less compelling. In this case, there are only fifty clusters and no matter how many observations are drawn from this population a researcher will only observe fifty clusters. Sampling with replacement will not draw more and more clusters.

Moreover, the dependence between observations belonging to the same cluster is traditionally modeled as a dependence in the random errors or "shocks" present in the outcome model. This is the "modeling-based" approach to clustering. In this paper, I consider other features of study design that takes place at the level of clustering, such as sampling and treatment.

The setting presented in this paper differs, however, from "fixed-G" treatments of clustering by presenting asymptotic results that might be well-approximated by populations with a finite number of clusters. This case is important because in some cases the number of clusters in a population is very large but a significant fraction of these clusters are nevertheless sampled. Therefore, population objects might be well-approximated by asymptotic results that take into consideration the proportion of sampled clusters.

When cluster numbers are treated as fixed and finite features of the population in this way, cases in which a significant fraction of population's clusters are sampled can be analyzed. This paper analyzes two cases: one in which observations are drawn from a superpopulation that contains a finite number of clusters, and one in which observations are drawn from a finite population which by necessity can only contain finite clusters. In this paper I show that in these cases the usual cluster-robust standard errors for the OLS estimator are conservative and wrong. I characterize the population variance as a function of the observed proportion of clusters. I propose an alternative estimator that robustly estimates the population variance.

This paper builds upon the growing literature of finite-population adjustments for OLS standard errors. Most directly, Abadie et al. (2023) show results analogous to those presented in this paper in a

finite-population, finite-cluster setting with a binary treatment without controls. I expand upon these results by extending the class of treatments to treatments that are continuous, binary, or mixed, as well as including controls. Moreover, I show that finite-population assumptions are not required to obtain many of the results: I present results for the setting in which the population is infinite and the number of clusters is finite.

Other authors such as Donald and Lang (2007) have considered the question of finite clusters when the number of groups is small. The examples they provide are two period/two group studies. The results in this paper apply to settings in which the number of groups in both the population and the sample are large. For instance, if clustering is performed at the level of U.S. county tracts, there are just over 84,000 clusters and so asymptotic approximations based on a population containing an infinite population might be credible. However, suppose all 84,000 clusters are sampled as well. The traditional approach considers a sequence of samples that each contain a negligible fraction of a population's clusters. I suppose both a sequence of populations and a sequence of samples, then consider the asymptotic behavior of the estimator computed from the sequence of samples. This preserves the ratio of a sample's clusters over the corresponding population's clusters, which is critical component of the asymptotic variance.

Xu and Yap (Working Paper, 2024) provide alternative standard errors for M-estimators under finite-population assumptions, noticing that the usual cluster-robust standard errors are too conservative. The standard errors proposed in that paper are still conservative. This paper demonstrates that if more structure is placed on the treatment, then robust standard errors can be obtained. In that paper, treatment is allowed to be independently but not identically (i.n.i.d.) distributed across all observations. The results of this paper require that treatment be i.i.d. within each clusters, while still allowing for treatment distribution to vary between clusters.

Much of the asymptotic theory in this paper is built upon the work in Abadie et al. (2020) which noticed that OLS standard errors are conservative in the case of heteroskedasticity under finite-population assumptions, and Hansen and Lee (2019) which establishes formal asymptotic theory for clustered samples.

2 Model

Consider a sequence of populations indexed by finite integer k. This sequence of populations and their indices k are theoretical constructions that allow me to characterize the asymptotic behavior of estimators of interest when sample sizes and cluster sizes are large enough. The researcher is only concerned with one population and only observes one sample. Each population is partitioned into G_k clusters and

contains n_k units. The number of units in cluster g is denoted $n_{k,g}$. The researcher observes a sample N_k from each population. For each sampled unit, indexed i, the researcher observes its cluster $g_{k,i}$, a $L \times 1$ set of covariates $X_{k,i}$, an outcome $Y_{k,i}$, and a treatment variable $K \times 1$ $W_{k,i}$. All of these variables can be continuous, discrete, or mixed.

Suppose the researcher is interested in the parameters of the following linear model:

$$Y_{k,i} = W'_{k,i}\theta_{k,i} + X'_{k,i}\gamma_k + \nu_{k,i}$$

where $\theta_{k,i}$ and $\gamma_{k,i}$ are $K \times 1$ and $L \times 1$ vectors of parameters, respectively, and $\nu_{k,i}$ is a random shocks. Notably, the linear treatment effects $\theta_{k,i}$ are allowed to vary across individuals.

The model proposed in this setting departs from the potential outcome framework discussed in the previous work on finite-population analysis. In both this setting and in previous settings, treatment effects are allowed to be fully heterogeneous across individuals. In previous work, however, the controls (or attributes) $X_{k,i}$ and the errors $\nu_{k,i}$ are modeled as nonrandom. The reason for the departure from this approach in this paper is due to the role of covariates: if the controls are nonrandom, how can they influence the treatment?

Abadie et al. (2020) model the effect of covariates on treatment by allowing the random treatments to be a function of the nonrandom covariates. Since each individual treatment can vary with each individual's attributes, the treatments are assumed to be distributed i.n.i.d.; in order for the results of this paper to hold (and the results of Abadie et al. (2023) who do not consider covariates beyond fixed effects), treatment must be distributed i.i.d., at least within each cluster.

Random errors $\nu_{k,i}$ are not strictly necessary for the theoretical results. Indeed, the potential outcome model proposed in Abadie et al. (2023) does not include random errors. However, the inclusion of random errors lends more plausibility to the linear model. Suppose the true model of $Y_{k,i}$ has the following form:

$$Y_{k,i} = W'_{k,i}\theta_{k,i} + f_{k,i}(X_{k,i}).$$

If the controls $X_{k,i}$ are nonrandom, then this model can be properly specified as the nonstochastic function of w:

$$Y^*_{k,i}(w) = w\theta_{k,i} + \xi_{k,i}$$

where $\xi_{k,i} = f_{k,i}(X_{k,i})$ and is nonrandom. This is the approach taken by Abadie et al. (2020) . However, if

$X_{k,i}$ is random,

$$Y_{k,i} = W_{k,i}^{'}\theta_{k,i} + X_{k,i}\gamma_k + \nu_{k,i}$$

where $\nu_{k,i} = f_{k,i}(X_{k,i}) - X_{k,i}\gamma_k$. Then, $\nu_{k,i}$ is random unless $f_{k,i}(X_{k,i}) - X_{k,i}\gamma_k$ is a constant for all X. Therefore, modeling the errors as nonrandom requires correctly modeling the relationship between the outcome and the controls; this paper only requires that after a linear function of the controls is partialed out from the treatment, $W_{k,i}\nu_{k,i}$ satisfies a suitable orthogonality condition.

Let $n_{k,g}$ denote the number of sampled observations belonging to each cluster. Define $X_{k,g}$, $Y_{k,g}$, and $W_{k,g}$ to be the matrices containing the controls, outcomes, and treatments, respectively, for all units in cluster g.

2.1 Sampling

Sampling follows a two-step process identical to that suggested by Abadie et al. (2023) .

First, clusters are sampled with probability $q_k \in (0, 1]$. Then, observations are sampled from the remaining clusters with unit sampling probability $p_k \in (0, 1]$. Notice that both of these intervals include 1. If $q_k = 1$, every cluster is sampled. If $p_k = 1$, every observation from every sampled cluster is sampled. If $q_k = p_k = 1$, then every observation in every cluster is sampled.

The researcher observes $Q_{k,g}$, the vector indicating sampling status for each cluster and $R_{k,i}$, the vector indicating sampling status for each unit in the population, including cluster sampling status.

This is most similar to variable probability sampling as discussed in Quesenberry and Jewell (1986) with each cluster forming a stratum. However, in typical variable probability sampling, the selection for each point in selected point stratum (cluster) g is determined individually. Here, $Q_{k,g}$ is a random variable that determines selection for every point in a stratum (cluster) g: the entire set of observations for that cluster is selected or discarded with probability q_k. Then, the sampling for each point within each sampled cluster is determined individually with probability p_k following a process identical to variable probability sampling with selection probability constant across all clusters.

Assumption 2.1. i) There exists sequences of sampling probabilities p_k and q_k s.t.

$$\Pr(R_k = r) = p_k^{\sum_{i=1}^{n_k} r_i}(1 - p_k)^{n_k - \sum_{i=1}^{n_k} r_i}$$
$$\Pr(Q_k = s) = q_k^{\sum_{i=1}^{G_k} s_i}(1 - q_k)^{G_k - \sum_{i=1}^{G_k} s_i}$$

for all vectors r of length n_k, where $r_i \in \{0,1\}$ and for all vectors s of length G_k, where $s_i \in \{0,1\}$. ii) The sequences of sampling rates p_k and q_k satisfy $n_k p_k \to \infty$ and $G_k q_k \to \infty$ and $\{p_k, q_k\} \to \{p, q\}$. p, q are contained in the interval $[0,1]$.

Assumption 2.2.ii guarantees that the expected sample size grows as k increases, as does the expected number of sampled clusters. This situates this paper in the "large-G" asymptotic case.

In the usual approach to inference, it is assumed that a negligible fraction of clusters are sampled from a population that has an infinite number of clusters; this special case is addressed in this paper's asymptotic results when $q \to 0$.

N_k, then, denotes the number of units observed after sampling and $N_{k,g}$ denote the number of sampled observations within each cluster. Both of these terms are random because they incorporate the sampling scheme.

Assumption 2.2.

1. $\liminf_{k\to\infty} p_k \min_g n_{k,g} > 0$
2. $\limsup_{k\to\infty} \frac{\max_g n_{k,g}}{\min_g n_{k,g}} < \infty$

Part 1 of Assumption 2.2 guarantees that the average number of units sampled in any sampled cluster never goes to 0. Part 2 restricts degree to which sampled cluster sizes can be unbalanced.

2.2 Clustering

There are three forms of uncertainty that I allow to occur at the cluster level, distinguished in Wooldridge (2023) as *sample-based*, *design-based*, and *model-based*. Sampling-based uncertainty is discussed in the previous section through the cluster-sampling mechanism.

Typically, clustering is modeled as dependence in the linear model random shocks $\nu_{k,g}$: that is, $\text{Cov}(\nu_{k,i,g}, \nu_{k,j,g}) \neq 0$ but $\text{Cov}(\nu_{k,i,g}, \nu_{k,j,h}) = 0$ when $g \neq h$. Errors within clusters are allowed to be correlated but errors across clusters are not. This is what is defined as model-based uncertainty. This paper discusses both this case and also cases in which clustering enters through the treatment and controls, or design-based uncertainty.

Assumption 2.3 (Assignment Mechanism)**.** Clustered treatment in the continuous case follows a two-step process:

1. Suppose there exists a random variable A_k with mean μ_k and variance $\sigma_k^2 < \infty$. Independently sample A_k for each group g to obtain the mean level of treatment $A_{k,g}$.

2. The treatment distribution W is common across all groups and $E(W|A_{k,g}) = A_{k,g}$. For a cluster g, treatment is drawn i.i.d. from the conditional distribution $W|A_{k,g}$. That is, the mean treatment level for each cluster is determined by the corresponding draw $A_{k,g}$.

The clustering mechanism proposed for binary treatment by Abadie et al. (2023) draws the mean treatment probability by cluster from some distribution A bound uniformly between 0 and 1. This is a special case of this framework because $E(W|A_{k,g}) = P(W = 1|A_{k,g})$ when W is binary.

Assumption 2.4.

1. The treatments $W_{k,1}, .., W_{k,n_k}$ are independent of the cluster sampling scheme $Q_{k,1}, ..., Q_{k,G_k}$ and the unit sampling scheme $R_{k,i}, ..., R_{k,n_k}$.

2. The controls $X_{k,i,g}$ are i.i.d. conditional on cluster assignment.

The treatment assignment scheme proposed in this paper is very flexible. For instance, the mechanism is agnostic to how A_k determines the mean level of the treatment. A_k is allowed to change features of the distribution other than the mean. For instance, if $W \sim \Gamma(\alpha, \beta)$, A_k might be a joint distribution of $\frac{\alpha}{\beta}$. A_k could determine α only, β only, or α and β, so long as β is bounded uniformly from 0.

Alternatively, clustering can be modeled as dependence in the error terms within clusters. I propose an alternative assumption to assumption 2.3 below.

