GMM UNDER FINITE-POPULATION ASYMPTOTICS: INSTRUMENTAL VARIABLES AND REGRESSION ADJUSTMENT

HARUO KAKEHI, YUKITOSHI MATSUSHITA, AND TAISUKE OTSU

Abstract. This paper extends the finite-population asymptotic approach by Abadie *et al.* (2014, 2020) to the generalized method of moments (GMM) estimator for moment condition models. Motivating examples include instrumental variable regressions and regression adjustments under randomized controlled trials. We study asymptotic properties of the GMM estimator and propose conservative variance estimators. Notably, even though the optimally weighted GMM estimation is infeasible under the finite-population asymptotics, the regression adjustment estimator in a randomized controlled trial is shown to be asymptotically equivalent to the optimally weighted GMM estimator. A simulation study and empirical example illustrate usefulness of our GMM theory.

1. Introduction

Since the seminal works by Abadie *et al.* (2014, 2020), there has been growing interest in inference methods under finite-population setups accounting for design-based uncertainty. The design-based perspective for investigating econometric or statistical methods has been prevalent in randomized experiments (e.g., Neyman, 1923; Rosenbaum, 2002; Freedman, 2008a, 2008b). However, this literature does not consider sampling-based uncertainty deriving from not observing the entire population since it is common to assume that random assignment is the only source of uncertainty in an experimental setting. On the other hand, extensive statistical literature exists on finite-population asymptotics, taking into account of sampling variation (see, Prášková and Sen, 2009, for an overview) although this body of work

Matsushita acknowledges financial support from the JSPS KAKENHI (23K01331).

omits consideration of design-based uncertainty. Abadie *et al.* (2020) developed an alternative inference framework in observational study settings by incorporating both design and sampling-based uncertainty. While their framework is restricted to the case of linear regression, recent literature considers design-based uncertainty in different settings; M-estimators (Xu, 2021), spatial correlation (Xu and Wooldridge, 2022), and network (Sakamoto and Shimizu, 2025). Our paper closely relates to the work by Xu (2021), which extends Abadie *et al.* (2020) to the M-estimation setting.

In this paper, we extend the above existing results on finite-population inference to the situations where estimands of interest are defined by overidentified moment conditions. In conventional empirical economic analyses using the infinite-population asymptotics, moment condition models are ubiquitous, and there is a rich literature on applied and theoretical econometric analyses for these models typically using the generalized method of moments (GMM) (see, e.g., Hall, 2005, for an overview). Therefore, it is of substantial interest to develop the GMM theory for finite-population inference problems.

We study the asymptotic properties of the GMM estimator for moment condition models under the finite-population asymptotics. We derive the consistency and asymptotic normality of the GMM estimator. In particular, we find that its finite-population asymptotic variance takes a different form from the conventional infinite-population asymptotic variance, which is an analogous finding in just-identified models (e.g., Abadie *et al.*, 2020; Xu, 2021). Since our asymptotic variance for the GMM estimator involves a component that is not consistently estimable, we propose two asymptotically conservative variance estimators: one is the conventional variance estimator, and the other is an adapted version of Abadie *et al.* (2020) variance estimator to the GMM context. Furthermore, we discuss the choice of the GMM weight matrix under the finite-population asymptotic framework and suggest a feasible data-dependent weight where the associated conservative variance estimator shows a desirable property.

Our primary motivating examples are instrumental variable (IV) regression models and regression or covariate adjustments under randomized controlled trials. The recent design-based literature has focused on experiments with noncompliance (e.g., Kang, Peck and Keele, 2018; Jiang et al., 2024; Ren, 2024; Rambachan and Roth, 2025). In this setting, our contribution is to extend the design-based framework to allow for non-binary and multiple instruments, and incorporate both sampling and design uncertainty in a unified GMM setting. Furthermore, the finite-population asymptotic analysis of the GMM estimator and the development of feasible inference methods are open issues in the literature. In contrast to M-estimation problems as studied by Xu (2021), the GMM requires a weight matrix, and currently there is no guidance on its choice under the finite-population asymptotics.

Another important implication of the proposed GMM theory is that it provides a framework to evaluate the efficiency gain by the regression adjustment to estimate the average treatment effect under a randomized controlled trial. Regression adjustments for baseline covariates are widely applied in practice and also highly encouraged by regulatory agencies to achieve more credible and efficient inference on the causal effects. Under the design-based setup, Lin (2013) suggested to run a regression of the observed outcome on the treatment variable, covariates, and their interactions to guarantee efficiency gain over the simple difference in means. Negi and Wooldridge (2021) established an analogous efficiency gain result for the regression adjustment under the sampling-based (or super-population) setup. As emphasized in the literature, such efficiency gain for the regression adjustment is guaranteed regardless of correct specification of the linear model, called the no-harm property. Indeed the regression adjustment estimator for the average treatment effect can be interpreted as a GMM estimator for certain (plug-in) moment conditions, and an application of our GMM theory reveals a new no-harm property of the regression adjustment under the finite-population asymptotics. Even though the optimally weighted GMM estimation is generally infeasible under the finite-population asymptotics, the regression adjustment estimator for the average treatment effect in a randomized controlled trial turns out to be asymptotically equivalent to the optimally weighted GMM estimator.

The rest of the paper is organized as follows. Section 2 presents the basic setup and GMM estimator. In Section 2.1, we discuss the regression adjustment estimator as an example. Section 3 presents our main theoretical results, asymptotic properties of the GMM estimator under the finite-population asymptotics. Furthermore, Section 3.1 extends our finite-population asymptotics to the situation where the moment conditions are misspecified, Section 3.2 discusses the choice of the GMM weight matrix and the power property of our recommended estimator, and Section 3.3 considers overidentifying restriction testing. In Section 4, we apply the proposed GMM theory to investigate the properties of the regression adjustment estimator in a randomized controlled trial. Section 5 illustrates the proposed GMM theory by a simulation study for the regression adjustment estimator (Section 5.1) and an empirical example to study the effect of institutional quality on income across former European colonies based on Acemoglu, Johnson, and Robinson (2001) (Section 5.2).

2. Setup and estimator

We first introduce our basic setup based on Abadie *et al.* (2020) and Xu (2021). For each unit $i = 1, \dots, M$ with population size M, consider the population $\{X_i, z_i, Y_i\}_{i=1}^M$, where X_i is a vector of assignment variables, z_i is a vector of i's attributions, and Y_i is a vector of outcome variables. We are interested in the effect of X_i rather than z_i. For example, X_i refers to a treatment status of an experiment or indicator for a policy intervention, while z_i includes age, gender, and socioeconomic status of the unit i. Throughout the paper, we assume that z_i is non-random, and the outcomes are written as $Y_i = y_i(X_i)$ for a potential outcome function $y_i(\cdot)$. Although these variables depend on the population size M to conduct asymptotic analysis for $M \to \infty$, we suppress such dependence to simplify the presentation. From the finite population $\{X_i, z_i, Y_i\}_{i=1}^M$, we observe a random sample by Bernoulli sampling. Let R_i be a Bernoulli random variable, which equals one if i is sampled, and zero otherwise. Let $N = \sum_{i=1}^M R_i$ be the sample size, which is random.[1]

This paper is concerned with the situation where the estimand of interest is defined as a solution of just- or over-identified moment conditions. Let $g_i(X_i, \theta) := g(X_i, z_i, y_i(X_i), \theta)$ with a k-dimensional vector of moment functions g and a p-dimensional vector of parameters θ. The estimand of interest θ_M^* is defined as a unique solution of

$$\frac{1}{M} \sum_{i=1}^M \mathbb{E}[g_i(X_i, \theta_M^*)] = 0, \tag{1}$$

[1]We note that in the finite-population asymptotic framework adopted in this paper (as well as Abadie *et al.,* 2020, and Xu, 2021), the sample size N and also the number of treated units $\sum_{i=1}^M R_i X_i$ are treated as random. On the other hand, in much of the statistics literature on design-based analysis (e.g., Neyman, 1923, Lin, 2013, and Li and Ding, 2017), the asymptotic analysis is typically conducted conditionally on the number of treated units.

for $k \geq p$. Note that the expectation $\mathbb{E}[\cdot]$ is taken with respect to the assignment variables X_i. It should be noted that the existing papers such as Abadie *et al.* (2020) and Xu (2021) only cover the case of $k = p$ (just-identified).[2]

Therefore, we mainly focus on the overidentified case and consider the GMM estimator

$$\hat{\theta}_N(W_N) = \arg\min_{\theta\in\Theta} \left\{ \frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \theta) \right\}' W_N \left\{ \frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \theta) \right\}, \tag{2}$$

where Θ is parameter space of θ_M^* and W_N is a $k \times k$ weight matrix. The choice of W_N will be discussed in Section 3.2 below. Before presenting its theoretical properties, we mention some motivating examples below and in the next subsection.

Example 1. [IV regression] An important example of the moment condition model (1) is IV regression. Kang, Peck and Keele (2018) studied estimation and inference methods for (just-identified) IV regressions under the design-based framework. To fit into our notation, let $X_i \in \{0, 1\}$ be a binary IV with $\mathbb{P}\{X_i = 1\} = p$, $y_i^{\mathrm{t}}(x)$ be the potential treatment for the instrument value $x \in \{0, 1\}$, and $y_i^{\mathrm{o}}(x, t)$ be the potential outcome for the instrument value $x \in \{0, 1\}$ and treatment value t. Then the moment function for the IV regression estimator considered Kang, Peck and Keele (2018) can be written as

$$g_i(X_i, \theta_M^*) = \left(\frac{X_i}{p} - \frac{1 - X_i}{1 - p} \right) \{ y_i^{\mathrm{o}}(X_i, y_i^{\mathrm{t}}(X_i)) - \theta_M^* y_i^{\mathrm{t}}(X_i) \},$$

Kang, Peck and Keele (2018) proposed a randomization-based inference method by using the IV estimator. This paper provides an alternative inference method for the parameters θ_M^*

[2]For the just-identified case, the method of moments estimator is written as $\arg\min_\theta \left\| \frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \theta) \right\|^2$ and this cannot be covered by the setup in Xu (2021) who studied the M-estimator $\arg\min_\theta \frac{1}{N} \sum_{i=1}^{M} R_i Q(X_i, z_i, y_i(X_i), \theta)$ for some criterion function $Q(\cdot)$.

by extending the finite-population asymptotics to this context. Furthermore, our framework accommodates the case of continuous and multiple instruments which implies overidentified moment conditions.