Assumption 2.5 (Error clustering)**.**

1. $\text{Cov}(\nu_{k,i,g}, \nu_{k,j,h}) = 0$ when $g \neq h$. Covariance in error terms within clusters are unrestricted.

2. Treatment status in cluster g for each observation i is drawn i.i.d. from a common distribution $W_{k,g}$. The treatment distribution $W_{k,g}$ are allowed to vary across g.

3. Treatment status is independent across clusters.

The mechanism through which within-cluster correlation enters the data is important to consider. As Wooldridge (2023) points out, the test proposed by MacKinnon et. al (2023) for which level to cluster at performs well when model-based uncertainty is the only component, but leads to overly conservative standard errors when uncertainty is design-based.

The multivocal nature of clustering also complicates the decision to cluster. In this section, I suppose that expected assignment levels, clustered sampling, and correlated errors occur at the level of a common group structure G_k. However, this need not be the case. For instance, treatment design could be at the level of county tract, while model error correlation could occur at the state level. Intuitively, clustering at

the coarser level should result in standard errors that estimate a population variance no larger than the true population variance, but even this needs to be shown and improvements upon this may be possible.

2.3 Setup and Notation

Assumption 2.6. $E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} X_{k,i}X_{k,i}^{'}\right]$ is full rank for all k large enough.

For ease of use, this paper works with the following transformation of the causal variable $W_{k,i}$, removing the correlation with the attributes:

$$U_{k,i} = W_{k,i} - \Lambda_k X_{n,i}$$

$$\Lambda_k = \left(E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} W_{k,i}X_{k,i}^{'}\right]\right)\left(E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} X_{k,i}X_{k,i}^{'}\right]\right)^{-1}.$$

While Λ_k is a function of the population moments and cannot be observed directly, it can be consistently estimated by

$$\hat{\Lambda}_k = \left(\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}W_{k,i}X_{k,i}^{'}\right)\left(\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}X_{k,i}X_{k,i}^{'}\right)^{-1}.$$

The next assumption is a regularity condition that bounds the moments of central variables.

Assumption 2.7. There exists some $\delta > 0$ s.t. the sequences

$$\frac{1}{n_k}\sum_{n_k=1}^{n_k} E\left[|Y_{k,i}|^{4+\delta}\right],\quad \frac{1}{n_k}\sum_{n_k=1}^{n_k} E\left[|U_{k,i}|^{4+\delta}\right],\quad \frac{1}{n_k}\sum_{n_k=1}^{n_k} E\left[|X_{k,i}|^{4+\delta}\right],\quad \frac{1}{n_k}\sum_{n_k=1}^{n_k} E\left[|\nu_{k,i}|^{4+\delta}\right]$$

are uniformly bounded.

Define

$$P_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i}\end{pmatrix}\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i}\end{pmatrix}^{'}, \qquad \Omega_k = \frac{1}{n_k}\sum_{i=1}^{n_k} E\left[\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i}\end{pmatrix}\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i}\end{pmatrix}^{'}\right].$$

$\Omega_k = E(P_k)$. Define $\chi_{k,i,g}$ to be the indicator function for each unit marking membership of unit i in cluster g for population k, defined as $\chi_{k,i,g} = \mathbf{1}\{g_{k,i} = g\}$. Let N_k be the number of sampled units or

$N_k = \sum_{m=1}^{G_k} Q_{k,m} \sum_{i=1}^{n_k} \chi_{g_{k,i}=m} R_{k,i}$. Then, define sampling analogues of P_k and Ω_k:

$$\tilde{P}_k = \frac{1}{N_k} \sum_{i=1}^{n_k} R_{k,i} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix}', \qquad \tilde{\Omega}_k = \frac{1}{N_k} \sum_{i=1}^{n_k} R_{k,i} E \left[\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix}' \right].$$

Throughout the rest of the paper, I will refer to partitions of these matrices using superscripts. For example,

$$P_k = \begin{pmatrix} P_k^{YY} & P_k^{YU} & P_k^{YX} & P_k^{Y\nu} \\ P_k^{UY} & P_k^{UU} & P_k^{UX} & P_k^{U\nu} \\ P_k^{XY} & P_k^{XU} & P_k^{XX} & P_k^{Y\nu} \\ P_k^{\nu Y} & P_k^{\nu U} & P_k^{\nu X} & P_k^{\nu\nu} \end{pmatrix}.$$

Lemma 2.1. Under assumptions 4.1,2.3, 2.4, and 2.7, or assumptions 4.1 2.5 and 2.7 $\tilde{P}_k \xrightarrow{p} \Omega_k$, $\tilde{\Omega}_k \xrightarrow{p} \Omega_k$, and $\tilde{P}_k \xrightarrow{p} P_k$

The next assumption imposes the condition that the expected value of the population second moments converge deterministically to some limit.

Assumption 2.8. $\Omega_k \to \Omega$, which has full rank.

2.4 Identification

The treament effects for each individual cannot be recovered in this framework. The average treatment effect within each cluster defined as $\theta_{k,g} = \frac{1}{n_{k,g}} \sum_{i=1}^{n_k} \chi_{k,i,g} \theta_{k,i}$. Each cluster treatment effect $\theta_{k,g}$ is identified in this framework and can be estimated consistently if the number of observations within each cluster grow. However, estimation and inference upon these parameters is challenging. Even if a sufficient number of observations are present to obtain a reliable estimate of $\theta_{k,g}$, inference is impossible due to the unrestricted covariance structure between observations within each cluster; much stronger assumptions about the covariance structure must be made in order to perform robust inference. Therefore, it is necessary to consider an estimand that combines information from all observed clusters. In this section I propose an identified estimand that is easily interpreted under an orthogonality condition I provide. This estimand is "causal" in the sense defined by Abadie et al. 2020 because it is a function of the population moments, not the observed realizations of the variables.

It is defined under the conditions above with probability approaching 1:

$$\begin{pmatrix} \theta_k^{\text{causal}} \\ \gamma_k^{\text{causal}} \end{pmatrix} = \begin{pmatrix} \Omega_k^{UU} & \Omega_k^{UX} \\ \Omega_k^{XU} & \Omega_k^{XX} \end{pmatrix}^{-1} \begin{pmatrix} \Omega_k^{UY} \\ \Omega_k^{XY} \end{pmatrix}.$$

This estimand satisfies the orthogonality condition defined in terms of the residuals with respect to θ_k^{causal}: $\varepsilon_k^{\text{causal}} = Y_{k,i} - U_{k,i}'\theta_k^{\text{causal}} - X_{k,i}'\gamma_k^{\text{causal}}$,

$$\frac{1}{n_k}\sum_{i=1}^{n_k} E\left[\begin{pmatrix} U_{k,i} \\ X_{k,i} \end{pmatrix} \varepsilon_{k,i}^{\text{causal}}\right] = 0.$$

The residuals $\epsilon_{k,i}^{\text{causal}}$ are not approximations of the model shocks $\nu_{k,i}$. To compare the two, observe

$$\nu_{k,i} = Y_{k,i} - W_{k,i}'\theta_{k,i} - X_{k,g}'\gamma_k.$$

That is, $\nu_{k,i}$ is a residual calculated using the unit-specific treatment effects. $\varepsilon_k^{\text{causal}}$ is the residual term in the misspecified regression modeling treatment effects as constant for each unit. I provide an orthogonality condition to interpret this estimand below.

Assumption 2.9.

$$E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} \chi_{k,i,g} W_{k,i}\nu_{k,i}\right] = 0 \ \forall g \quad \text{and}$$
$$E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} \chi_{k,i,g} X_{k,i}\nu_{k,i}\right] = 0 \ \forall g$$

This assumption is not strictly necessary for interpreting θ_k^{causal}; all that is required is that $E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right] = 0$, which the above assumption implies. However, to consistently estimate the population variance discussed later in the paper, the stronger orthogonality condition on each cluster is required.

Theorem 2.1. Under assumptions 4.1, 2.1, 2.3, 2.4, 2.7, 2.8, and 2.9

$$\theta_k^{\text{causal}} = \frac{1}{n_k}\sum_{i=1}^{n_k} \frac{1}{n_k}\theta_{k,i}$$

If assumption 2.3 is replaced by assumption 2.5

$$\theta_k^{\text{causal}} = E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} U_{k,i}U_{k,i}^{'}\right]^{-1}\sum_{m=1}^{G_k}\frac{n_{k,m}}{n_k}E[U_{k,m}U_{k,m}^{'}]\theta_{k,m}.$$

That is, θ_k^{causal} is a weighted average of all of the cluster treatment effects, with the weights determined by the variation in treatment distribution across clusters and the number of samples belong to each cluster. If $W_{k,i}$ is i.i.d., because $X_{k,i}$ is i.i.d., the above can be written as:

$$\theta_k^{\text{causal}} = \sum_{m=1}^{G_k}\frac{n_{k,m}}{n_k}\theta_{k,m}$$

as well.

The moment representation of θ_k^{causal} naturally suggests the sample analog OLS estimator, which will be the estimator of interest discussed in the rest of the paper defined as

$$\begin{pmatrix}\hat{\theta}_k\\ \hat{\gamma}_k\end{pmatrix} = \begin{pmatrix}\tilde{W}_k^{UU} & \tilde{W}_k^{UX}\\ \tilde{W}_k^{XU} & \tilde{W}_k^{XX}\end{pmatrix}^{-1}\begin{pmatrix}\tilde{W}_k^{UY}\\ \tilde{W}_k^{XY}\end{pmatrix}.$$

2.5 Asymptotic Properties of the OLS estimator

In this section I analyze the asymptotic properties of the OLS estimator as an estimator of θ_k^{causal}, with special attention to the asymptotic variance.

The key insight is noticing that when treatment effects vary across units $E[U_{k,i}\varepsilon_{k,i}^{\text{causal}}] = E[U_{k,i}U_{k,i}^{'}](\theta_{k,i} - \theta_k^{\text{causal}})$ which is not generally 0. However, the average across all units is 0 as discussed in the previous section.

Let $V_{k,i} = R_{k,i}U_{k,i}\varepsilon_{k,i}^{\text{causal}}$ and $V_{k,g} = \sum_{i=1}^{n_k}\chi_{k,i,g}V_{k,i}$. Then the "sandwich" term in the variance of the OLS estimator is $\frac{1}{n_k}\sum_{m=1}^{G_k}Q_{k,m}V_{k,m}^{'}$. Since these terms are independent across clusters,

$$\begin{aligned}
\text{Var}\left(\frac{1}{\sqrt{n_k}}\sum_{m=1}^{G_k} Q_{k,m}V_{k,m}^{'}\right) &= \frac{1}{n_k}\sum_{m=1}^{G_k} Var\left(Q_{k,m}V_{k,m}\right) \\
&= \frac{1}{n_k}\sum_{m=1}^{G_k} E\left[Q_{k,m}^2 V_{k,m}V_{k,m}^{'}\right] - E\left[Q_{k,m}V_{k,m}\right]E\left[Q_{k,m}V_{k,m}\right]^{'} \\
&= \frac{1}{n_k}\sum_{m=1}^{G_k} q_k E\left[V_{k,m}V_{k,m}^{'}\right] - q_k^2 E\left[V_{k,m}\right]E\left[V_{k,m}\right]^{'} \\
&= q_k * \left(\frac{1}{n_k}\sum_{m=1}^{G_k} q_k \text{Var}(V_{k,m}) + (1-q_k)E\left[V_{k,m}V_{k,m}^{'}\right]\right).
\end{aligned}$$

$\text{Var}(V_{k,m}) \leq E\left[V_{k,m}V_{k,m}^{'}\right]$ in the matrix sense and generally this inequality is strict. Since the usual cluster-robust standard errors only estimates the second-moment component of the variance, this explains why these errors are conservative in this setting. The discussion above is ad hoc; the intuition is revealing but must be formalized. Define the following limits of the population variances and covariances:

$$\begin{aligned}
&\lim_{k\longrightarrow\infty}\frac{1}{n_k}\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}\text{Var}(U_{k,i}\varepsilon_{k,i}^{\text{causal}}) + (1-p_k)\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}E(U_{k,i}^2\varepsilon_{k,i}^2)\right. \\
&\left.+2p_k\sum_{i=1}^{n_k}\sum_{j\neq i}\chi_{g_{k,i}=m}\chi_{g_{k,j}=m}\text{Cov}(U_{k,i}\varepsilon_{k,i}^{\text{causal}}, U_{k,j}\varepsilon_{k,j}^{\text{causal}})\right) = \lim_{k\longrightarrow\infty}\Delta_k^{\text{CCV}} \rightarrow \Delta^{\text{CCV}}
\end{aligned}$$

and

$$\lim_{k\longrightarrow\infty}\frac{1}{n_k}\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}E(U_{k,i}(\varepsilon_{k,i}^{\text{causal}})^2U_{k,i}^{'}) + p_k\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}\chi_{g_{k,j}=m}E(U_{k,i}\varepsilon_{k,i}^{\text{causal}}\varepsilon_{k,j}^{\text{causal}}U_{k,j}^{'})\right) = \lim_{k\longrightarrow\infty}\Delta_k^{\text{Cluster}} \rightarrow \Delta^{\text{Cluster}}.$$

The difference between the two matrices can be expressed as:

$$\Delta^{\mu} = \Delta^{\text{Cluster}} - \Delta^{\text{CCV}} = \lim_{k\rightarrow\infty}\frac{1}{n_k}\sum_{i=1}^{G_k} E[V_{k,m}]E[V_{k,m}]^{'}$$

which is positive semidefinite.

Assumption 2.10. Δ^{CCV} and Δ^{Cluster} exist and are positive semidefinite.

Theorem 2.2. Let $\Gamma = \Omega^{WW}$. Under assumptions 2.1-2.4 and assumptions 2.6-2.10, or under the same set of assumptions with assumption 2.5 replacing assumption 2.3,

$$\sqrt{N_k}(\hat{\theta}_k - \theta_k^{\text{causal}}) \xrightarrow{d} N\left(0, \Gamma^{-1}\left(q\Delta^{\text{CCV}} + (1-q)\Delta^{\text{Cluster}}\right)\Gamma^{-1}\right).$$

When $q = 0$, the inner term of the asymptotic variance reduces to the limit of the expected outer product. This is precisely what is estimated by the usual cluster-robust variance estimator and therefore this estimator will not be conservative in this case. This is a reasonable perspective in cases of panel data in which each individual constitutes their own cluster and only a negligible fraction of the population is observed. When $q \neq 0$, the difference between the usual cluster-robust variance object and the finite population variance is equal to $q\Gamma^{-1}(\Delta^{\text{Cluster}} - \Delta^{\text{CCV}})\Gamma^{-1}$ which is positive semidefinite.

2.6 The Variance Under Correct Specification

Suppose

Assumption 2.11.

$$Y_{k,i} = W_{k,i}\theta_k + X_{k,i}\gamma_{k,i} + \nu_{k,i}.$$

This assumption strengthens the linear model with treatment effects varying across cluster by imposing a constant treatment effect θ_k for every unit. Under this assumption, $\theta_k^{\text{causal}} = \theta_k$ and $E[U_{k,i}\varepsilon_{k,i}^{\text{causal}}] = 0 \; \forall i$. This in turn implies $\Delta^{\mu} = \Delta^{\text{Cluster}} - \Delta^{\text{CCV}} = 0$. This leads to the following result.

Theorem 2.3. Let $\Gamma = \Omega^{WW}$. Under assumptions 2.1-2.4 and assumptions 2.6-2.11, or under the same set of assumptions with assumption 2.5 replacing assumption 2.3, . Then,

$$\sqrt{N_k}(\hat{\theta}_k - \theta_k^{\text{causal}}) \xrightarrow{d} N(0, \Gamma^{-1}\Delta^{\text{Cluster}}\Gamma^{-1})$$

irrespective of q.

Since the typical cluster robust variance estimator consistently estimates Δ^{Cluster}, it is not conservative in this case.

3 Estimating the Variance

I now discuss the problem of estimating the variance of the causal estimand. In this section I will use the following short hand: $\mathcal{V} = \Gamma^{-1}(q\Delta^{\text{CCV}}+(1-q)\Delta^{\text{Cluster}})\Gamma^{-1}$, $\mathcal{V}^{\text{CCV}} = \Gamma^{-1}(\Delta^{\text{CCV}})\Gamma^{-1}$, and $\mathcal{V}^{\text{Cluster}} = \Gamma^{-1}(\Delta^{\text{Cluster}})\Gamma^{-1}$.

Instead of regression-based approaches taken by Abadie et al. (2020) and Xu and Yap (Working Paper, 2024) , I propose a direct estimator of the variance matrix similar to the estimator proposed by Abadie et al. (2023) . The results in Xu and Yap apply to cases that are not covered by the setting in this paper:

specifically, cases in which treatment distribution is allowed to be i.n.i.d. across all observations. Moreover, these authors provide results for clustered standard errors for a general class of M-estimators, not just OLS. Therefore, these results are not a strict improvement on the results discussed in that paper.

First, I discuss estimation of q, Γ^{-1} and Δ^{Cluster}. q is straightforward: if g_k is the number of sampled clusters, $\hat{q}_k = \frac{g_k}{G_k}$. This estimator depends on knowing the number of clusters in the population regardless of whether they are sampled or not. This is feasible in many cases; for instance, a researcher might observe a sample from ten U.S. states and therefore know $g_k = 10$ and $G_k = 50$. To estimate Γ^{-1}, calculate $\hat{U}_{k,i} = W_{k,i} - \hat{\Lambda}_{k,i} X_{k,i}$ and $\hat{\Gamma} = \frac{1}{N_k} \sum_{i=1}^{n_k} R_{k,i} \hat{U}_{k,i} \hat{U}_{k,i}'$.

Estimating Δ^{Cluster} is also straightforward and familiar. First, estimate residuals for the sampled units $\hat{\varepsilon}_{k,i} = Y_{k,i} - \hat{U}_{k,i}' \hat{\theta}_k - X_{k,i}' \hat{\gamma}_k$. Let $g_{k,i}$ indicate which cluster unit i in population k belongs to. Then

$$\hat{\Delta}_k^{\text{Cluster}} = \frac{1}{N_k} \sum_{m=1}^{G_k} \left[\left(\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} R_{k,i} \hat{U}_{k,i} \hat{\varepsilon}_{k,i} \right) \left(\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} R_{k,i} \hat{U}_{k,i} \hat{\varepsilon}_{k,i} \right)' \right]$$

Lemma 3.1. Let assumptions 2.1-2.4, 2.7 with $\delta > 4$ and 2.8-2.11, or under the above assumptions with 2.4 in place of 2.5. Then,

$$\hat{\Gamma}_k^{-1} \hat{\Delta}_k^{\text{Cluster}} \hat{\Gamma}_k^{-1} \xrightarrow{p} \mathcal{V}^{\text{Cluster}}.$$

Perhaps the most natural way to estimate Δ^{CCV} would be to estimate the difference $\Delta^{\text{Cluster}} - \Delta^{\text{CCV}} = \Delta^{\mu}$ directly and subtract it from the EHW estimate of Δ^{Cluster}. However, direct estimates of Δ^{μ} exhibit large estimation error in simulations, leading to substantial bias and even negative variance estimates in some cases. Therefore, I follow the approach of Abadie et al. (2023) and propose a very similar estimator that holds up in simulations.

One downside to the proposed estimator is that it exhibits substantial substantial upwards bias when the linear model is correctly specified; that is, when treatment effects are constant across units. Moreover, the estimator proposed relies on accurately estimating the within-cluster treatment effects. This requires both variation in within-cluster treatment and a large number of observations in each cluster. In this case, the usual cluster-robust errors consistently estimate the true population variance and should be used instead.

3.1 The Case With All Clusters Observed

First, I focus on the case in which all clusters are sampled. That is, $q_k = 1$. In this case,

$$\sqrt{N_k}(\hat{\theta}_k - \theta_{\text{causal}}) = \frac{1}{\sqrt{n_k p_k}}\Gamma^{-1}\sum_{m=1}^{G_k} C_{k,m} + o_p(1)$$
$$C_{k,g} = \sum_{i=1}^{n_k} \chi_{g_{k,i}=g} R_{k,i} U_{k,i} \varepsilon_{k,i}^{\text{causal}}.$$

As observed above,

$$E[C_{k,g}] = n_{k,g} p_k E[U_{k,g}^2](\theta_{k,g} - \theta_k^{\text{Causal}}).$$

Summing these terms across all clusters yields:

$$\sum_{m=1}^{G_k} E\left[C_{k,m}\right] = E[U_k^2]\sum_{m=1}^{G_k} p_k n_{k,g}(\theta_{k,m} - \theta_k^{\text{Causal}}) = 0$$

when treatment distribution is i.i.d.; when treatment distribution is allowed to vary across clusters

$$\sum_{m=1}^{G_k} E\left[C_{k,m}\right] = \sum_{m=1}^{G_k} E\left[U_{k,m} U_{k,m}^{'}\right] p_k n_{k,g}(\theta_{k,m} - \theta_k^{\text{Causal}}).$$

Isolating the second term, theorem 2.1 yields:

$$\sum_{m=1}^{G_k} p_k n_{k,g} E\left[U_{k,g} U_{k,g}^{'}\right]\theta_k^{\text{Causal}} = \sum_{m=1}^{G_k} p_k n_{k,g} E\left[U_{k,m} U_{k,m}^{'}\right] * E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} U_{k,i} U_{k,i}\right]^{-1} * \sum_{g=1}^{G_k} \frac{n_{k,g} E[U_{k,g} U_{k,g}]}{n_k}\theta_{k,g}$$
$$= p_k \sum_{g=1}^{G_k} n_{k,g} E[U_{k,g} U_{k,g}]\theta_{k,g}.$$

Therefore, in either case,

$$\sum_{m=1}^{G_k} E\left[C_{k,m}\right] = 0$$

The above shows that while the $E[C_{k,g}]$ terms are not generally equal to 0 (they will not be equal to 0 unless $\theta_{k,m} = \theta_k^{\text{Causal}}$ or $W_{k,g}$ is identically 0), their sum is equal to 0. Therefore,

$$\sqrt{N_k}(\hat{\theta}_k - \theta_{\text{causal}}) = \frac{1}{\sqrt{n_k}}\Gamma^{-1}\sum_{m=1}^{G_k}(C_{k,m} - E(C_{k,m})) + o_p(1)$$

Now observe that $C_{k,m} - E[C_{k,m}] = C_{k,m,1} + C_{k,m,2}$ where

$$C_{k,m,1} = \sum_{i=1}^{n_k} \chi_{g_{k,i}=m}(R_{k,i} - p_k)(\theta_{k,m} - \theta_k)E[U_{k,m}^2]$$

and

$$C_{k,m,2} = \sum_{i=1}^{n_k} \chi_{g_{k,i}=m} R_{k,i}(U_{k,i}\varepsilon_{k,i}^{\text{causal}} - (\theta_{k,m} - \theta_k)E[U_{k,m}^2])$$

and

$$\sqrt{N_k}(\hat{\theta}_k - \theta_{\text{causal}}) = \frac{1}{\sqrt{n_k p_k}}\Gamma^{-1}\sum_{m=1}^{G_k}(C_{k,m,1} + C_{k,m,2})) + o_p(1).$$

$C_{k,m,1}$ and $C_{k,m,2}$ have mean 0, are uncorrelated with each other, and are uncorrelated across clusters. The variance of $\frac{\Gamma^{-1}C_{k,m,1}}{\sqrt{n_k p_k}}$ is

$$(1 - p_k)\sum_{m=1}^{G_k}\frac{n_{k,m}}{n_k}(\theta_{k,m} - \theta_k)^2$$

and can be directly estimated by replacing θ_k with the OLS estimate across the sample, replacing $\theta_{k,m}$ with the OLS estimate across each cluster, and replacing p_k with the estimate of the sampling probability $\frac{N_k}{n_k}$. This requires knowledge of the population size n_k.