Example 2. [Auxiliary information] Another motivating example is the situation where the researcher has auxiliary information on the population moments as investigated in Imbens and Lancaster (1994). Under the finite-population asymptotic framework, the proportion of the sample size to the population size is non-negligible and it may be plausible that researchers have access to some population moments. Examples include large-scale experiments where sample representativeness is important (Muralidharan and Niehaus, 2017; Duflo and Banerjee, 2017) and the Integrated Public Use Microdata Series (IPUMS) data, which is the 10% sample of the U.S. Census. In these cases, researchers may incorporate moments based on the entire states or counties to improve point estimation and associated inference.

2.1. **Regression adjustment.** An important application of our GMM theory under the finite-population asymptotics is regression adjustment to estimate the average treatment effect under a randomized controlled trial. In this subsection, let $X_i \in \{0, 1\}$ be an indicator for assignment ($X_i = 1$ corresponds to the treatment, and $X_i = 0$ corresponds to the control), and $y_i(x)$ be a potential outcome for the treatment $x \in \{0, 1\}$. For each unit i with $R_i = 1$, we observe the randomly assigned treatment X_i, the observable outcome $Y_i = y_i(X_i)$, and a vector of pre-treatment covariates z_i (without a constant term) treated as non-random.

In this setup, we wish to conduct inference on the average treatment effect $\theta_M^* = \frac{1}{M}\sum_{i=1}^M \{y_i(1) - y_i(0)\}$ using the observables $\{X_i, Y_i, z_i\}_{i:R_i=1}$. Let

$$\begin{aligned}\bar{Y}_1 &= \frac{1}{N_1}\sum_{i=1}^M R_i X_i Y_i, \qquad \bar{Y}_0 = \frac{1}{N_0}\sum_{i=1}^M R_i(1-X_i)Y_i, \\ \bar{z}_1 &= \frac{1}{N_1}\sum_{i=1}^M R_i X_i z_i, \qquad \bar{z}_0 = \frac{1}{N_0}\sum_{i=1}^M R_i(1-X_i)z_i,\end{aligned}$$

where $N_1 = \sum_{i=1}^M R_i X_i$ and $N_0 = \sum_{i=1}^M R_i(1-X_i)$ are the numbers of sampled treatment and control units, respectively. Popular estimators of θ_M^* are:

$$\begin{aligned}\hat{\theta}_{\mathrm{DIM}} &= \bar{Y}_1 - \bar{Y}_0, \qquad \text{(difference-in-means)} \\ \hat{\theta}_{\mathrm{RA}} &= \bar{Y}_1 - \bar{Y}_0 - (\hat{\beta}_1', -\hat{\beta}_0')\begin{pmatrix} \bar{z}_1 - \bar{z} \\ \bar{z}_0 - \bar{z} \end{pmatrix}, \qquad \text{(regression adjustment)} \qquad (3)\end{aligned}$$

where

$$\begin{aligned}\hat{\beta}_1 &= \left(\sum_{i=1}^M R_i X_i (z_i - \bar{z}_1)(z_i - \bar{z}_1)'\right)^{-1} \sum_{i=1}^M R_i X_i (z_i - \bar{z}_1) Y_i, \\ \hat{\beta}_0 &= \left(\sum_{i=1}^M R_i (1-X_i)(z_i - \bar{z}_0)(z_i - \bar{z}_0)'\right)^{-1} \sum_{i=1}^M R_i (1-X_i)(z_i - \bar{z}_0) Y_i.\end{aligned}$$

Under the design-based setup, Lin (2013) showed that the regression adjustment estimator $\hat{\theta}_{\mathrm{RA}}$ is asymptotically more efficient than the difference-in-means estimator $\hat{\theta}_{\mathrm{DIM}}$ without assuming correct specification of the linear regression model. Under the sampling-based (or super-population) setup, Negi and Wooldridge (2021) established analogous efficiency guarantee for the regression adjustment. It is of substantial interest whether such efficiency improvement for the regression adjustment continues to hold under the present setup of the finite-population asymptotics.

The GMM theory in this paper provides a unified framework to study asymptotic properties of $\hat{\theta}_{\mathrm{DIM}}$ and $\hat{\theta}_{\mathrm{RA}}$ under the finite-population asymptotics. In particular, letting $\hat{\pi} = N_1/N$, we consider the moment function:

$$g_i(X_i, \theta_M^*) = \begin{pmatrix} \frac{X_i Y_i}{\hat{\pi}} - \frac{(1-X_i)Y_i}{1-\hat{\pi}} - \theta_M^* \\ \left(\frac{X_i}{\hat{\pi}} - \frac{1-X_i}{1-\hat{\pi}}\right) z_i \end{pmatrix}.$$

Then the difference-in-means estimator $\hat{\theta}_{\mathrm{DIM}}$ and an asymptotically equivalent counterpart of the regression adjustment estimator $\hat{\theta}_{\mathrm{RA}}$ can be interpreted as the GMM estimators with suitably chosen weight matrices. In Section 4 below, we apply the GMM theory under the finite-population asymptotics to show that $\hat{\theta}_{\mathrm{RA}}$ is asymptotically more efficient than $\hat{\theta}_{\mathrm{DIM}}$ and is indeed optimal in the class of the GMM estimators.

3. GMM under finite-population asymptotics

We now study large sample properties of the GMM estimator $\hat{\theta}_N(W_N)$ in our asymptotic framework. We impose the following assumptions.

Assumption.

(1) $\{X_i\}_{i=1}^M$ *is independent but not necessarily identically distributed.* $\{R_i\}_{i=1}^M$ *is an independent and identically distributed sequence of Bernoulli random variables with* $\rho_M = \mathbb{P}\{R_i = 1\}$ *satisfying* $\rho_M \to \rho \in (0,1]$ *as* $M \to \infty$. *Furthermore,* $\{X_i\}_{i=1}^M$ *and* $\{R_i\}_{i=1}^M$ *are independent.*

(2) Θ *is compact.* W_N *converges in probability to a positive semi-definite matrix* W. $\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[g_i(X_i,\theta)] = 0$ *is uniquely satisfied at* $\theta = \theta^* := \lim_{M\to\infty}\theta_M^*$. $g_i(x,\theta)$ *is continuous at each* $\theta\in\Theta$ *for almost every* x, *and* $\sup_{i,M}\mathbb{E}[\sup_{\theta\in\Theta}||g_i(X_i,\theta)||^4] < \infty$. *There exist functions* $h_1(\cdot)$ *and* $b_{1i}(\cdot)$ *such that* $\lim_{u\to 0}h_1(u) = 0$, $\sup_{i,M}\mathbb{E}[b_{1i}(X_i)] < \infty$, *and* $||g_i(X_i,\theta) - g_i(X_i,\theta_1)|| \le b_{1i}(X_i)h_1(||\theta-\theta_1||)$ *for each* $\theta,\theta_1\in\Theta$.

(3) $\theta^*\in\mathrm{int}(\Theta)$. $g_i(x,\theta)$ *is continuously differentiable on* $\mathrm{int}(\Theta)$ *for almost every* x, *and* $G := \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\frac{\partial g_i(X_i,\theta_M^*)}{\partial\theta'}\right]$ *has full column rank. There exist functions* $h_2(\cdot)$ *and* $b_{2i}(\cdot)$ *and a neighborhood* $\mathcal{N}$ *around* θ^* *such that* $\sup_{i,M}\mathbb{E}[\sup_{\theta\in\mathcal{N}}||\partial g_i(X_i,\theta)/\partial\theta'||^2] < \infty$, $\lim_{u\to 0}h_2(u) = 0$, $\sup_{i,M}\mathbb{E}[b_{2i}(X_i)] < \infty$, *and* $||\partial g_i(X_i,\theta)/\partial\theta' - \partial g_i(X_i,\theta_1)/\partial\theta'|| \le b_{2i}(X_i)h_2(||\theta-\theta_1||)$ *for each* $\theta,\theta_1\in\mathcal{N}$.

Assumption (1) is on the sampling framework, which is also employed by Abadie *et al.* (2020) and Xu (2021). This implies the sample size $N = \sum_{i=1}^{M} R_i$ is random, and its expectation $\mathbb{E}[N] = M\rho_M$ diverges at the same rate as $M\to\infty$. Assumptions (2) and (3) collect regularity conditions on the weight matrix W_N and the moment function g_i. These are natural adaptations of the conventional GMM theory to our finite-population setup. Assumption (2) is used to derive the consistency of the GMM estimator, and Assumption (3) contains additional conditions to establish asymptotic normality.

Under the above assumptions, the asymptotic properties of the GMM estimator $\hat\theta_N(W_N)$ are obtained as follows.

Theorem 1.

(1) *Under Assumptions (1)-(2), it holds* $\hat\theta_N(W_N) - \theta_M^* \xrightarrow{p} 0$.