The variance of $\frac{\Gamma^{-1}C_{k,m,2}}{\sqrt{N_k}}$ is

$$\Gamma^{-1}\sum_{m=1}^{G_k}(p_{k,g} - p_{k,g}^2)(\theta_{k,m} - \theta_k)^2 E[U_{k,m}U_{k,m}^{'}]^2 * \Gamma^{-1}.$$

Define $\hat{\Gamma}_{k,g} = \frac{1}{n_{k,g}}\sum_{k=1}^{n_k}\chi_{g_{k,i}=g}U_{k,i}U_{k,i}^{'}$ The variance of $\sum_{m=1}^{G_k} C_{k,m} - E[C_{k,m}]$ can be directly estimated by:

$$\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}R_{k,i}\left(\hat{U}_{k,i}\hat{\epsilon}_{k,i} - (\hat{\theta}_{k,m} - \hat{\theta}_k)\hat{\Gamma}_{k,g}\right)\right)^2.$$

However, this direct estimator is biased in practice because the estimation error of its components are correlated. Therefore, I employ the sample-splitting approach proposed by Abadie et al. (2023) to correct this bias. The sample-splitting technique is described in the steps below:

1. Split the sample randomly into two subsamples. Let $Z_{k,i} \in \{0,1\}$ indicate whether unit i belongs to

the second subsample and let $\overline{Z}_k$ be the mean of $Z_{k,i}$

2. Using the subsample with $Z_{k,i} = 0$, obtain estimates $\hat{\theta}^*_{k,m}$, $\hat{\gamma}^*_k$, and $\hat{\theta}^*_k$ of $\theta_{k,m}$, γ_k, and θ^{causal}_k respectively.

3. For observations with $Z_{k,i} = 1$, calculate $\hat{\varepsilon}^*_{k,i} = Y_{k,i} - \hat{U}'_{k,i}\hat{\theta}^*_k - X'_{k,i}\hat{\gamma}^*_k$

Then, the variance can be estimated with

$$\begin{aligned}\hat{\mathcal{V}}^{\text{CCV}}_k = \frac{1}{N_k}\sum_{m=1}^{G_k}\hat{\Gamma}_k^{-1}\Bigg[\frac{1}{\overline{Z}_k^2}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}R_{k,i}Z_{k,i}\left(\hat{U}_{k,i}\hat{\varepsilon}^*_{k,i} - (\hat{\theta}_{k,m} - \hat{\theta}_k)\hat{\Gamma}_{k,g}\right)\right)\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}R_{k,i}Z_{k,i}\left(\hat{U}_{k,i}\hat{\varepsilon}^*_{k,i} - (\hat{\theta}_{k,m} - \hat{\theta}_k)\hat{\Gamma}_{k,g}\right)\right)'\\ - \frac{1-\overline{Z}_k}{\overline{Z}_k^2}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}R_{k,i}Z_{k,i}\left(\hat{U}_{k,i}\hat{\varepsilon}^*_{k,i} - (\hat{\theta}_{k,m} - \hat{\theta}_k)\hat{\Gamma}_{k,g}\right)\left(\hat{U}_{k,i}\hat{\varepsilon}^*_{k,i} - (\hat{\theta}_{k,m} - \hat{\theta}_k)\hat{\Gamma}_{k,g}\right)'\\ + (1-\hat{p}_k)\sum_{m=1}^{G_k}\frac{N_{k,m}}{N_k}(\hat{\theta}_{k,m} - \hat{\theta}_k)(\hat{\theta}_{k,m} - \hat{\theta}_k)'.\end{aligned}$$

For clusters with no variation in the treatment in the entire sample, I replace $\hat{\theta}_{k,m}$ with $\hat{\theta}_k$. For clusters with no variation in the treatment in the relevant split, I replace $\hat{\theta}^*_{k,m}$ with $\hat{\theta}^*_k$.

To improve the precision of the sample-splitting estimator, I re-estimate this object using an average across multiple splits (redrawing $Z_{k,i}$ each time). In the simulations presented below, the variance estimator is re-estimated four times with $E[\bar{Z}_k] = .5$ in each case.

3.2 The Case When Not All Clusters are Sampled

Theorem 2.2 shows that

$$\mathcal{V} = q\mathcal{V}^{\text{CCV}} + (1-q)\mathcal{V}^{\text{Cluster}}.$$

When $q = 1$, $\mathcal{V} = \mathcal{V}^{\text{CCV}}$ so $\hat{\mathcal{V}}^{\text{CCV}}$ is an estimator of $\mathcal{V}$. However, when $q < 1$, the convex combination form of $\mathcal{V}$ suggests the following estimator which estimates the weights of this convex combination:

$$\hat{\mathcal{V}}_k = \hat{q}_k\hat{\mathcal{V}}^{\text{CCV}} + (1-\hat{q}_k)\hat{\mathcal{V}}^{\text{Cluster}}_k.$$

In order to consistently estimate each element of $\hat{\mathcal{V}}^{\text{CCV}}$, the number of observations in each cluster must increase with k.

Assumption 3.1.

$$n_{k,g} \to \infty \quad \forall g$$

Theorem 3.1. Let assumptions 2.1-2.4, 2.7 with $\delta > 4$, 2.8-2.10, and 2.12 hold, or let the above assumptions hold with 2.4 in place of 2.5. Then,

$$\hat{\mathcal{V}}_k \xrightarrow{p} \mathcal{V}$$

4 Superpopulation Approach

The results presented so far in this paper are based on the results presented in a finite-population setting in Abadie et al. (2023). The above section shows that when a significant fraction of a population's clusters are sampled, the population variance is a convex combination between a variance matrix consistently estimated by the usual cluster-robust variance estimator and a smaller matrix. However, the size of the diagonal terms of that latter matrix depend on the unit sampling rate p_k. In particular, when a significant fraction of a population's units are sampled, the diagonal terms are smaller.

This section examines how important it is to observe a significant fraction of a population's units in order to obtain a smaller, valid standard errors. This is important because in many settings it is considerably easier to observe a significant number of clusters than it is to observe a significant number of observations. I show in this section than in these settings substantial variance reduction can be obtained. This section also bridges the gap between existing theory on sampling from an infinite population and the developing theory on finite populations. Finally, the results from this section tease out potential issues with how sampling from clusters is usually modeled.

4.1 Finite-Population Model

Consider a sequence of populations indexed by k. Each population is partioned into G_k clusters. The researcher observes a sample N_k from each population. For each sampled unit, indexed i, the researcher observes its cluster $g_{k,i}$, a $L \times 1$ set of covariates $X_{k,i}$, an outcome $Y_{k,i}$, and a treatment variable $K \times 1$ $W_{k,i}$. All of these variables can be continuous, discrete, or mixed.

Let $n_{k,g}$ denote the number of sampled observations belonging to each cluster. Define $X_{k,g}$, $Y_{k,g}$, and $W_{k,g}$ to be the matrices containing the controls, outcomes, and treatments, respectively, for all units in cluster g.

Suppose the researcher is interested in the parameters of the following linear model:

$$Y_{k,g} = W'_{k,g}\theta_{k,g} + X'_{k,g}\gamma_k + \nu_{k,g}$$

where $\theta_{k,g}$ and $\gamma_{k,g}$ are $K \times 1$ and $L \times 1$ vectors of parameters, respectively, and $\nu_{k,g}$ is a $n_{k,g}$ length vector of random shocks. Notably, the linear treatment effects $\theta_{k,g}$ are allowed to vary across cluster.

4.2 Sampling

Sampling follows a very similar process to the process discussed in the previous section, with the omission of unit sampling. Suppose samples $n_{k,g}$ are taken from each cluster.

Then, clusters are sampled with probability $q_k \in (0,1]$. If $q_k = 1$, every cluster is sampled. The researcher observes $Q_{k,g}$, the vector that indicates the sampling status for each group. If $Q_{k,g} = 1$, the observations within that cluster are kept. Otherwise, they are discarded. Therefore, $N_{k,g}$, the number of samples a research observes within each cluster, is either $n_{k,g}$ or 0 and $N_k = \sum_{m=1}^{G_k} Q_{k,m} n_{k,g}$. Cluster sampling q_k continues to satisfy the appropriate components of assumption 2.1, listed below.

Assumption 4.1. i) There exists a sequence of sampling probabilities and q_k s.t.

$$\Pr(Q_k = s) = q_k^{\sum_{i=1}^{G_k} s_i}(1-q_k)^{G_k - \sum_{i=1}^{G_k} s_i}$$

for all vectors s of length G_k, where $s_i \in \{0,1\}$. ii) The sequence of sampling rates q_k satisfies $G_k q_k \to \infty$ and $q_k \to q$. q is contained in the interval $[0,1]$.

4.3 Identification

Clustering can be modeled either as correlated treatment as in 2.3 or as correlated errors as in 2.3. Moreover, in this section, I will use the notation of $U_{k,i}, P_k$ and Ω_k exactly as defined in section 2.3, with the minor modification that the sampling indicator $R_{k,i}$ in $\tilde{P}_k$ and $\tilde{\Omega}_k$ only incorporates cluster sampling. I provide a similar result to Lemma 2.1:

Lemma 4.1. Under assumptions 4.2, 2.3, 2.4, and 2.7, or assumptions 4.2, 2.5, and 2.7 $\tilde{P}_k \xrightarrow{p} \Omega_k$, $\tilde{\Omega}_k \xrightarrow{p} \Omega_k$, and $\tilde{P}_k \xrightarrow{p} P_k$.

Then, θ_k^{causal} is defined identically to the matrix of moments defined in the finite-population section, except now sampling is only a function of the cluster sampling device Q_k. The next result demonstrates how interpretation of this estimand differs across settings.

Theorem 4.1. Under assumptions 4.1, 2.1, 2.3, 2.4, and 2.9

$$\theta_k^{\text{causal}} = \sum_{m=1}^{G_k} \frac{n_{k,m}}{n_k} \theta_{k,m}$$

If assumption 2.3 is replaced by assumption 2.5

$$\theta_k^{\text{causal}} = E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} U_{k,i}U_{k,i}^{'}\right]^{-1} \sum_{m=1}^{G_k} \frac{n_{k,m}}{n_k} E[U_{k,m}U_{k,m}^{'}]\theta_{k,m}.$$

As before, if treatment assignment $W_{k,i}$ is i.i.d. conditional on cluster assignment and if $X_{k,i}$ is i.i.d. conditional on cluster assignment, the above can be written as

$$\theta_k^{\text{causal}} = \frac{1}{n_k}\sum_{i=1}^{n_k} \theta_{k,i}$$

where $\theta_{k,i}$ in this setting is defined as the treatment effect for the cluster to which an observation belongs. For instance, $\theta_{k,i} = \theta_{k,g}$ if unit i belongs to cluster g. The OLS estimator is defined using the same notation as in the superpopulation setting using the partitions of $\tilde{W}_k$ that incorporate unit sampling as well as cluster sampling.

4.4 Asymptotic Variance

The key difference between the superpopulation setting and the finite population setting is the omission unit sampling. This section shows that the lack of unit sampling affects the estimation of the population variance. In particular, the unit variance becomes a special case of the previous section with $p_k = 1$

Define the following limits of the population variances and covariances:

$$\lim_{k\longrightarrow\infty} \frac{1}{n_k}\sum_{m=1}^{G_k}\sum_{m=1}^{G_k} Var(V_{k,m}) = \lim_{k\longrightarrow\infty} \Delta_k^{\text{causal}} \rightarrow \Delta^{\text{causal}}$$

and

$$\lim_{k\to\infty} \frac{1}{n_k}\sum_{m=1}^{G_k} E[V_{k,m}V_{k,m}^{'}]. = \lim_{k\longrightarrow\infty} \Delta_k^{\text{Cluster}} \rightarrow \Delta^{\text{Cluster}}.$$

The difference between the two matrices can be expressed as:

$$\Delta^{\mu} = \Delta^{\text{Cluster}} - \Delta^{\text{CCV}} = \lim_{k\to\infty} \frac{1}{n_k}\sum_{i=1}^{G_k} E[V_{k,m}]E[V_{k,m}]^{'}$$

Assumption 4.2. Δ^{Cluster} and Δ^{CCV} exist and are positive semidefinite.

which is positive semidefinite.

Theorem 4.2. Under assumptions 4.1,4.2, 2.3, 2.4, 2.7-2.10, and assumption 4.2:

$$\sqrt{N_k}(\hat{\theta}_k - \theta_k^{\text{causal}}) \xrightarrow{d} N(0, \Gamma^{-1}(q\Delta^{\text{CCV}} + (1-q)\Delta^{\text{Cluster}})\Gamma^{-1}).$$

4.5 Estimation

Estimation proceeds in a very similar fashion to the previous section. For instance, $\hat{\Delta}_k^{\text{Cluster}}$ in this section is defined as the special case where $p_k = 1$, and therefore the direct estimator still consistently estimates this component.

Lemma 4.2. Under assumptions 4.1,4.2, 2.3, 2.4, 2.7 with $\delta > 4$, 2.8-2.10, 2.12, and assumption 4.2

$$\hat{\mathcal{V}}_k^{\text{Cluster}} \to \Gamma^{-1}(\Delta^{\text{Cluster}})\Gamma^{-1}.$$

To estimate Δ^{CCV}, I focus on the case in which all clusters are sampled. That is, $q_k = 1$.

$$\sqrt{N_k}(\hat{\theta}_k - \theta_{\text{causal}}) = \frac{1}{\sqrt{N_k}}\Gamma^{-1}\sum_{m=1}^{G_k} C_{k,m} + o_p(1)$$
$$C_{k,g} = \sum_{i=1}^{N_k} \chi_{m_{k,i}=g} U_{k,i}\varepsilon_{k,i}^{\text{causal}}.$$

Then

$$E[C_{k,g}] = n_{k,g}E[U_{k,g}^2](\theta_{k,g} - \theta_k^{\text{causal}}).$$

Summing these terms across all clusters yields:

$$\sum_{m=1}^{G_k} E\left[C_{k,m}\right] = E[U_k^2]\sum_{m=1}^{G_k} n_{k,g}(\theta_{k,m} - \theta_k^{\text{Causal}}) = 0$$

when treatment distribution is i.i.d.; when treatment distribution is allowed to vary across clusters

$$\sum_{m=1}^{G_k} E\left[C_{k,m}\right] = \sum_{m=1}^{G_k} E\left[U_{k,m}U_{k,m}'\right] p_{k,m}n_k(\theta_{k,m} - \theta_k^{\text{Causal}}).$$

Isolating the second term, theorem 2.1 yields:

$$\sum_{m=1}^{G_k} E\left[U_{k,g}U_{k,g}^{'}\right]\theta_k^{\text{Causal}} = \sum_{m=1}^{G_k} p_{k,m}n_k E\left[U_{k,m}U_{k,m}^{'}\right] * E\left[\frac{1}{n_k}\sum_{i=1}^{n_k} U_{k,i}U_{k,i}\right]^{-1} * \sum_{g=1}^{G_k}\frac{n_{k,g}E[U_{k,g}U_{k,g}]}{n_k}\theta_{k,g}$$
$$= \sum_{g=1}^{G_k} n_{k,g}E[U_{k,g}U_{k,g}]\theta_{k,g}.$$

Therefore, in either case,

$$\sum_{m=1}^{G_k} E\left[C_{k,m}\right] = 0$$

The above shows that while the $E[C_{k,g}]$ terms are not generally equal to 0 (they will not be equal to 0 unless $\theta_{k,m} = \theta_k^{\text{Causal}}$ or $W_{k,g}$ is identically 0), their sum is equal to 0. Therefore,

$$\sqrt{N_k}(\hat{\theta}_k - \theta_{\text{causal}}) = \frac{1}{\sqrt{n_k}}\Gamma^{-1}\sum_{m=1}^{G_k}(C_{k,m} - E(C_{k,m})) + o_p(1)$$

Now observe that

$$C_{k,m} - E[C_{k,m}] = \sum_{i=1}^{n_k}\chi_{g_{k,i}=m}(U_{k,i}\varepsilon_{k,i}^{\text{causal}} - (\theta_{k,m} - \theta_k)E[U_{k,m}^2]).$$

Notice that this is exactly the $C_{k,m,2}$ term defined the finite population section and can be estimated in exactly the same way through an identical sample-splitting process. The final estimator is

$$\tilde{\mathcal{V}}_k^{\text{CCV}} = \frac{1}{N_k}\sum_{m=1}^{G_k}\hat{\Gamma}^{-1}\left[\frac{1}{\overline{Z}_k^2}\left(\sum_{i=1}^{n_k}\chi_{m_{k,i}=m}R_{k,i}Z_{k,i}\left(\hat{U}_{k,i}\hat{\varepsilon}_{k,i}^{*} - (\hat{\theta}_{k,m} - \hat{\theta}_k)\hat{\Gamma_{k,g}^{'}}\right)\right)^2\right.$$
$$\left. - \frac{1-\overline{Z}_k}{\overline{Z}_k^2}\sum_{i=1}^{n_k}\chi_{m_{k,i}=m}R_{k,i}Z_{k,i}\left(\hat{U}_{k,i}\hat{\varepsilon}_{k,i}^{*} - (\hat{\theta}_{k,m} - \hat{\theta}_k)\hat{\Gamma}\right)^2\right]\hat{\Gamma}_{k,g}^{-1}$$

where estimates marked with an asterisk indicate that they are estimated via sampling splitting (see section 4.4 for more details).

When $q_k < 1$, the correct estimator for the population variance is given by

$$\tilde{\mathcal{V}}_k = \hat{q}_k\tilde{\mathcal{V}}^{\text{CCV}} + (1-\hat{q}_k)\tilde{\mathcal{V}}_k^{\text{Cluster}}.$$

Theorem 4.3. Under assumptions 4.1,4.2, 2.3, 2.4, 2.7 with $\delta > 4$, 2.8-2.10, 2.12, and assumption 4.2, or the same set of assumption with assumption 2.3 replaced by assumption 2.5,

$$\tilde{\mathcal{V}}_k \xrightarrow{p} \mathcal{V}.$$

4.6 Discussion

$\tilde{\mathcal{V}}^{\text{CCV}}$ is almost identical to $\hat{\mathcal{V}}^{\text{CCV}}$; the only difference is the term

$$(1-\rho_k)\sum_{m=1}^{G_k}\frac{N_{k,m}}{N_k}(\theta_{k,m}-\theta_k)^2$$

which is only present in the former and depends upon the sampling rate redefined here for clarity as p_k, the fraction of observations sampled from the population. In the case of an infinite population, the sampling sizes in each cluster $n_{k,g}$ are assumed to be nonrandom. Therefore, $\hat{\mathcal{V}}^{\text{CCV}_k}$ does not include this term.

This poses some conceptual challenges. The case $p_k \to 0$ is supposed to reflect the case where a negligible fraction of a population's units is sampled. In the infinite population case, a negligible fraction of a population's units are always sampled, yet the results reflect the case when $p_k \to 1$. This is an artifact of how cluster sampling is typically modeled, where cluster sample sizes are treated as fixed components of sampling design. Therefore, the only two realizations of the variable $N_{k,g}$ representing sampled cluster observations are $n_{k,g}$ or 0. In the finite population case, each observation has an individual variable determining sampling status. This adds additional sampling uncertainty.

A possible solution to bridge the gap between the results and intuition is to suppose each sampled unit has some probability of belonging to cluster g, $\rho_{k,g}$. In this case, the number of sampled units within each cluster will demonstrate similar behavior to unit sampling in the finite-population case. However, because this approach introduces some dependence in observations between clusters. I leave further discussion of this issue and potential solutions to it for future work.

5 Simulations

Simulations demonstrating the results of this paper in the finite-population case proceed according to the following two-step process. Step one produces population features that are constant across all iterations. Step two produces population features that are randomized across iterations.

Step 1 Performed once per parameter set.

1. Separate observations into G clusters.
2. Generate mean treatment effect by cluster from uniform distribution with variance $\sigma^2_{\tau_g}$.
3. Generate individual treatment effects θ_i by cluster from a unit-variance normal distribution centered at the corresponding cluster mean.

Step 2 These steps are iterated 10000 times per parameter set:

- Generate an n-vector of attributes $X_{k,i} \sim N(0,1)$.
- Generate random shocks $\xi_{k,i} \sim N(0,1)$
- Generate $A_{k,g} \sim U[0, \sqrt{12}]$. Generate n/G-vector $W_{k,g} \sim N(A_{k,g}, 0)$
- Generate outcome vector $Y_{k,i}$ using equation $Y^*_{k,i} = W_{k,i}\theta_{k,i} + \xi_{k,i}$
- Randomly sample clusters and observations within clusters with probabilities q and p, respectively.
- Calculate $\hat{\theta}$, $\hat{V}^{\text{CCV}}$, and $\hat{V}^{\text{Cluster}}$
- Check whether the estimated 95 % intervals around $\hat{\tau}$ using $\hat{V}^{\text{CCV}}$, and $\hat{V}^{\text{Cluster}}$ cover τ^{Causal}

The results are presented in Table 1. The process for generating simulations for the infinite population case proceed in exactly the same way, except the mean cluster treatment effect is assigned to every unit in the cluster. That is, individuals do not possess their own idiosyncratic treatment effect. Moreover, unit sampling does not occur. These results are presented in Table 2.

6 Conclusion

This paper builds on the literature examining finite-population settings. In particular, it shows that in the finite-population setting when clustering is continuous, clustering need not depend on the presence of dependent structure in the error term, but may also be created by dependence in the treatment and in the sampling scheme. I analyze the asymptotic variance of the OLS estimator in this setting and show

that the usual cluster-robust standard errors are conservative estimates of the true population variance. I propose a new variance estimator that robustly estimates the true variance matrix.

This paper only considers very simple variable probability weighting schemes in which every observation and cluster is sampled with equal probability. Future work would extend the results of this paper by analyzing the population variance when the unit and cluster sampling probabilities are allowed to vary across clusters. Moreover, sometimes variable probability schemes are determined by observed attributes of each observation. This case should be explored in future work as well.

This paper only discusses one-way clustering in a fairly narrow context. Future research would investigate how the multi-way clustering is impacted in finite-setting clusters with treatment effect heterogeneity.

The performance of the proposed variance estimator depends on the presence of treatment effect heterogeneity across cluster. When there is little treatment effect heterogeneity, the estimator tends to be overly conservative. Future research would investigate why this occurs and propose formal tests that guide researchers as to whether or not the method proposed in this paper should be used. Moreover the estimator is not guaranteed to be positive; in some simulation cases, a negative estimate is rarely produced. Future work would perform a thorough investigation of the properties of this estimator.

7 Figures

Table 1: Error Clustering Simulation Results

n	10000	150000	100000	50000	100000	10000
q	1	.5	1	1	1	1
G	10	150	100	50	100	10
$\sigma^2_{\tau_g}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{25}{3}$	$\frac{25}{3}$	0
MC Standard Errors	0.016	0.05	0.005	0.02	0.015	0.014
$(\bar{V}^{\text{Cluster}}/N)^{1/2}$	0.171	0.068	0.056	0.436	0.323	0.013
$(\bar{V}^{\text{CCV}}/N)^{1/2}$	0.022	0.048	0.007	0.021	0.016	0.02
$\frac{(\bar{V}^{\text{CCV}}/N)^{1/2}}{(\bar{V}^{\text{Cluster}}/N)^{1/2}}$	0.126	0.71	0.121	0.048	0.048	1.587
CR using $\bar{V}^{\text{Cluster}}$	1	0.993	1	1	1	0.911
CR using $\bar{V}^{\text{CCV}}$	0.977	0.937	0.989	0.956	0.95	0.987

Table 2: Simulation Results

n	10000	150000	100000	50000	100000	10000
ρ	.5	.5	.5	.5	.5	.5
q	1	.5	1	1	1	1
G	10	150	100	50	100	10
$\sigma^2_{\tau_g}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{25}{3}$	$\frac{25}{3}$	0
$(\bar{V}^{\text{Cluster}}/N)^{1/2}$	0.179	0.094	0.079	0.605	0.386	0.027
$(\bar{V}^{\text{CCV}}/N)^{1/2}$	0.147	0.083	0.059	0.429	0.272	0.067
MC Standard Errors	0.135	0.083	0.056	0.435	0.268	0.028
$\frac{(\bar{V}^{\text{Cluster}}/N)^{1/2}}{(\bar{V}^{\text{CCV}}/N)^{1/2}}$	0.82	0.878	0.745	0.708	0.703	2.491
CR using $\bar{V}^{\text{Cluster}}$	0.971	0.971	0.992	0.988	0.994	0.91
CR using $\bar{V}^{\text{CCV}}$	0.97	0.943	0.947	0.954	0.948	1

8 Appendix and Proofs

Proof of Lemma 2.1. I begin by showing $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$. First notice that since

$$E\left[\frac{N_k}{n_k q_k p_k}\right] = 1$$

and

$$\text{Var}\left[\frac{\sum_{m=1}^{G_k} Q_{k,m} \sum_{i=1}^{n_k} R_{k,i}}{q_k p_k n_k}\right] = \frac{\sum_{m=1}^{G_k} q_k \left(n_{k,m}(p_k)(1-p_k)\right)}{q_k^2 p_k^2 n_k^2} + \frac{\sum_{m=1}^{G_k} q_k n_{k,m}^2 (1-q_k)(p_k)^2}{q_k^2 p_k^2 n_k^2}.$$

The first term converges to 0. Analyzing the second term further:

$$\frac{\sum_{m=1}^{G_k} q_k n_{k,m}^2 (1-q_k)(p_k)^2}{q_k^2 p_k^2 n_k^2} \leq \frac{\sum_{m=1}^{G_k} n_{k,m}^2}{q_k n_k^2} \leq \frac{\sum_{m=1}^{G_k} n_{k,m} * \max_g(n_k, g)}{q_k n_k^2}$$
$$\leq \frac{\max_g(n_{k,g})}{\min_g n_{k,g}} * \frac{1}{G_k q_k}.$$

Since $\limsup_{k\to\infty} \frac{\max_g(n_{k,g})}{\min_g n_{k,g}} < \infty$ and $G_k q_k \to \infty$, this term converges to 0 as well. Therefore,

$$\text{Var}\left[\frac{\sum_{m=1}^{G_k} Q_{k,m} \sum_{i=1}^{n_k} R_{k,i}}{q_k p_k n_k}\right] \to 0$$

and $\frac{N_k}{q_k p_k n_k} \to 1$.