(2) *Under Assumptions (1)-(3), it holds*

$$\sqrt{N}(\hat{\theta}_N(W_N) - \theta^*_M) \xrightarrow{d} N(0, V_{\mathrm{GMM}}(W)), \tag{4}$$

where

$$\begin{aligned} V_{\mathrm{GMM}}(W) &= (G'WG)^{-1}G'W(\Omega - \rho\Delta)WG(G'WG)^{-1}, \\ \Omega &= \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta^*_M)g_i(X_i, \theta^*_M)'], \\ \Delta &= \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta^*_M)]\mathbb{E}[g_i(X_i, \theta^*_M)]'. \end{aligned}$$

Theorem 1 (1) says that the GMM estimator $\hat{\theta}_N(W_N)$ is consistent for the population parameter θ^*_M, and Theorem 1 (2) derives its asymptotic distribution. Compared to the conventional infinite-population asymptotics, the main difference is the presence of the additional term "$\rho\Delta$" in the asymptotic variance. Letting $V_{\mathrm{C}}(W) = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$ be the asymptotic variance of the GMM estimator under the conventional infinite-population asymptotics, and $V_{\mathrm{A}}(W) = \rho(G'WG)^{-1}G'W\Delta WG(G'WG)^{-1}$ be an additional component, the finite-population asymptotic variance can be written as $V_{\mathrm{GMM}}(W) = V_{\mathrm{C}}(W) - V_{\mathrm{A}}(W)$. Since Δ is positive semi-definite, $V_{\mathrm{A}}(W)$ is also positive semi-definite and $V_{\mathrm{GMM}}(W)$ is always smaller than the conventional variance $V_{\mathrm{C}}(W)$ in the matrix sense (denoted by $V_{\mathrm{GMM}}(W) \leq_{\mathrm{pd}} V_{\mathrm{C}}(W)$). Although $V_{\mathrm{C}}(W)$ can be consistently estimated (as shown below), the component Δ and thus the variance $V_{\mathrm{GMM}}(W)$ cannot be consistently estimable in general.

As in Abadie *et al.* (2020) and Xu (2021), we propose conservative estimators for $V_{\text{GMM}}(W)$. The first variance estimator is a consistent estimator of the conventional variance $V_{\text{C}}(W)$, that is

$$\hat{V}_{\text{C}}(W_N) = (\hat{G}'W_N\hat{G})^{-1}\hat{G}'W_N\hat{\Omega}W_N\hat{G}(\hat{G}'W_N\hat{G})^{-1},$$

where $\hat{G} = \frac{1}{N}\sum_{i=1}^{M} R_i\partial g_i(X_i, \hat{\theta}_N(W_N))/\partial\theta'$ and $\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{M} R_i g_i(X_i, \hat{\theta}_N(W_N))g_i(X_i, \hat{\theta}_N(W_N))'$. The second variance estimator is constructed by estimating a lower bound for Δ, that is

$$\hat{\Delta}_Z = \frac{1}{N}\sum_{i=1}^{M} R_i\hat{P}'z_iz_i'\hat{P},$$

where $\hat{P} = \left(\sum_{i=1}^{M} R_iz_iz_i'\right)^{-1}\left(\sum_{i=1}^{M} R_iz_ig_i(X_i, \hat{\theta}_N(W_N))\right)$. Then the second variance estimator is written as

$$\hat{V}_Z(W_N) = (\hat{G}'W_N\hat{G})^{-1}\hat{G}'W_N(\hat{\Omega} - \rho\hat{\Delta}_Z)W_N\hat{G}(\hat{G}'W_N\hat{G})^{-1}.$$

The asymptotic properties of these variance estimators are presented as follows.

Theorem 2.

(1): *Under Assumptions (1)-(3),* $\hat{V}_{\text{C}}(W_N) \xrightarrow{p} V_{\text{C}}(W)$.

(2): *In addition to Assumptions (1)-(3), assume that* $\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[g_i(X_i, \theta_M^*)]z_i'$ *exists,* $\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M} z_iz_i'$ *exists and non-singular, and* $\sup_{i,M}||z_i|| < \infty$. *Then* $\hat{\Delta}_Z$ *converges in probability to a positive semi-definite matrix* Δ_Z *such that* $\Delta_Z \leq_{\text{pd}} \Delta$.

(3): $V_{\text{GMM}}(W) \leq_{\text{pd}} V_Z(W) \leq_{\text{pd}} V_{\text{C}}(W)$, *where* $V_Z(W) = (G'WG)^{-1}G'W(\Omega - \rho\Delta_Z)WG(G'WG)^{-1}$. *Furthermore,* $V_Z((\Omega - \rho\Delta_Z)^{-1}) \leq_{\text{pd}} V_Z(\Omega^{-1})$.

The proof of the second statement of Theorem 2 (3) is provided in Appendix A.2. Since proofs of the other parts are similar to the ones in Xu (2021, Theorems 2.2 and 3.1), they are omitted. Theorem 2 (1) says that the conventional variance estimator $\hat{V}_{\mathrm{C}}(W_N)$ is still consistent for the variance component $V_{\mathrm{C}}(W)$ under the finite-population asymptotics. Theorem 2 (2) guarantees conservativeness of the second variance estimator $\hat{V}_Z(W_N)$ for $V_{\mathrm{GMM}}(W)$. Theorem 2 (3) clarifies the relationships of the limits of the variance estimators. Although we cannot consistently estimate the asymptotic variance $V_{\mathrm{GMM}}(W)$ of the GMM estimator, we can provide asymptotically conservative estimators $\hat{V}_{\mathrm{C}}(W_N)$ and $\hat{V}_Z(W_N)$. We recommend using the second estimator $\hat{V}_Z(W_N)$ under the finite-population asymptotic framework because it is less conservative than the first one $\hat{V}_{\mathrm{C}}(W_N)$.

3.1. **Misspecified case.** When the moment conditions are overidentified ($k > p$), one may be concerned with the situation where the moment conditions are misspecified, i.e., $\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta)] \neq 0$ for any $\theta \in \Theta$. Indeed, our finite-population asymptotic analysis can be extended to allow for such misspecified models. To this end, we modify Assumption (2) as follows.

Assumption.

(2)': *Assumption (2) except for the third sentence on identification of θ^* holds true. Furthermore, assume that*

$$\theta_M^{\#}(W) := \arg\min_{\theta\in\Theta} \left(\frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta_M^*)]\right)' W \left(\frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta_M^*)]\right),$$

and $\theta^{\#}(W) := \lim_{M\to\infty} \theta_M^{\#}(W)$ are unique.

(3)': $\theta^{\#}(W) \in \text{int}(\Theta)$. *$g_i(x, \theta)$ is twice continuously differentiable on* $\text{int}(\Theta)$ *for almost every x. Define*

$$\begin{aligned}
\mu^{\#} &= \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta_M^{\#})], \qquad G^{\#} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}\left[\frac{\partial g_i(X_i, \theta_M^{\#})}{\partial \theta'}\right], \\
G^{(2)\#} &= \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}\left[\frac{\partial}{\partial \theta'} \text{vec}\left\{\frac{\partial g_i(X_i, \theta_M^{\#})}{\partial \theta'}\right\}\right], \\
H &= G^{\#\prime} W G^{\#}, \qquad M = (\mu^{\#\prime} W \otimes I) G^{(2)\#}.
\end{aligned}$$

$G^{\#}$ has full column rank. H and $I - H^{-1}M$ are nonsingular. There exist functions $h_2(\cdot)$, $h_3(\cdot)$, $b_{2i}(\cdot)$, and $b_{3i}(\cdot)$, and a neighborhood $\mathcal{N}$ around $\theta^{\#}$ such that $\sup_{i,M} \mathbb{E}[\sup_{\theta\in\mathcal{N}} ||\partial g_i(X_i, \theta)/\partial \theta'||^2] < \infty$, $\sup_{i,M} \mathbb{E}[\sup_{\theta\in\mathcal{N}} ||\partial \text{vec}\{\partial g_i(X_i, \theta)/\partial \theta'\}/\partial \theta'||^2] < \infty$, $\lim_{u\to 0} h_2(u) = 0$, $\lim_{u\to 0} h_3(u) = 0$, $\sup_{i,M} \mathbb{E}[b_{2i}(X_i)] < \infty$, $\sup_{i,M} \mathbb{E}[b_{3i}(X_i)] < \infty$, $||\partial g_i(X_i, \theta)/\partial \theta' - \partial g_i(X_i, \theta_1)/\partial \theta'|| \leq b_{2i}(X_i) h_2(||\theta - \theta_1||)$, *and* $||\partial \text{vec}\{\partial g_i(X_i, \theta)/\partial \theta'\}/\partial \theta' - \partial \text{vec}\{\partial g_i(X_i, \theta_1)/\partial \theta'\}/\partial \theta'|| \leq b_{3i}(X_i) h_3(||\theta - \theta_1||)$ *for each $\theta, \theta_1 \in \mathcal{N}$. Furthermore, assume*

$$\begin{pmatrix} \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_i \{g_i(X_i, \theta_M^{\#}) - \mathbb{E}[g_i(X_i, \theta_M^{\#})]\} \\ \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_i \left\{\frac{\partial g_i(X_i, \theta_M^{\#})}{\partial \theta'} - \mathbb{E}\left[\frac{\partial g_i(X_i, \theta_M^{\#})}{\partial \theta'}\right]\right\}' W \mu^{\#} \\ \sqrt{N}\{W_N - W\} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{Z}_g \\ \mathcal{Z}_G \\ \mathcal{Z}_W \end{pmatrix} \sim N(0, \mathcal{V}).$$

Assumption (2)' introduces the pseudo-true values in our context and imposes their uniqueness. It should be noted that in contrast to the correctly specified case, the pseudo true values $\theta_M^{\#}(W)$ and $\theta^{\#}(W)$ depend on the weight matrix W, and their interpretations need to be clarified for each application (see, Hall and Inoue, 2003, for analogous results under the conventional sampling-based framework). Assumption (3)' lists additional assumptions to derive the asymptotic distribution of the GMM estimator. These assumptions are adapted from Hall and Inoue (2003, Theorem 2).