This implies, by the continuous mapping theorem, $\left(\frac{p_k q_k n_k}{N_k}\right)^{1/2} \to 1$ and it suffices to show that

$$\frac{1}{n_k q_k p_k} \sum_{m=1}^{G_k} \sum_{i=1}^{n_k} R_{k,i} E\left[\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix}'\right] - \Omega_k = 0.$$

Since $E[R_{k,i}] = qp$. Since each of these terms are uniformly integrable by 2.7 and since assumption 1 of Hansen and Lee (2019) is assumed in assumption 2.1, apply Theorem 1 to obtain:

$$\frac{1}{n_k q_k p_k} \sum_{m=1}^{G_k} \sum_{i=1}^{n_k} R_{k,i} E\left[\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix}'\right] - \Omega_k \xrightarrow{p} \frac{1}{n_k q_k p_k} \sum_{m=1}^{G_k} \sum_{i=1}^{n_k} q_k p_k E\left[\begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ \nu_{k,i} \end{pmatrix}'\right] - \Omega_k$$
$$= \Omega_k - \Omega_k.$$

By another application Theorem 1 of Hansen and Lee, $P_k \to \Omega_k$. Finally, to show $\tilde{P}_k$ converges to Ω_k,

note again that it suffices to show that

$$\frac{1}{n_k p_k q_k} \sum_{i=1}^{n_k} R_{k,i} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ v_{k,i} \end{pmatrix} \begin{pmatrix} Y_{k,i} \\ U_{k,i} \\ X_{k,i} \\ v_{k,i} \end{pmatrix}'$$

□

converges to Ω_k. Since $E[R_{k,i}] = p_k q_k$ and since the sampling probability is independent of the realizations of any of the variables, the result is obtained.

Proof. □

Proof of Theorem 2.1. Since $\Omega_k^{UX} = 0$,

$$\begin{aligned} \theta_k^{\text{causal}} &= (\Omega_k^{UU})^{-1} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} E(U_{k,i} Y_{k,i}) \right) \\ &= (\Omega_k^{UU})^{-1} \left(\frac{1}{n_k} \sum_{m=1}^{G_k} \sum_{i=1}^{n_k} \chi_{g_{k,i}=m} \Big[E(U_{k,m} U_{k,m}) \theta_{k,i} + E(U_{k,m} X_{k,m}) \gamma_k + E(U_{k,m} v_{k,i}) \Big] \right) \end{aligned}$$

Since $\sum_{i=1}^{n_k} E(U_{k,i} X_{k,i}) = 0$ and $\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} E(U_{k,m} v_{k,i}) = 0$,

$$\theta_k^{\text{causal}} = (\Omega_k^{UU})^{-1} \left(\frac{1}{n_k} \sum_{m=1}^{G_k} \sum_{i=1}^{n_k} \chi_{g_{k,i}=m} E(U_{k,m} U_{k,m}') \theta_{k,i} \right)$$

Under assumption 2.3 and identically distributed treatment, $E(U_{k,m} U_{k,m}') = \Omega_k^{UU}$ and

$$\theta_k^{\text{causal}} = \frac{1}{n_k} \sum_{i=1}^{n_k} \theta_{k,i}$$

Under assumption 2.5 and treatment distribution varying across cluster,

$$\begin{aligned} \theta_k^{\text{causal}} &= (\Omega_k^{UU})^{-1} \left(\frac{1}{n_k} \sum_{m=1}^{G_k} n_{k,m} E(U_{k,m} U_{k,m}') \theta_{k,m} \right) \\ &= (\Omega_k^{UU})^{-1} \left(\sum_{m=1}^{G_k} \frac{n_{k,m}}{n_k} E(U_{k,m} U_{k,m}') \theta_{k,m} \right) \end{aligned}$$

□

Lemma 8.1. Let $V_{n,i}$ be a row-wise independent triangular array and $\mu_{n,i} = E[V_{n,i}]$. Suppose that $Q_{k,1}, Q_{k,2}, ...Q_{k,G_k}$ are independent of $V_{n,1}, V_{n,2}, ..., V_{n,G_k}$ and that Assumption 2.2 holds. $V_{k,m} = \sum_{i=1}^{n_k} \chi_{g_{k,i}=m} R_{k,i} D_{k,i}$ Assume that

$$\frac{1}{n} \sum_{m=1}^{G_k} E\left[\left|\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} D_{k,i}\right|^{2+\delta}\right]$$

is bounded for $\delta > 0$,

$$\sum_{m=1}^{G_k} \mu_{k,m} = 0,$$

$$\frac{1}{n_k} \sum_{m=1}^{G_k} \left(\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} \text{Var}(D_{k,i}) + (1-p_k) \sum_{i=1}^{n_k} \chi_{g_{k,i}=m} E(D_{k,i}^2) + 2p_k \sum_{i=1}^{n_k} \sum_{j \neq i} \chi_{g_{k,i}=m} \chi_{g_{k,j}=m} \text{Cov}(D_{k,i}, D_{k,j}) \right) = r_k \to r$$

and

$$\frac{1}{n_k} \sum_{m=1}^{G_k} \left(\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} E(D_{k,i}^2) + p_k \sum_{i=1}^{n_k} \chi_{g_{k,i}=m} \chi_{g_{k,j}=m} E(D_{k,i} D_{k,j}) \right) = s_k \to s,$$

where $q * r + (1-q)s > 0$.

$$\frac{1}{\sqrt{N_k}} \sum_{m=1}^{G_k} (Q_k V_{k,m}) \xrightarrow{d} \mathcal{N}(0, qr + (1-q)s),$$

where $N_k = \sum_{m=1}^{G_k} Q_k \sum_{i=1}^{n_k} R_{k,i}$.

Proof.

$$\text{Var}\left[\frac{\sum_{m=1}^{G_k} Q_{k,m} \sum_{i=1}^{n_k} R_{k,i}}{q_k p_k n_k}\right] = \frac{\sum_{m=1}^{G_k} q_k \left(n_{k,m}(p_k)(1-p_k)\right)}{q_k^2 p_k^2 n_k^2} + \frac{\sum_{m=1}^{G_k} q_k n_{k,m}^2 (1-q_k)(p_k)^2}{q_k^2 p_k^2 n_k^2}.$$

The first term converges to 0. Analyzing the second term further:

$$\begin{aligned} \frac{\sum_{m=1}^{G_k} q_k n_{k,m}^2 (1-q_k)(p_k)^2}{q_k^2 p_k^2 n_k^2} &\leq \frac{\sum_{m=1}^{G_k} n_{k,m}^2}{q_k n_k^2} \leq \frac{\sum_{m=1}^{G_k} n_{k,m} * \max_g(n_k, g)}{q_k n_k^2} \\ &\leq \frac{\max_g(n_{k,g})}{\min_g n_{k,g}} * \frac{1}{G_k q_k}. \end{aligned}$$

Since $\limsup_{k\to\infty} \frac{\max_g(n_{k,g})}{\min_g n_{k,g}} < \infty$ and $G_k q_k \to \infty$, this term converges to 0 as well. Therefore,

$$\text{Var}\left[\frac{\sum_{m=1}^{G_k} Q_{k,m} \sum_{i=1}^{n_k} R_{k,i}}{q_k p_k n_k}\right] \to 0$$

and $\frac{N_k}{q_k p_k n_k} \to 1$.

This implies, by the continuous mapping theorem, $\left(\frac{p_k q_k n_k}{N_k}\right)^{1/2} \to 1$. Then, it is sufficient to prove

$$\frac{1}{\sqrt{n_k}} \sum_{m=1}^{G_k} \frac{1}{\sqrt{q_k p_k}} Q_{k,m} V_{k,m} \xrightarrow{d} \mathcal{N}(0, qr + (1-q)s).$$

Notice,

$$\text{Var}\left(\frac{1}{\sqrt{n_k q_k p_k}} \sum_{m=1}^{G_k} Q_{k,m} V_{k,m}\right) = \frac{1}{q_k p_k n_k} \sum_{m=1}^{G_k} q_k \left(E(V_{k,m}^2) - q_k E(V_{k,m})^2\right) = \frac{1}{p_k n_k} \sum_{m=1}^{G_k} \left(q_k \text{Var}(V_{k,m}) + (1-q_k) E(V_{k,m}^2)\right).$$

By direct calculation,

$$\frac{1}{n_k p_k} \sum_{m=1}^{G_k} \text{Var}(V_{k,m}) = r_k$$
$$\frac{1}{n_k p_k} \sum_{m=1}^{G_k} E(V_{k,m}^2) = s_k$$

Let k be large enough such that s_k is bounded away from 0. Then, for all $i = 1, ..., n_k$

$$E\left[\frac{Q_{k,m} V_{k,m} - q_k \mu_{k,m}}{(q_k r_k + (1-q_k) s_k) \sqrt{n_k p_k q_k}}\right] = 0$$

and

$$Var\left[\frac{Q_{k,m} V_{k,m} - q_k \mu_{k,m}}{(q_k r_k + (1-q_k) s_k) \sqrt{n_k p_k q_k}}\right] = 1.$$

The rest of the argument follows from a Liapunov-type argument. □

Proof of Theorem 4.2. Consider $V_{k,g} = \sum_{i=1}^{n_k} a' \chi_{g_{k,i}=g} R_{k,i} U_{k,i} \varepsilon_{k,i}^{\text{causal}}$.

$$\frac{1}{n_k}\sum_{m=1}^{G_k} E\left[|\sum_{i=1}^{n_k} a^{'}\chi_{g_{k,i}=g}U_{k,i}\varepsilon_{k,i}^{\text{causal}}|^{2+\delta}\right] \leq \frac{||a||^{2+\delta}}{n_k}\sum_{m=1}^{G_k} E\left[\left|\sum_{k=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right|^{2+\delta}\right].$$

By Minkowski's inequality,

$$\begin{aligned}\frac{||a||^{2+\delta}}{n_k}\sum_{m=1}^{G_k} E\left[\left|\sum_{k=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right|^{2+\delta}\right]^{\frac{1}{2+\delta}} &\leq \frac{||a||^{2+\delta}}{n_k}\sum_{m=1}^{G_k}\sum_{i=1}^{n_k} E\left[\chi_{g_{k,i=m}}|U_{k,i}\varepsilon_{k,i}^{\text{causal}}|^{2+\delta}\right]^{\frac{1}{2+\delta}} \\ &\leq \frac{||a||^{2+\delta}}{n_k}\sum_{i=1}^{n_k} E[\chi_{g_{k,i}=m}||U_{k,i}||^{2+\delta}(|Y_{k,i}| + ||U_{k,i}||||\theta_{k,i}|| + ||X_{k,i}||||\gamma_k|| + |\varepsilon_{k,i}^{\text{causal}}|)^{2+\delta}]^{\frac{1}{2+\delta}}\end{aligned}$$

By Minkowski's equality and assumption 2.7, this term is bounded. In addition,

$$\sum_{m=1}^{G_k} a'\mu_{k,i} = \sum_{i=1}^{n_k} a^{'}E[U_{k,i}\varepsilon_{k,i}^{\text{causal}}] = 0.$$