Based on these assumptions, the asymptotic properties of the GMM estimator $\hat{\theta}_N(W_N)$ for the misspecified case can be presented as follows.

Corollary 1.

(1) *Under Assumptions (1) and (2)', it holds* $\hat{\theta}_N(W_N) - \theta^{\#}_M(W) \xrightarrow{p} 0$.

(2) *Under Assumptions (1), (2)', and (3)', it holds*

$$\sqrt{N}(\hat{\theta}_N(W_N)-\theta^{\#}_M(W)) \xrightarrow{d} (I-H^{-1}M)^{-1}H^{-1}G^{\#\prime}W\mathcal{Z}_g+(I-H^{-1}M)^{-1}H^{-1}\mathcal{Z}_G+(I-H^{-1}M)^{-1}H^{-1}G^{\#\prime}\mathcal{Z}_W.$$

The asymptotic distribution is now centered around the pseudo true value $\theta^{\#}_M(W)$, and the asymptotic distribution is analogous to the one in Hall and Inoue (2003, Theorem 2). However, the limiting variances $\mathrm{Var}(\mathcal{Z}_g)$ and $\mathrm{Var}(\mathcal{Z}_G)$ typically depend on the sampling ratio ρ. For example, an application of Abadie *et al.* (2020, Lemma A.1) implies $\mathcal{Z}_g \sim N(0, \Omega - \rho\Delta)$. The asymptotic variance can be estimated by taking the sample counterparts and estimating Δ by $\hat{\Delta}_Z$ as in the last subsection.

3.2. **Choice of GMM weight.** The asymptotic analysis in the previous subsection focuses on the case where the weight matrix W_N for the GMM estimation is given. Given the different form of the asymptotic variance $V_{\mathrm{GMM}}(W)$ from the one under the conventional infinite-population asymptotics, it is interesting to investigate the choice of the weight matrix under the current setup.

First of all, the variance $V_{\mathrm{GMM}}(W)$ is minimized by $W_{\mathrm{opt}} = (\Omega - \rho\Delta)^{-1}$ in the matrix sense. However, due to the component Δ, a consistent estimator of W_{opt} is not available in general.

Motivated by the discussion in the previous subsection, we can consider two feasible weights, $\hat{\Omega}^{-1}$ and $(\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1}$. Theorems 1 and 2 imply

$$\sqrt{N}(\hat{\theta}_N(\hat{\Omega}^{-1}) - \theta^*_M) \xrightarrow{d} N(0, V_{\text{GMM}}(\Omega^{-1})),$$
$$\sqrt{N}(\hat{\theta}_N((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1}) - \theta^*_M) \xrightarrow{d} N(0, V_{\text{GMM}}((\Omega - \rho\Delta_Z)^{-1})).$$

We note that these asymptotic variances are not directly comparable.[3] However, Theorem 2 (3) implies

$$V_{\text{GMM}}((\Omega - \rho\Delta_Z)^{-1}) \leq_{\text{pd}} V_Z((\Omega - \rho\Delta_Z)^{-1}) \leq_{\text{pd}} V_Z(\Omega^{-1}) \leq_{\text{pd}} V_{\text{C}}(\Omega^{-1}). \tag{5}$$

Based on these relationships, to conduct inference on the parameters θ^*_M, we recommend using the point estimator $\hat{\theta}_N((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$ combined with the asymptotic variance estimator $\hat{V}_Z((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$.

Remark 1. [Power property] To illustrate the power property of inference by using $\hat{\theta}_N((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$ and $\hat{V}_Z((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$, let us consider the case where θ^*_M is scalar, and compare the t-ratios:

$$T_\rho = \frac{\hat{\theta}_N((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1}) - c}{\sqrt{\hat{V}_Z((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})/N}}, \qquad T_0 = \frac{\hat{\theta}_N(\hat{\Omega}^{-1}) - c}{\sqrt{\hat{V}_Z(\hat{\Omega}^{-1})/N}},$$

for testing the null hypothesis $H_0 : \theta^*_M = c$ against the one-sided alternative $H_1 : \theta^*_M > c$. Under the local alternative hypothesis $H_{1n} : \theta^*_M = c + N^{-1/2}h$ for some $h > 0$, an adaptation

[3] To illustrate, consider the case of $k = 2$, $p = 1$, and $\rho = 0.8$ with

$$\Omega = \begin{bmatrix} 5 & 0.5 \\ 0.5 & 5 \end{bmatrix}, \qquad \Delta = \begin{bmatrix} 4 & 0.1 \\ 0.1 & 5 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

which yields $V_{\text{GMM}}(\Omega^{-1}) \approx 0.91$. On the other hand, $V_{\text{GMM}}((\Omega - \rho\Delta_Z)^{-1})$ varies with Δ_Z. For example, let $\Delta_Z = \begin{bmatrix} c & 0.1 \\ 0.1 & 2 \end{bmatrix}$ for some $c > 0$ satisfying $0 \leq \Delta_Z \leq_{\text{pd}} \Delta$. When $c = 1$, it holds $V_{\text{GMM}}((\Omega - \rho\Delta_Z)^{-1})) \approx 0.87 < V_{\text{GMM}}(\Omega^{-1})$. However, when $c = 3$, it holds $V_{\text{GMM}}((\Omega - \rho\Delta_Z)^{-1})) \approx 0.98 > V_{\text{GMM}}(\Omega^{-1})$.

of the proofs of Theorems 1 and 2 implies

$$T_\rho \xrightarrow{d} N(\mu_\rho, V_\rho), \qquad T_0 \xrightarrow{d} N(\mu_0, V_0),$$

where $\mu_\rho = \frac{h}{\sqrt{V_Z((\Omega-\rho\Delta_Z)^{-1})}}$, $\mu_0 = \frac{h}{\sqrt{V_Z(\Omega^{-1})}}$, $V_\rho = \frac{V_{\mathrm{GMM}}((\Omega-\rho\Delta_Z)^{-1})}{V_Z((\Omega-\rho\Delta_Z)^{-1})}$, and $V_0 = \frac{V_{\mathrm{GMM}}(\Omega^{-1})}{V_Z(\Omega^{-1})}$. Thus, if we employ the same critical value ξ, the limiting local power functions of the t-tests by these statistics are written as

$$\mathbb{P}\{T_\rho \geq \xi\} \to 1 - \Phi\left(\frac{\xi - \mu_\rho}{\sqrt{V_\rho}}\right), \qquad \mathbb{P}\{T_0 \geq \xi\} \to 1 - \Phi\left(\frac{\xi - \mu_0}{\sqrt{V_0}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the $N(0,1)$ distribution. Although the non-centrality parameters satisfy $\mu_\rho \geq \mu_0$, we cannot rank the variance terms V_ρ and V_0. If $V_\rho \leq V_0$, the local power of the test $\mathbb{I}\{T_\rho \geq \xi\}$ is always no worse than the alternative test $\mathbb{I}\{T_0 \geq \xi\}$. On the other hand, if $V_\rho > V_0$, the local power of $\mathbb{I}\{T_\rho \geq \xi\}$ is no worse than $\mathbb{I}\{T_0 \geq \xi\}$ as far as $\frac{\xi-\mu_\rho}{\sqrt{V_\rho}} \leq \frac{\xi-\mu_0}{\sqrt{V_0}}$. $\square$

3.3. **Overidentification test.** Similar to parameter hypothesis testing, the asymptotic behavior of the overidentification test is also different from the conventional super-population setting. In this subsection, we consider the following overidentifying restriction test statistic

$$J_N(W_N) = N\left(\frac{1}{N}\sum_{i=1}^{M} R_i g_i(X_i, \hat{\theta}_N(W_N))\right)' W_N \left(\frac{1}{N}\sum_{i=1}^{M} R_i g_i(X_i, \hat{\theta}_N(W_N))\right).$$

The asymptotic property of $J_N(W_N)$ is obtained as follows.

Corollary 2. *Suppose Assumptions (1)-(3) hold true.*

(1) *If $W_N \xrightarrow{p} W = (\Omega - \rho\Delta)^{-1}$ as $M \to \infty$, then $J_N(W_N) \xrightarrow{d} \chi^2_{k-p}$.*

(2) *If* $W_N \xrightarrow{p} W \neq (\Omega - \rho\Delta)^{-1}$ *as* $M \to \infty$, *then* $J_N(W_N) \xrightarrow{d} \sum_{j=1}^{k-p} \lambda_j \chi^2_{1,j}$, *where* $\{\chi^2_{1,j}\}_{j=1}^{k-p}$ *are mutually independent* χ^2_1 *random variables and* $\{\lambda_j\}_{j=1}^{k-p}$ *are the non-zero eigenvalues of* $(\Omega - \rho\Delta)\{W - WG(G'WG)^{-1}G'W\}$.

Corollary 2 shows that the limiting distribution of the statistic $J_N(W_N)$ is characterized by a weighted sum of the chi-squared random variables in general. Due to the component Δ, we cannot consistently estimate the eigenvalues $\{\lambda_j\}_{j=1}^{k-p}$. A practical approach is to substitute Δ_Z and work with the eigenvalues of $(\Omega - \rho\Delta_Z)\{W - WG(G'WG)^{-1}G'W\}$, denoted by $\{\lambda_{Z,j}\}_{j=1}^{k-p}$. By Theorem 2 and the Courant-Fischer-Weyl min-max principle, we have $\lambda_{Z,j} \geq \lambda_j$ for all $j = 1, \ldots, k-p$. Thus, we can employ the estimated quantiles of $\sum_{j=1}^{k-p} \lambda_{Z,j}\chi^2_{1,j}$ as conservative critical values for testing the overidentifying restrictions (i.e., eq. (1) holds true at some $\theta^*_M \in \Theta$).

4. Regression adjustment

In this section, we apply the GMM theory presented in the last section to compare the difference-in-means estimator $\hat{\theta}_{\mathrm{DIM}}$ and regression adjustment estimator $\hat{\theta}_{\mathrm{RA}}$ for the average treatment effect under random assignment of the binary treatment. Recall the setup and notation in Section 2.1, where the moment function is

$$g_i(X_i, \theta^*_M) = \begin{pmatrix} \frac{X_i Y_i}{\hat{\pi}} - \frac{(1-X_i)Y_i}{1-\hat{\pi}} - \theta^*_M \\ \left(\frac{X_i}{\hat{\pi}} - \frac{1-X_i}{1-\hat{\pi}}\right) z_i \end{pmatrix}, \tag{6}$$

with $\hat{\pi} = N_1/N$. Based on this moment function, we consider the GMM estimator for θ^*_M:

$$\hat{\theta}(W) = \arg\min_{\theta} \left\{ \frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \theta) \right\}' W \left\{ \frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \theta) \right\},$$

where $W = \begin{pmatrix} W_1 & W_{12}' \\ W_{12} & W_2 \end{pmatrix}$ with scalar W_1. Let $\bar{z}_M = \frac{1}{M}\sum_{i=1}^M z_i$, $\theta^*_{Mk} = \frac{1}{M}\sum_{i=1}^M y_i(k)$, and $v_i(k) = y_i(k) - \theta^*_{Mk}$ for $k = 0, 1$. Then mild regularity conditions guarantee

$$\frac{1}{\sqrt{N}}\sum_{i=1}^M R_i g_i(X_i, \theta^*_M) = \frac{1}{\sqrt{N}}\sum_{i=1}^M R_i \begin{pmatrix} \frac{X_i v_i(1)}{\pi} - \frac{(1-X_i)v_i(0)}{1-\pi} \\ \frac{(X_i-\pi)(z_i-\bar{z}_M)}{\pi(1-\pi)} \end{pmatrix} + o_p(1), \tag{7}$$

by using $\hat{\pi} = \frac{\sum_{i=1}^M R_i X_i}{\sum_{i=1}^M R_i}$, $\frac{1}{N_1}\sum_{i=1}^M R_i X_i Y_i = \theta^*_{M1} + o_p(1)$, and $\frac{1}{N_0}\sum_{i=1}^M R_i(1-X_i)Y_i = \theta^*_{M0} + o_p(1)$ (due to the law of large numbers). By combining this and $G = \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^M \mathbb{E}\left[\frac{\partial g_i(X_i,\theta^*_M)}{\partial\theta}\right] = (-1, 0')'$ for (6), an analogous argument to Theorem 1 implies

$$\begin{aligned} \text{asy.var}[\sqrt{N}(\hat{\theta}_N(W) - \theta^*_M)] &= (G'WG)^{-1}G'W\Omega^* WG(G'WG)^{-1} \\ &= (1, W_1^{-1}W_{12}') \begin{pmatrix} \Omega^*_1 & \Omega^{*\prime}_{12} \\ \Omega^*_{12} & \Omega^*_2 \end{pmatrix} \begin{pmatrix} 1 \\ W_1^{-1}W_{12} \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \Omega^*_1 &= \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^M \left\{\frac{v_i(1)^2}{\pi} + \frac{v_i(0)^2}{1-\pi}\right\} - \rho \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^M (v_i(1) - v_i(0))^2 \\ \Omega^*_2 &= \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^M \frac{(z_i - \bar{z}_M)^2}{\pi(1-\pi)}, \\ \Omega^*_{12} &= \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^M \left\{\frac{(z_i - \bar{z}_M)v_i(1)}{\pi} + \frac{(z_i - \bar{z}_M)v_i(0)}{1-\pi}\right\}. \end{aligned}$$

Although we cannot consistently estimate the (inverse of) optimal weight $\Omega^* = \begin{pmatrix} \Omega^*_1 & \Omega^{*\prime}_{12} \\ \Omega^*_{12} & \Omega^*_2 \end{pmatrix}$ (especially Ω^*_1), an inspection of the above asymptotic variance reveals that $\hat{\theta}_N(W)$ can achieve the same asymptotic efficiency as the (infeasible) optimally weighted GMM estimator $\hat{\theta}_N(\Omega^{*-1})$ when the weight W satisfies

$$W_1^{-1}W_{12}' = -\Omega_{12}^{*\prime}\Omega_2^{*-1}. \tag{8}$$

A key observation is that the condition (8) does not involve Ω_1^*, which is typically impossible to estimate consistently under the present setup. Indeed, we can show that the (probability limit of) weight

$$W_{\mathrm{RA}} = \begin{pmatrix} 1 & \frac{1}{N}\sum_{i=1}^M R_i\left(\frac{X_i}{\hat{\pi}} - \frac{1-X_i}{1-\hat{\pi}}\right)^2 Y_i(z_i-\bar{z})' \\ \frac{1}{N}\sum_{i=1}^M R_i\left(\frac{X_i}{\hat{\pi}} - \frac{1-X_i}{1-\hat{\pi}}\right)^2 Y_i(z_i-\bar{z}) & \frac{1}{N}\sum_{i=1}^M R_i\left(\frac{X_i}{\hat{\pi}} - \frac{1-X_i}{1-\hat{\pi}}\right)^2 (z_i-\bar{z})(z_i-\bar{z})' \end{pmatrix}^{-1}, \tag{9}$$

satisfies the condition in (8) so that the resulting GMM estimator $\hat{\theta}_N(W_{\mathrm{RA}})$ is asymptotically as efficient as the optimally weighted GMM estimator $\hat{\theta}_N(\Omega^{*-1})$. We can also see that $\hat{\theta}_N(W_{\mathrm{RA}})$ is asymptotically equivalent to the regression adjustment estimator $\hat{\theta}_{\mathrm{RA}}$ in (3).

Moreover, since the difference-in-means estimator $\hat{\theta}_{\mathrm{DIM}}$ corresponds to $\hat{\theta}_N(W_{\mathrm{DIM}})$ with $W_{\mathrm{DIM}} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, we can conclude that the difference-in-means estimator is less efficient than $\hat{\theta}_{\mathrm{RA}}$ for all $\rho \in (0,1]$ under the finite-population asymptotics. These results are summarized in the following proposition. Recall $V_{\mathrm{GMM}}(W) = (G'WG)^{-1}G'W\Omega^*WG(G'WG)^{-1}$.

Proposition 1. *Consider the setup of Section 2.1 and this section. Suppose (7) holds true and* $\tilde{g}_i(X_i, \theta_M^*) = \begin{pmatrix} \frac{X_i v_i(1)}{\pi} - \frac{(1-X_i)v_i(0)}{1-\pi} \\ \frac{(X_i-\pi)(z_i-\bar{z}_M)}{\pi(1-\pi)} \end{pmatrix}$ *satisfies Assumptions (1)-(3). Then*

(1) $\sqrt{N}(\hat{\theta}_N(W_{\mathrm{RA}}) - \theta_M^*) \xrightarrow{d} N(0, V_{\mathrm{GMM}}(\Omega^{*-1}))$ and $\sqrt{N}(\hat{\theta}_N(W_{\mathrm{RA}}) - \hat{\theta}_{\mathrm{RA}}) = o_p(1)$.

(2) $\sqrt{N}(\hat{\theta}_N(W_{\mathrm{DIM}}) - \theta_M^*) \xrightarrow{d} N(0, V_{\mathrm{GMM}}(W_{\mathrm{DIM}}))$ and $V_{\mathrm{GMM}}(W_{\mathrm{DIM}}) \geq V_{\mathrm{GMM}}(\Omega^{*-1})$.

The proof of this proposition follows by an adaptation of that of Theorem 1 to $\tilde{g}_i(X_i, \theta_M^*)$ combined with (7). The proof of the statement $\sqrt{N}(\hat{\theta}_N(W_{\mathrm{RA}}) - \hat{\theta}_{\mathrm{RA}}) = o_p(1)$ follows by applying the same argument in Tsiatis *et al.* (2008) to the finite-population asymptotic setup. In contrast to the general case discussed in Section 3.2, we can achieve the asymptotically optimal variance $V_{\mathrm{GMM}}(\Omega^{*-1})$ by the feasible regression adjustment estimator $\hat{\theta}_N(W_{\mathrm{RA}})$ (or $\hat{\theta}_{\mathrm{RA}}$). Given this optimal property of $\hat{\theta}_N(W_{\mathrm{RA}})$, we can also conclude that the regression adjustment estimator is guaranteed to be asymptotically more efficient than the difference-in-means estimator regardless of correct specification of the linear regression model and for any sampling ratio $\rho \in (0, 1]$. For the case of $\rho = 0$ (i.e., the sampling-based or super-population setup), Negi and Wooldridge (2021) derived such efficiency gain. Our result shows that such efficiency gain continues to hold for any value of $\rho \in (0, 1]$ under the finite-population asymptotics.

Note that even though the point estimator $\hat{\theta}_N(W_{\mathrm{RA}})$ or $\hat{\theta}_{\mathrm{RA}}$ achieves asymptotic optimality in the class of GMM estimators, its asymptotic variance $V_{\mathrm{GMM}}(\Omega^{*-1}) = (G'\Omega^{*-1}G)^{-1}$ involves Ω^*, which cannot be consistently estimated in general. Thus, the resulting inference methods are conservative as in the general case.