When $a \neq 0$,

$$\begin{aligned}\frac{1}{n_k}\sum_{m=1}^{G_k}\text{Var}(V_{k,g}) = a'\left(\frac{1}{n_k}\sum_{m=1}^{G_k}\text{Var}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)\right)a \rightarrow a^{'}\Delta^{\text{CCV}}a \\ \frac{1}{n_k}\sum_{m=1}^{G_k}\mu_{k,i}^2 = a'\left(\frac{1}{n_k}\sum_{m=1}^{G_k}E\left[\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right]E\left[\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}\varepsilon_{k,i}^{\text{causal}}U_{k,i}^{'}\right]\right)a \rightarrow a^{'}\Delta^{\mu}a\end{aligned}$$

Then,

$$a'\left(\frac{1}{\sqrt{N}}\sum_{m=1}^{G_k} Q_{k,m}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)\xrightarrow{d} N(0, a'\left(\Delta^{\text{CCV}} + (1-q)\Delta^{\mu})a\right)$$

and by the Cramer-Wold theorem,

$$\left(\frac{1}{\sqrt{N}}\sum_{m=1}^{G_k} Q_{k,m}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)\xrightarrow{d} N(0, \left(\Delta^{\text{CCV}} + (1-q)\Delta^{\mu})\right).$$

$$\begin{pmatrix}\hat{\theta}_k^{\text{Causal}}\\ \hat{\gamma}_k^{\text{Causal}}\end{pmatrix} = \begin{pmatrix}\tilde{W}_k^{UU} & \tilde{W}_k^{UX}\\ \tilde{W}_k^{XU} & \tilde{W}_k^{XX}\end{pmatrix}^{-1} \frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\begin{pmatrix}U_{k,i}Y_{k,i}\\ X_{k,i}Y_{k,i}\end{pmatrix}$$
$$= \begin{pmatrix}\theta_k^{\text{Causal}}\\ \gamma_k^{\text{Causal}}\end{pmatrix} + \begin{pmatrix}\tilde{W}_k^{UU} & \tilde{W}_k^{UX}\\ \tilde{W}_k^{XU} & \tilde{W}_k^{XX}\end{pmatrix}^{-1} \frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\begin{pmatrix}U_{k,i}\epsilon_{k,i}^{\text{Causal}}\\ X_{k,i}\epsilon_{k,i}^{\text{Causal}}\end{pmatrix}.$$

Then write

$$\begin{pmatrix}\hat{\theta}_k^{\text{Causal}} - \theta_k^{\text{Causal}}\\ \hat{\gamma}_k^{\text{Causal}} - \gamma_k^{\text{Causal}}\end{pmatrix} = \begin{pmatrix}\tilde{W}_k^{UU} & \tilde{W}_k^{UX}\\ \tilde{W}_k^{XU} & \tilde{W}_k^{XX}\end{pmatrix}^{-1} \frac{1}{\sqrt{N_k}} R_{k,i}\begin{pmatrix}U_{k,i}\epsilon_{k,i}^{\text{Causal}}\\ X_{k,i}\epsilon_{k,i}^{\text{Causal}}\end{pmatrix}$$
$$= \begin{pmatrix}\Omega_k^{UU} & \Omega_k^{UX}\\ \Omega_k^{XU} & \Omega_k^{XX}\end{pmatrix}^{-1} \frac{1}{\sqrt{N_k}} R_{k,i}\begin{pmatrix}U_{k,i}\epsilon_{k,i}^{\text{Causal}}\\ X_{k,i}\epsilon_{k,i}^{\text{Causal}}\end{pmatrix} + r_{k,i}$$

where

$$r_{k,i} = \left[\begin{pmatrix}\Omega_k^{UU} & \Omega_k^{UX}\\ \Omega_k^{XU} & \Omega_k^{XX}\end{pmatrix} - \begin{pmatrix}\tilde{W}_k^{UU} & \tilde{W}_k^{UX}\\ \tilde{W}_k^{XU} & \tilde{W}_k^{XX}\end{pmatrix}\right]^{-1} \frac{1}{\sqrt{N_k}} R_{k,i}\begin{pmatrix}U_{k,i}\epsilon_{k,i}^{\text{Causal}}\\ X_{k,i}\epsilon_{k,i}^{\text{Causal}}\end{pmatrix}.$$

$\Omega_k^{UX} =$, the first term above is $o_p(1)$, and $\frac{1}{\sqrt{N_k}} R_{k,i} U_{k,i}\epsilon_{k,i}^{\text{Causal}}$ is $O_p(1)$ by the discussion above. Therefore,

$$\sqrt{N}(\hat{\theta}_k - \theta_k^{\text{Causal}}) = (\Omega_k^{UU})^{-1}\frac{1}{\sqrt{N_k}}\sum_{i=1}^{n_k} U_{k,i}\varepsilon_{k,i}^{\text{causal}}$$

if

$$\frac{1}{\sqrt{N_k}}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}X_{k,i}\varepsilon_{k,i}^{\text{causal}} = O_p(1).$$

Above I showed that

$$\left(\frac{p_k q_k n_k}{N_k}\right)^{1/2} \to 1.$$

Therefore, it suffices to show

$$\frac{1}{\sqrt{n_k}}\sum_{m=1}^{G_k}\frac{Q_{k,m}}{\sqrt{q_k}}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}\frac{R_{k,i}}{\sqrt{p_k}}X_{k,i}\epsilon_{k,i}^{\text{causal}} = O_p(1).$$

For ease of notation, assume that $L = 1$. For a larger set of controls, simply apply the following argument elementwise. These terms are bounded in expectation by Assumption 2.7. Therefore, all that is required is that the variance of this term is bounded by using Chebyshev's inequality. Because these terms have 0 covariance across m,

$$\begin{aligned}
\text{Var}\left(\frac{1}{\sqrt{n_k}}\sum_{m=1}^{G_k}\frac{Q_{k,m}}{\sqrt{q_k}}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}\frac{R_{k,i}}{\sqrt{p_k}}X_{k,i}\varepsilon_{k,i}^{\text{causal}}\right) &= \frac{1}{n_k}\sum_{m=1}^{G_k}\text{Var}\left(\frac{Q_{k,m}}{\sqrt{q_k}}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}X_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)\\
&\leq \frac{1}{n_k}\sum_{m=1}^{G_k}E\left[\left(\frac{Q_{k,m}}{\sqrt{q_k}}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}\frac{R_{k,i}}{\sqrt{p_k}}X_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)^2\right]\\
&= \frac{1}{n_k}\sum_{m=1}^{G_k}E\left[\sum_{i=1}^{n_k}(\chi_{g_{k,i}=m}\chi_{g_{k,j}=m}X_{k,i})^2 + p_k\sum_{i=1}^{n_k}\sum_{j\neq i}\chi_{g_{k,i}=m}\chi_{g_{k,j}=m}X_{k,i}\varepsilon_{k,i}^{\text{causal}}X_{k,j}\varepsilon_{k,j}^{\text{causal}}\right]
\end{aligned}$$

The last line follows since $E(Q_{k,m}) = q_k$ and $E(R_{k,i}) = p_k$. Now, by the same Minkowski's inequality arguments as above and Assumption 2.7, every component of this term is bounded. Since all that was required was to show that the variance was bounded, this completes the proof. □

Proof of Theorem 2.3. This follows directly from the fact that $E[\frac{1}{n_k}\sum_{i=1}^{n_k}\chi_{g_{k,i}=g}U_{k,i}\varepsilon_{k,i}^{\text{causal}}] = 0 \quad \forall g$ □

Proof of Lemma 3.1. First, notice that by lemma 2.1, $\hat{\Lambda}_k \to \Lambda_k$. Therefore

$$\hat{\Gamma}_k - \tilde{P}_k^{UU} = (\hat{\Lambda}_k - \Lambda_k)\tilde{P}^{XX}(\hat{\Lambda}_k - \Lambda_k)' - \tilde{P}_n^{WX}(\hat{\Lambda}_k - \Lambda_k)' - (\hat{\Lambda}_k - \Lambda_k)\tilde{P}_n^{XW} \xrightarrow{p} 0.$$

Lemma 2.1 and assumption 2.8 imply $\hat{\Gamma}_k - \Gamma \xrightarrow{p} 0$. Moreover, theorem 2.2 implies $\hat{\theta}_k \to \theta_k^{\text{causal}}$. Lemma 2.1 implies $\hat{\gamma}_k \to \gamma_k$. Define

$$\check{\Delta}_k^{\text{Cluster}} = \frac{1}{N_k}\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k}\chi_{g_k=m}R_{k,i}U_{k,i}\hat{\varepsilon}_{k,i}\right)\left(\sum_{i=1}^{n_k}\chi_{g_k=m}R_{k,i}U_{k,i}\hat{\varepsilon}_{k,i}\right)', \qquad \tilde{\Delta}_k^{\text{Cluster}} = \frac{1}{N_k}\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k}\chi_{g_k=m}R_{k,i}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)\left(\sum_{i=1}^{n_k}\chi_{g_k=m}R_{k,i}U_{k,i}\varepsilon_{k,i}^{\text{ca}}\right.$$

and

$$\Delta_k^{\text{Cluster}} = \frac{1}{n_k}\sum_{m=1}^{G_k}E\left[\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}(\varepsilon_{k,i}^{\text{causal}})^2U_{k,i}' + 2p_k\sum_{i=1}^{n_k}\sum_{j\neq i}^{n_k}\chi_{g_{k,i}=m}\chi_{g_{k,j}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\varepsilon_{k,j}^{\text{causal}}U_{k,j}'\right]$$

If assumption 2.7 holds with $\delta > 4$, then by theorem 1 of Hansen and Lee (2019) $\tilde{\Delta}_k^{\text{Cluster}} - \Delta_k^{\text{Cluster}} \xrightarrow{p} 0$. By the same argument and the convergence of $\hat{\theta}_k$ and $\hat{\gamma}_k$, $\hat{\Delta}_k^{\text{Cluster}} - \check{\Delta}_k^{\text{Cluster}} \xrightarrow{p} 0$ and $\check{\Delta}_k^{\text{Cluster}} - \tilde{\Delta}_k^{\text{Cluster}} \xrightarrow{p} 0$. Then: $\hat{\Delta}_k^{\text{Cluster}} - \Delta_k^{\text{Cluster}} = (\hat{\Delta}_k^{\text{Cluster}} - \check{\Delta}_k^{\text{Cluster}}) + (\check{\Delta}_k^{\text{Cluster}} - \tilde{\Delta}_k^{\text{Cluster}}) + (\tilde{\Delta}_k^{\text{Cluster}} - \Delta_k^{\text{Cluster}}) + (\Delta_k^{\text{Cluster}} - \Delta^{\text{Cluster}}) \xrightarrow{p} 0$ where the last difference goes to 0 by assumption 2.10. □

Proof of Theorem 3.1. First, notice that for variance estimators of the form:

$$E\left[\left(\sum_{i=1}^{n} V_i\right)^2\right]$$

based on a subsample consisting of units such that $Z_i = 1$ where Z_i is a binary, i.i.d. variable with $P(Z_i = 1) = p_z$ and independent of V_i, by section A.4 of Abadie et al. (2023):

$$E\left[\left(\sum_{i=1}^{n} V_i\right)^2\right] = \frac{1}{p_Z^2} E\left[\left(\sum_{i=1}^{n} Z_i V_i\right)^2\right] - \frac{1-p_z}{p_Z^2} \sum_{i=1}^{n} E[Z_i V_i^2].$$

Note that by assumption [check] $\hat{\theta}_{k,m} \xrightarrow{p} \theta_{k,m}$, $\hat{\theta}_k \xrightarrow{p} \theta_k$, and $\hat{\Gamma}_{k,m} \xrightarrow{p} \Gamma_m$. Let

$$V_i = R_{k,i}(U_{k,i}\varepsilon_{k,i}^{\text{causal}} - (\theta_{k,m} - \theta_m)\Gamma_g), \quad \check{V}_i = R_{k,i}(\hat{U}_{k,i}\hat{\varepsilon}_{k,i}^{*} - (\hat{\theta}_{k,m} - \hat{\theta}_m)\hat{\Gamma}_{k,g})$$

and

$$\tilde{V}_{k,i} = R_{k,i}(U_{k,i}\hat{\varepsilon}_{k,i}^{*} - (\hat{\theta}_{k,m} - \hat{\theta}_m)\Gamma_g).$$