Finally, the above results can be extended to stratified randomized experiments (e.g., Wang, Wang and Liu, 2023) with a finite number of strata, where for each group $g = 1, \ldots, G$, the assignment mechanism satisfies $\mathbb{P}\{X_i = 1 | G_i = g\} = \pi_g$ with the group indicator G_i. In this case, the moment function is replaced with

$$g_i(X_i, \theta_M^*) = \left(\begin{array}{c} \sum_{g=1}^{G} \frac{N_g}{N} \mathbb{I}\{G_i = g\} \left(\frac{X_i Y_i}{\hat{\pi}_g} - \frac{(1-X_i)Y_i}{1-\hat{\pi}_g} \right) - \theta_M^* \\ \left\{ \mathbb{I}\{G_i = g\} \left(\frac{X_i}{\hat{\pi}_g} - \frac{1-X_i}{1-\hat{\pi}_g} \right) z_i \right\}_{g=1}^{G} \end{array} \right),$$

where N_g is the size of the g-th stratum and $\hat{\pi}_g$ is the estimated assignment probability. An analogous argument yields efficiency guarantee of the regression adjustment estimator.

5. Numerical illustration

5.1. **Simulation.** This subsection illustrates our theoretical results on the regression adjustment estimator through Monte Carlo simulation. Our simulation design is motivated by Lin (2013) and Negi and Wooldridge (2021). For each population size $M \in \{500, 1000, 2000\}$, we generate finite population of M units with covariates $z_{1i} \sim U[-4, 4]$ and $z_{2i} \sim N(\mu, \Sigma)$, where $\mu = (0.5, 1)$ and $\Sigma = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 2 \end{pmatrix}$. Potential outcomes are generated as

$$\begin{aligned} y_i(0) &= \frac{\exp(z_{1i}) + \exp(z_{1i}/2)}{4} + z_{2i}'\beta_0 + u_{0i}, \\ y_i(1) &= \frac{-\exp(z_{1i}) + \exp(z_{1i}/2)}{4} + z_{2i}'\beta_1 + u_{1i}, \end{aligned}$$

where $\beta_0 = (-2, -2)'$, $\beta_1 = (2, 2)'$, and $u_{0i}, u_{1i} \sim N(0, 1)$ independent of $\{z_i\}_{i=1}^M$. The estimand of interest is the average treatment effect $\theta_M^* = M^{-1} \sum_{i=1}^M \{y_i(1) - y_i(0)\}$. In each Monte Carlo replication and for each sampling ratio $\rho \in \{0.25, 0.5, 0.75\}$, we implement Bernoulli sampling with the indicator $R_i \sim \text{Bernoulli}(\rho)$ from the population, and let $\mathcal{S} = \{i : R_i = 1\}$ with $N = |\mathcal{S}|$. Within the realized sample, treatment is assigned independently as $X_i \sim \text{Bernoulli}(\pi)$ with $\pi = 0.75$. In our preliminary simulation, we also consider the cases of $\pi = 0.25$ and 0.5, but the results are similar.

As a point estimator for θ_M^*, we consider the difference-in-means estimator $\hat{\theta}_{\text{DIM}}$ and regression adjustment estimator in $\hat{\theta}_{\text{RA}}$ in (3) with (z_{1i}, z_{2i}'). Furthermore, we consider the GMM estimator $\hat{\theta}_N(W_{\text{RA}})$ with the weight matrix W_{RA} in (9), which is shown to be asymptotically equivalent to $\hat{\theta}_{\text{RA}}$. The results are based on 10000 Monte Carlo replications.

Table 1 reports the absolute biases and standard deviations of the point estimators, average standard errors, and coverages of the 95% confidence intervals by these standard errors. We also report the standard deviations $\sqrt{V_{\text{GMM}}(W_{\text{DIM}})/N}$ and $\sqrt{V_{\text{GMM}}(W_{\text{RA}})/N}$ based on knowledge of the data generating process.

First, we compare the point estimators. In all designs, $\hat{\theta}_{\text{RA}}$ and $\hat{\theta}_N(W_{\text{RA}})$ are slightly biased compared to $\hat{\theta}_{\text{DIM}}$, which is expected due to unbiasedness of $\hat{\theta}_{\text{DIM}}$. The magnitudes of the biases generally decrease as M and ρ increase. The Monte Carlo standard deviations show efficiency gains by the regression adjustment. The standard deviation of $\hat{\theta}_{\text{RA}}$ is about $17.32 - 40.21\%$ smaller than that of $\hat{\theta}_{\text{DIM}}$, and the standard deviation of $\hat{\theta}_N(W_{\text{RA}})$ is about $18.19 - 40.58\%$ smaller than that of $\hat{\theta}_{\text{DIM}}$. Moreover, the Monte Carlo standard deviations of $\hat{\theta}_{\text{DIM}}$ and $\hat{\theta}_{\text{RA}}$ (or $\hat{\theta}_N(W_{\text{RA}})$) are close to the standard deviations $\sqrt{V_{\text{GMM}}(W_{\text{DIM}})/N}$ and $\sqrt{V_{\text{GMM}}(W_{\text{RA}})/N}$, respectively.

Next, we compare the variance estimators. For both W_{DIM} and W_{RA}, the reported standard errors are typically close to, but slightly above, the Monte Carlo standard deviations. Across all (M, ρ), the regression adjustment yields smaller standard errors than the difference-in-means by $12.39 - 14.55\%$ under V_{C} and $14.83 - 23.38\%$ under V_Z, respectively. In addition, the design-based standard errors V_Z are uniformly smaller than the conventional ones V_{C}.

TABLE 1. Simulation results ($\pi = 0.75$)

	$M = 500$			$M = 1000$			$M = 2000$		
	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
$\lvert\bar{\hat{\theta}}_{\mathrm{DIM}} - \theta^*_M\rvert$	0.0037	0.0008	0.0008	0.0046	0.0062	0.0052	0.0006	0.0010	0.0004
$\lvert\bar{\hat{\theta}}_{\mathrm{RA}} - \theta^*_M\rvert$	0.0597	0.0296	0.0178	0.0207	0.0100	0.0059	0.0107	0.0054	0.0039
$\lvert\bar{\hat{\theta}}_N(W_{\mathrm{RA}}) - \theta^*_M\rvert$	0.0352	0.0126	0.0105	0.0270	0.0138	0.0090	0.0091	0.0045	0.0027
SD $\hat{\theta}_{\mathrm{DIM}}$	1.0091	0.6283	0.4468	0.6683	0.4204	0.2948	0.4793	0.3050	0.2115
SD $\hat{\theta}_{\mathrm{RA}}$	0.8557	0.5108	0.3218	0.5700	0.3402	0.2103	0.4085	0.2457	0.1531
SD $\hat{\theta}_N(W_{\mathrm{RA}})$	0.8238	0.5032	0.3200	0.5609	0.3375	0.2097	0.4055	0.2446	0.1529
$\sqrt{V_{\mathrm{GMM}}(W_{\mathrm{DIM}})/N}$	0.9893	0.6227	0.4373	0.6643	0.4168	0.2911	0.4802	0.3022	0.2120
$\sqrt{V_{\mathrm{GMM}}(W_{\mathrm{RA}})/N}$	0.8367	0.4987	0.3142	0.5630	0.3342	0.2082	0.4079	0.2434	0.1535
SE $V_{\mathrm{C}}(W_{\mathrm{DIM}})$	1.0862	0.7689	0.6279	0.7330	0.5175	0.4230	0.5284	0.3733	0.3047
SE $V_{\mathrm{C}}(W_{\mathrm{RA}})$	0.9282	0.6651	0.5447	0.6350	0.4515	0.3693	0.4611	0.3268	0.2670
SE $V_Z(W_{\mathrm{DIM}})$	1.0061	0.6515	0.4771	0.6771	0.4352	0.3170	0.4902	0.3171	0.2326
SE $V_Z(W_{\mathrm{RA}})$	0.8406	0.5309	0.3670	0.5723	0.3561	0.2429	0.4176	0.2614	0.1810
Coverage $V_{\mathrm{C}}(W_{\mathrm{DIM}})$	0.9659	0.9811	0.995	0.9671	0.9828	0.9936	0.9721	0.9821	0.9951
Coverage $V_{\mathrm{C}}(W_{\mathrm{RA}})$	0.9713	0.9891	0.9992	0.9718	0.9900	0.9993	0.9733	0.9909	0.9988
Coverage $V_Z(W_{\mathrm{DIM}})$	0.9489	0.9554	0.9623	0.9525	0.9566	0.9646	0.9596	0.9559	0.9672
Coverage $V_Z(W_{\mathrm{RA}})$	0.9542	0.9592	0.9743	0.9549	0.9592	0.9774	0.9582	0.9622	0.9773

The relative reduction in standard errors is about $7.22 - 25.06\%$ for W_{DIM} and $9.43 - 32.63\%$ for W_{RA}, and the reduction becomes larger as ρ increases.

Finally, Table 1 reports the empirical coverages of the 95% confidence intervals. In general, the coverages are above 0.95 and increases with ρ. Comparing V_Z and V_{C}, the over-coverage is notable in V_{C} compared to V_Z under both W_{DIM} and W_{RA}. For instance, $V_Z(W_{\text{RA}})$ moves coverage closer to 0.95 by about $1.51 - 3.08$ percentage points than $V_{\text{C}}(W_{\text{RA}})$. Comparing weights W_{RA} and W_{DIM}, $V_Z(W_{\text{RA}})$ typically attains larger coverage than $V_Z(W_{\text{DIM}})$ by about 0.38–1.28 percentage points.