Now, define

$$\begin{aligned}\Delta_k^{\text{CCV}} &= \frac{1}{n_k p_k q_k}\Gamma^{-1} E\left[\sum_{i=1}^{n_k} V_{k,i}^2\right]\Gamma^{-1} = \frac{1}{n_k p_k q_k}\Gamma^{-1}\left(\frac{1}{p_Z^2} E\left[\left(\sum_{i=1}^{n_k} Z_{k,i}V_{k,i}\right)^2\right] - \frac{1-p_z}{p_Z^2}\sum_{i=1}^{n_k} E[Z_{k,i}V_{k,i}^2]\right)\Gamma^{-1} \\ &= \frac{1}{n_k p_k q_k}\Gamma^{-1}\left(\frac{1}{p_Z^2}\sum_{m=1}^{G_k} E\left[\left(\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} Z_{k,i}V_{k,i}\right)^2\right] - \frac{1-p_z}{p_Z^2}\sum_{i=1}^{n_k} E[Z_{k,i}V_{k,i}^2]\right)\Gamma^{-1},\end{aligned}$$

$$\check{\Delta}_k^{\text{CCV}} = \frac{1}{N_k}\Gamma^{-1}\left(\frac{1}{p_Z^2}\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k} \chi_{g_{k,i}=m} Z_{k,i}\check{V}_{k,i}\right)^2 - \frac{1-p_z}{p_Z^2}\sum_{i=1}^{n_k} Z_{k,i}\check{V}_{k,i}^2\right)\Gamma^{-1},$$

$$\tilde{\Delta}_k^{\text{CCV}} = \frac{1}{N_k}\Gamma^{-1}\left(\frac{1}{p_Z^2}\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}Z_{k,i}\tilde{V}_{k,i}\right)^2 - \frac{1-p_z}{p_Z^2}\sum_{i=1}^{n_k}Z_{k,i}\tilde{V}_{k,i}^2\right)\Gamma^{-1}.$$

and

$$\ddot{\Delta}_k^{\text{CCV}} = \frac{1}{N_k}\Gamma^{-1}\left(\frac{1}{p_Z^2}\sum_{m=1}^{G_k}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}Z_{k,i}V_{k,i}\right)^2 - \frac{1-p_z}{p_Z^2}\sum_{i=1}^{n_k}Z_{k,i}V_{k,i}^2\right)\Gamma^{-1}.$$

If assumption 2.7 holds with $\delta > 4$, $\ddot{\Delta}_k^{\text{CCV}} - \Delta_k^{\text{CCV}} \xrightarrow{p} 0$. By the same argument and convergence of $\hat{\theta}_{k,m}^*$, $\hat{\theta}_k$, $\hat{\theta}_k^*$, $\hat{\gamma}_k$ and $\hat{\Gamma}_{k,g}$:when $p_k = 1$ $\hat{V}_k^{\text{CCV}} - \check{\Delta}_k^{\text{CCV}} \xrightarrow{p} 0$, $\check{\Delta}_k^{\text{CCV}} - \tilde{\Delta}_k^{\text{CCV}} \xrightarrow{p} 0$ and $\tilde{\Delta}_k^{\text{CCV}} - \ddot{\Delta}_k^{\text{CCV}} \xrightarrow{p} 0$. Then when $p_k = 1$: $\hat{V}_k^{\text{CCV}} - V^{\text{CCV}} = (\hat{V}_k^{\text{CCV}} - \check{\Delta}_k^{\text{CCV}}) + (\check{\Delta}_k^{\text{CCV}} - \tilde{\Delta}_k^{\text{CCV}}) + (\tilde{\Delta}_k^{\text{CCV}} - \ddot{\Delta}_k^{\text{CCV}}) + (\ddot{\Delta}_k^{\text{CCV}} - \Delta_k^{\text{CCV}}) + (\Delta_k^{\text{CCV}} - \Delta^{\text{CCV}}) \xrightarrow{p} 0$ where the last difference goes to 0 by assumption 2.10.

The last thing to check is to show

$$(1-\hat{p}_k)\sum_{m=1}^{G_k}\frac{N_{k,m}}{N_k}(\hat{\theta}_{k,m} - \hat{\theta}_k)^2$$

converges. Notice that

$$E\left[\frac{N_{k,m}}{p_k q_k n_{k,m}}\right] =$$

and

$$\text{Var}\left[\frac{N_{k,m}}{p_k n_{k,m}}\right] = \frac{n_{k,m}p_k(1-p_k)}{n_{k,m}p_k^2} \to 0.$$

Therefore it suffices to show

$$(1-\hat{p}_k)\sum_{m=1}^{G_k}\frac{n_{k,m}}{n_k}\frac{Q_k}{q_k}(\hat{\theta}_{k,m} - \hat{\theta}_k)^2$$

converges. The result follows from the convergence of $\hat{\theta}_{k,m}$, $\hat{\theta}_k$, $\frac{Q_k}{q_k} \to 1$ and the fact that $N_k \xrightarrow{p} p_k q_k n_k$ so when $q_k = 1$

$$\hat{p}_k = \frac{q_k N_k}{n_k} \xrightarrow{p} p_k.$$

which follows since $\frac{N_k}{n_k p_k q_k} \xrightarrow{p} 1$. □

Lemma 8.2. Let $V_{n,i}$ be a row-wise independent triangular array and $\mu_{n,i} = E[V_{n,i}]$. Suppose that $Q_{k,1}, Q_{k,2}, ...Q_{k,G_k}$ are independent of $V_{n,1}, V_{n,2}, ..., V_{n,G_k}$ and that Assumption 2.2 holds. $V_{n,g} = \sum_{i=1}^{n_k} \chi_{g_{k,i}=g} D_{k,i}$. Assume that

$$\frac{1}{n_k} \sum_{m=1}^{G_k} E[|V_{k,m}|^{2+\delta}]$$

is bounded for $\delta > 0$,

$$\sum_{m=1}^{G_k} \mu_{k,m} = 0,$$
$$\frac{1}{n_k} \sum_{m=1}^{G_k} Var(V_{k,m}) \to \sigma^2$$

and

$$\frac{1}{n_k} \sum_{m=1}^{G_k} \mu_{k,m}^2 \to \kappa^2,$$

where $\sigma^2 + (1-q)\kappa^2 > 0$.

$$\frac{1}{\sqrt{N_k}} \sum_{m=1}^{G_k} (Q_{k,m} V_{k,m}) \xrightarrow{d} \mathcal{N}(0, \sigma^2 + (1-q)\kappa^2),$$

where $N_k = \sum_{m=1}^{G_k} Q_{k,m} n_{k,m}$.

Proof. I showed above that $\frac{N_k}{n_k q_k} \xrightarrow{p} 1$This implies, by the continuous mapping theorem, $\left(\frac{q_k n_k}{N_k}\right)^{1/2} \to 1$. Then, it is sufficient to prove

$$\frac{1}{\sqrt{n_k}} \sum_{m=1}^{G_k} \frac{1}{\sqrt{q_k}} Q_{k,m} V_{k,m} \xrightarrow{d} \mathcal{N}(0, \sigma^2 + (1-q)\kappa^2),$$

Let

$$s_k = \frac{1}{n_k} \sum_{m=1}^{G_k} \Big(\text{Var}(V_{k,m}) + (1-q_k)\mu_{k,m}^2 \Big).$$

Let k be large enough such that s_k is bounded away from 0. Then, for all $i = 1, ..., n_k$

$$E\left[\frac{Q_{k,m}V_{k,m} - q_k\mu_{k,m}}{s_k\sqrt{n_k q_k}}\right] = 0$$

$$\begin{aligned}\text{Var}(Q_k V_{k,g} - q_k\mu_{k,g}) &= q_k E[V_{k,g}^2] - q_k^2\mu_{k,g}^2 \\ &= q_k\left(\text{Var}(V_{k,g}) + (1-q_k)\mu_{k,g}^2\right).\end{aligned}$$

Therefore,

$$\sum_{m=1}^{G_k}\text{Var}\left(\frac{Q_{k,i}V_{k,i} - q_k\mu_{k,i}}{s_k\sqrt{q_k n_k}}\right) = 1$$

The rest of the argument follows from a Liapunov-type argument. For details, see the proof for Lemma A.1 in Abadie et al. (2020). □

Proof of Theorem 4.2. Consider $V_{k,g} = \sum_{i=1}^{n_k} a' \chi_{g_{k,i}=g} U_{k,i}\varepsilon_{k,i}^{\text{causal}}$.

$$\frac{1}{n_k}\sum_{m=1}^{G_k} E\left[|V_{k,m}|^{2+\delta}\right] \leq \frac{||a||^{2+\delta}}{n_k}\sum_{m=1}^{G_k} E\left[\left|\sum_{k=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right|^{2+\delta}\right].$$

By Minkowski's inequality,

$$\begin{aligned}\frac{||a||^{2+\delta}}{n_k}\sum_{m=1}^{G_k} E\left[\left|\sum_{k=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right|^{2+\delta}\right]^{\frac{1}{2+\delta}} &\leq \frac{||a||^{2+\delta}}{n_k}\sum_{m=1}^{G_k}\sum_{i=1}^{n_k} E\left[\chi_{g_{k,i=m}}|U_{k,i}\varepsilon_{k,i}^{\text{causal}}|^{2+\delta}\right]^{\frac{1}{2+\delta}} \\ &\leq \frac{||a||^{2+\delta}}{n_k}\sum_{i=1}^{n_k} E[\chi_{g_{k,i}=m}||U_{k,i}||^{2+\delta}(|Y_{k,i}| + ||U_{k,i}||||\theta_{k,i}|| + ||X_{k,i}||||\gamma_k|| + |\varepsilon_{k,i}^{\text{causal}}|)^{2+\delta}]^{\frac{1}{2+\delta}}.\end{aligned}$$

By Minkowski's equality and assumption 2.7, this term is bounded. In addition,

$$\sum_{m=1}^{G_k} a'\mu_{k,i} = \sum_{i=1}^{n_k} a' E[U_{k,i}\varepsilon_{k,i}^{\text{causal}}] = 0.$$

When $a \neq 0$,

$$\frac{1}{n_k}\sum_{m=1}^{G_k}\text{Var}(V_{k,g}) = a'\left(\frac{1}{n_k}\sum_{m=1}^{G_k}\text{Var}\left(\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)\right)a \to a'\Delta^{\text{CCV}}a$$

$$\frac{1}{n_k}\sum_{m=1}^{G_k}\mu_{k,i}^2 = a'\left(\frac{1}{n_k}\sum_{m=1}^{G_k}E\left[\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right]E\left[\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}\varepsilon_{k,i}^{\text{causal}}U_{k,i}'\right]\right)a \to a'\Delta^{\mu}a$$

Then,

$$a'\left(\frac{1}{\sqrt{N}}\sum_{m=1}^{G_k}Q_{k,m}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{causal}}\right)\xrightarrow{d}N(0,a'\left(\Delta^{\text{CCV}}+(1-q)\Delta^{\mu}\right)a)$$

and by the Cramer-Wold theorem,

$$\left(\frac{1}{\sqrt{N}}\sum_{m=1}^{G_k}Q_{k,m}\sum_{i=1}^{n_k}\chi_{g_{k,i}=m}U_{k,i}\varepsilon_{k,i}^{\text{Causal}}\right)\xrightarrow{d}N(0,\left(\Delta^{\text{CCV}}+(1-q)\Delta^{\mu}\right)).$$

The rest of the argument follows identically from the proof of theorem 2.2. □

Proof. Lemma 4.2 See proof for Lemma 3.1. □

Proof. Theorem 4.3 See proof for Theorem 3.1. □

References

[1] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M. Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

[2] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M. Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

[3] Stephen Donald and Kevin Lang. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233, 2007.

[4] Bruce E. Hansen and Seojeong Lee. Asymptotic theory for clustered samples. *Journal of Econometrics*, 210(2):268–290, 2019.

[5] Jeffrey M. Wooldridge. What is a standard error? (and how should we compute it?). *Journal of Econometrics*, 237(2, Part A):105517, 2023.

[6] James G MacKinnon, Morten Ørregaard Nielsen, and Matthew D Webb. Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2):272–299, 2023.

[7] Charles P Quesenberry Jr and Nicholas P Jewell. Regression analysis based on stratified samples. *Biometrika*, 73(3):605–614, 1986.

[8] Ruonan Xu and Luther Yap. Clustering with potential multidimensionality: Inference and practice, 2024.