5.2. **Real data example.** In this section, we present an empirical illustration of the proposed GMM method. We revisit an influential work by Acemoglu, Johnson, and Robinson (2001) to study the effect of institutional quality on income across former European colonies by using settler mortality as an instrument for contemporary institutions. Their main specification relates log GDP per capita in 1995 to an index of protection against expropriation risk, instrumented by log settler mortality and reports large and statistically significant effects of institutions of former European colonies on income.

In our implementation, motivated by their robustness checks, we augment their baseline specification with a set of geography and health controls to address the concern that disease environment and related fundamentals may directly affect economic performance. Specifically, we include absolute latitude, the malaria index in 1994, life expectancy and infant mortality in 1995, continent dummies (Asia and Africa), and the share of a country's territory within 100 km of the coast. In all specifications, we estimate the model controlling for these covariates.

We view the dataset as a sample from a finite population of countries. Specifically, the data contain $N = 163$ countries while the United Nations has $M = 193$ member states, so we set the sampling ratio to $\rho = 0.84$ and vary the instrument set. We consider three specifications: (1) uses log settler mortality only, (2) adds yellow fever presence and democracy in the first year of independence, and (3) further adds European settlers in 1900 and constraints on the executive in 1900. For each specification, we report two GMM point estimates (i.e., $\hat{\theta}_N(\hat{\Omega}^{-1})$ and $\hat{\theta}_N((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$ in the previous section). We also report three corresponding standard errors: the conventional variance $V_C(\hat{\Omega}^{-1})$, the design-based variance $V_Z(\hat{\Omega}^{-1})$, and the design-based variance using the updated weight $V_Z((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$.

TABLE 2. Empirical illustration (with p-value in parentheses)

	(1)	(2)		(3)	
	$\hat{\theta}_N(\hat{\Omega}^{-1})$	$\hat{\theta}_N(\hat{\Omega}^{-1})$	$\hat{\theta}_N((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$	$\hat{\theta}_N(\hat{\Omega}^{-1})$	$\hat{\theta}_N((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$
Average risk protection	0.7744	0.3843	0.3906	0.3045	0.3331
SE $V_C(\hat{\Omega}^{-1})$	0.4712	0.1646	-	0.0921	-
	(0.1003)	(0.0196)		(0.0094)	
SE $V_Z(\hat{\Omega}^{-1})$	0.4403	0.1465	-	0.0878	-
	(0.0786)	(0.0087)		(0.0053)	
SE $V_Z((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$	-	-	0.1463	-	0.0867
			(0.0076)		(0.0012)
Over-identification test					
J-statistic	-	0.9325	1.0023	2.9098	3.3000
5% critical value of χ^2_{k-p}	-	5.9915	5.9915	9.4877	9.4877
		(0.6273)	(0.6058)	(0.5730)	(0.5089)
5% critical value of $\sum_{j=1}^{k-p} \lambda_{Z,j}\chi^2_{1,j}$	-	2.0987	2.4900	2.1443	2.5343
		(0.2544)	(0.28393)	(0.0136)	(0.0168)

Table 2 reports the GMM estimates, standard errors, and p-values in parentheses. Column (1) corresponds to the just-identified case with mortality as the only instrument. Column (2) uses three instruments and reports results under both the conventional and finite-population weights, and column (3) does the same with five instruments. Two findings are notable. First, in Column (1), the design-based standard error is smaller than the conventional one $V_C(\hat{\Omega}^{-1})$, which is about 0.47, while $V_Z(\hat{\Omega}^{-1})$ is about 0.44, a reduction of roughly 6%. The smaller design-based standard error may change the conclusion: the coefficient is not significant at the 10% level with the conventional variance, but it is significant at the 10% level with the design-based variance. Second, in column (2), the point estimates are similar under the two weight matrices, but the design-based standard error is approximately 11% smaller than the conventional standard error. As in the just-identified case, this reduction changes the conclusion: the coefficient on institutional quality, which is significant at the 5% level under $V_C(\hat{\Omega}^{-1})$, becomes significant at the 1% level under $V_Z(\hat{\Omega}^{-1})$ and $V_Z((\hat{\Omega} - \rho\hat{\Delta}_Z)^{-1})$. In Column (3), we increase the number of instruments and the magnitude of the point estimates changes little, but the standard errors fall sharply, by about 40.06–44.04% relative to column (2).

However, the overidentification tests highlight the limits of adding instruments in finite populations. We compute 5% critical values both for the standard chi-square distribution with degrees of freedom equal to the number of overidentification restrictions and for the design-based weighted chi-square distribution $\sum_{j=1}^{k-p} \lambda_{Z,j}\chi^2_{1,j}$, where one million Monte Carlo simulation draws approximate the latter distribution. Using the conventional χ^2_{k-p} critical values, the overidentification restrictions are essentially never rejected, even in the five-instrument specification. By contrast, when we use the correct weighted chi-square critical values, column (2) continues to pass the overidentification test, but the five-instrument specification in column (3) is rejected at the 5% level although we use the conservative critical value of $\sum_{j=1}^{k-p} \lambda_{Z,j}\chi^2_{1,j}$. This pattern cautions against mechanically adding instruments: additional instruments may shrink standard errors but may fail the overidentification test.

Appendix A. Mathematical appendix

A.1. **Proof of Theorem 1.**

A.1.1. *Proof of Part (1).* It is sufficient to verify the conditions in Newey and McFadden (1994, Theorem 2.1). Their condition (i) is satisfied due to the uniqueness of θ^* and the positive definiteness of W in Assumption (2). Their condition (ii) (i.e., compactness of the parameter space) is directly imposed.

To verify their conditions (iii) and (iv), it is sufficient to show that

$$\sup_{\theta \in \Theta} \left\| \frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \theta) - \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta)] \right\| \xrightarrow{p} 0, \tag{10}$$

and $\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta)]$ is continuous at each $\theta \in \Theta$. The continuity of $\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta)]$ follows by the dominated convergence theorem and Jensen's inequality due to the conditions on $g_i(X_i, \theta)$ in Assumption (2). For (10), we first note that Abadie *et al.* (2014, Lemma A.2) implies the pointwise convergence

$$\frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \theta) - \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[g_i(X_i, \theta)] \xrightarrow{p} 0, \quad \text{for each } \theta \in \Theta.$$

Then Newey (1991, Corollary 2.2) combined with the Lipschitz condition in Assumption (2) implies the uniform convergence in (10).

Since we verify all the conditions of Newey and McFadden (1994, Theorem 2.1), the conclusion follows.

A.1.2. *Proof of Part (2).* Let $\hat{G}_N(\theta) = \frac{1}{N} \sum_{i=1}^{M} R_i \frac{\partial g_i(X_i, \theta)}{\partial \theta'}$. By the consistency of $\hat{\theta}_N(W_N)$ and Assumption (3) ($\theta^* \in \mathrm{int}(\Theta)$ and differentiability of $g_i(x, \theta)$), the estimator $\hat{\theta}_N(W_N)$

satisfies the first-order condition

$$\hat{G}_N(\hat{\theta}_N(W_N))'W_N \left\{ \frac{1}{N} \sum_{i=1}^{M} R_i g_i(X_i, \hat{\theta}_N(W_N)) \right\} = 0,$$

with probability approaching one. By expanding $g_i(X_i, \hat{\theta}_N(W_N))$ around $\hat{\theta}_N(W_N) = \theta^*_M$ and solving for $\hat{\theta}_N(W_N) - \theta^*_M$, we obtain

$$\sqrt{N}(\hat{\theta}_N(W_N) - \theta^*_M) = [\hat{G}_N(\hat{\theta}_N(W_N))'W_N \hat{G}_N(\tilde{\theta}_N)]^{-1} \hat{G}_N(\hat{\theta}_N(W_N))'W_N \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_i g_i(X_i, \theta^*_M), \tag{11}$$

where $\tilde{\theta}_N$ is a point on the line joining $\hat{\theta}_N(W_N)$ and θ^*_M. Since $\hat{\theta}_N(W_N) - \theta^*_M \xrightarrow{p} 0$ and $\tilde{\theta}_N - \theta^*_M \xrightarrow{p} 0$, it is sufficient for the conclusion to show that

$$\sup_{\theta \in \mathcal{N}} \left\| \hat{G}_N(\theta) - \mathbb{E}\left[\frac{\partial g_i(X_i, \theta)}{\partial \theta'} \right] \right\| \xrightarrow{p} 0, \tag{12}$$

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_i g_i(X_i, \theta^*_M) \xrightarrow{d} N(0, \Omega - \rho\Delta), \tag{13}$$

for some neighborhood $\mathcal{N}$ around θ^*.

For (12), Abadie *et al.* (2014, Lemma A.2) under Assumption (3) implies the pointwise convergence $\hat{G}_N(\theta) - \mathbb{E}\left[\frac{\partial g_i(X_i,\theta)}{\partial \theta'}\right] \xrightarrow{p} 0$ for each $\theta \in \mathcal{N}$. Then Newey (1991, Corollary 2.2) combined with the Lipschitz condition in Assumption (3) implies the uniform convergence in (12). For (13), it follows directly from Abadie *et al.* (2020, Lemma A.1) under Assumptions (2)-(3). Therefore, the conclusion follows.

A.2. **Proof of $V_Z((\Omega - \rho\Delta_Z)^{-1}) \leq_{\text{pd}} V_Z(\Omega^{-1})$ in Theorem 2 (3).** Let $A = \Omega^{-1}G(G'\Omega^{-1}G)^{-1}$ and $B = (\Omega - \rho\Delta_Z)^{-1}G(G'(\Omega - \rho\Delta_Z)^{-1}G)^{-1}$ so that $V_Z(\Omega^{-1}) = A'(\Omega - \rho\Delta_Z)A$ and $V_Z((\Omega -$

$\rho\Delta_Z)^{-1}) = B'(\Omega - \rho\Delta_Z)B$. Observe that

$$\begin{aligned} V_Z(\Omega^{-1}) &= \{B + (A - B)\}'(\Omega - \rho\Delta_Z)\{B + (A - B)\} \\ &= V_Z((\Omega - \rho\Delta_Z)^{-1}) + (A - B)'(\Omega - \rho\Delta_Z)(A - B), \end{aligned}$$

where the second equality follows from

$$B'(\Omega - \rho\Delta_Z)(A - B) = (G'(\Omega - \rho\Delta_Z)^{-1}G)^{-1} - (G'(\Omega - \rho\Delta_Z)^{-1}G)^{-1} = 0.$$

Since $(A - B)'(\Omega - \rho\Delta_Z)(A - B) \geq_{\text{pd}} 0$, we obtain the conclusion.

A.3. **Proof of Corollary 1.**

A.3.1. *Proof of Part (1).* It is sufficient to verify the conditions in Newey and McFadden (1994, Theorem 2.1). Their condition (i) is now satisfied due to the uniqueness of the pseudo true value $\theta^{\#}(W)$ in Assumption (2)'. Verifications of the other conditions are identical to those of Theorem 1 (1).

A.3.2. *Proof of Part (2).* The proof is an adaptation of that of Hall and Inoue (2003, Theorem 2). To simplify the presentation, denote $\hat{\theta}_N = \hat{\theta}_N(W_N)$ and $\theta^{\#}_M = \theta^{\#}_M(W)$. First, the same argument to (11) yields

$$
\begin{aligned}
\sqrt{N}(\hat{\theta}_N - \theta_M^{\#}) &= H_N^{-1}\hat{G}_N(\hat{\theta}_N)'W_N\frac{1}{\sqrt{N}}\sum_{i=1}^{M} R_i g_i(X_i, \theta_M^{\#}) \\
&= H_N^{-1}\hat{G}_N(\hat{\theta}_N)'W_N\frac{1}{\sqrt{N}}\sum_{i=1}^{M} R_i\{g_i(X_i, \theta_M^{\#}) - \mathbb{E}[g_i(X_i, \theta_M^{\#})]\} \\
& +H_N^{-1}\sqrt{N}\{\hat{G}_N(\hat{\theta}_N) - \hat{G}_N(\theta_M^{\#})\}'W_N\frac{1}{N}\sum_{i=1}^{M} R_i\mathbb{E}[g_i(X_i, \theta_M^{\#})] \\
& +H_N^{-1}\sqrt{N}\{\hat{G}_N(\theta_M^{\#}) - G_N(\theta_M^{\#})\}'W_N\frac{1}{N}\sum_{i=1}^{M} R_i\mathbb{E}[g_i(X_i, \theta_M^{\#})] \\
& +H_N^{-1}G_N(\theta_M^{\#})'\sqrt{N}\{W_N - W\}\frac{1}{N}\sum_{i=1}^{M} R_i\mathbb{E}[g_i(X_i, \theta_M^{\#})] \\
& +H_N^{-1}G_N(\theta_M^{\#})'W\frac{1}{\sqrt{N}}\sum_{i=1}^{M} R_i\mathbb{E}[g_i(X_i, \theta_M^{\#})] \\
&=: T_1 + \cdots + T_5,
\end{aligned}
$$

where $H_N = \hat{G}_N(\hat{\theta}_N)'W_N\hat{G}_N(\tilde{\theta}_N)$, $\tilde{\theta}_N$ is a point on the line joining $\hat{\theta}_N$ and $\theta_M^{\#}$, and $G_N(\theta_M^{\#}) = \frac{1}{N}\sum_{i=1}^{M} R_i\mathbb{E}\left[\frac{\partial g_i(X_i,\theta_M^{\#})}{\partial\theta'}\right]$.

By an expansion around $\hat{\theta}_N = \theta_M^{\#}$, T_2 is written as

$$
T_2 = H_N^{-1}M_N\sqrt{N}(\hat{\theta}_N - \theta_M^{\#}),
$$

where $M_N = \left(\frac{1}{N}\sum_{i=1}^{M} R_i\mathbb{E}[g_i(X_i, \theta_M^{\#})]'W_N \otimes I\right) G_N^{(2)}(\dot{\theta}_N)$, $G_N^{(2)}(\theta) = \frac{1}{N}\sum_{i=1}^{M} R_i\frac{\partial}{\partial\theta'}\text{vec}\left\{\frac{\partial g_i(X_i,\theta)}{\partial\theta'}\right\}$, and $\dot{\theta}_N$ is a point on the line joining $\hat{\theta}_N$ and $\theta_M^{\#}$. Also the first-order condition of $\theta_M^{\#}$ implies

$T_5 = 0$. Thus, we obtain

$$
\begin{aligned}
\sqrt{N}(\hat{\theta}_N - \theta_M^*) &= (I - H_N^{-1} M_N)^{-1} \{T_1 + T_3 + T_4\} \\
&= (I - H^{-1} M)^{-1} H^{-1} G^{\#\prime} W \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_i \{g_i(X_i, \theta_M^{\#}) - \mathbb{E}[g_i(X_i, \theta_M^{\#})]\} \\
&\quad + (I - H^{-1} M)^{-1} H^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{M} R_i \left\{ \frac{\partial g_i(X_i, \theta_M^{\#})}{\partial \theta'} - \mathbb{E}\left[\frac{\partial g_i(X_i, \theta_M^{\#})}{\partial \theta'} \right] \right\}' W \mu^{\#} \\
&\quad + (I - H^{-1} M)^{-1} H^{-1} G^{\#\prime} \sqrt{N} (W_N - W) \mu^{\#},
\end{aligned}
$$

where the second equality follows from $W_N \xrightarrow{p} W$ (Assumption (2)') and $\frac{1}{N} \sum_{i=1}^{M} R_i \mathbb{E}[g_i(X_i, \theta_M^{\#})] \xrightarrow{p} \mu^{\#}$, $\hat{G}_N(\hat{\theta}_N) \xrightarrow{p} G^{\#}$, and $G_N^{(2)}(\dot{\theta}_N) \xrightarrow{p} G^{(2)\#}$ (obtained by the uniform law of large numbers under Assumption (3)'). Therefore, the conclusion follows by Assumption (3)'.

A.4. **Proof of Corollary 2.** Let $\bar{g}_N(X_i, \theta_M^*) = \frac{1}{N} \sum_{i=1}^{M} R_i g(X_i, \theta_M^*)$ and $G_N(X_i, \theta_M^*) = \frac{1}{N} \sum_{i=1}^{M} R_i \frac{\partial g_i(X_i, \theta_M^*)}{\partial \theta'}$. Assumptions (1)–(3) imply $G_N(X_i, \theta_M^*) \xrightarrow{p} G$, $W_N \xrightarrow{p} W$, and

$$
\sqrt{N} \bar{g}_N(X_i, \theta_M^*) \xrightarrow{d} Z \sim N(0, \Omega - \rho \Delta).
$$

Thus, an expansion of the first-order condition of $\hat{\theta}_N$ yields

$$
\begin{aligned}
\sqrt{N}(\hat{\theta}_N - \theta_M^*) &= -\{G_N(X_i, \hat{\theta}_N)' W_N G_N(X_i, \bar{\theta})\}^{-1} G_N(X_i, \hat{\theta}_N)' W_N \sqrt{N} \bar{g}_N(X_i, \theta_M^*) + o_p(1) \\
&\xrightarrow{d} -(G' W G)^{-1} G' W Z.
\end{aligned}
$$

Combining these results, we obtain

$$\begin{aligned}\sqrt{N}\bar{g}_N(X_i,\hat{\theta}_N) &= [I_k - \{G_N(X_i,\hat{\theta}_N)'W_N G_N(X_i,\bar{\theta})\}^{-1}G_N(X_i,\hat{\theta}_N)'W_N]\sqrt{N}\bar{g}_N(X_i,\theta_M^*) + o_p(1) \\ &\xrightarrow{d} \{I_k - G(G'WG)^{-1}G'W\}Z.\end{aligned}$$

Therefore, the test statistic satisfies

$$J_N(W_N) \xrightarrow{d} Z'\{W - WG(G'WG)^{-1}G'W\}Z =: Q(W).$$

For Part (i) of this corollary, let $W = (\Omega - \rho\Delta)^{-1}$ and $\tilde{Z} = (\Omega - \rho\Delta)^{-1/2}Z \sim N(0, I_k)$. Then we have $Q(W) = \tilde{Z}'(I_k - P)\tilde{Z}$, where

$$P = (\Omega - \rho\Delta)^{-1/2}G(G'(\Omega - \rho\Delta)^{-1}G)^{-1}G'(\Omega - \rho\Delta)^{-1/2}.$$

Since $I_k - P$ is the symmetric idempotent matrix of rank $k - p$, we obtain the conclusion $Q(W) \sim \chi^2_{k-p}$.

For Part (ii) of this corollary, Hansen (2021, Lemma 1) implies that if $\{\lambda_j\}_{j=1}^{k-p}$ are the nonzero eigenvalues of $(\Omega - \rho\Delta)\{W - WG(G'WG)^{-1}G'W\}$, then $Q(W) \sim \sum_{j=1}^{k-p}\lambda_j\chi^2_{1,j}$.

References

[1] Abadie, A., Athey, S., Imbens, G. W. and J. M. Wooldridge (2014) Finite population causal standard errors, Working paper no. 20325, National Bureau of Economic Research.

[2] Abadie, A., Athey, S., Imbens, G. W. and J. M. Wooldridge (2020) Sampling-based versus design-based uncertainty in regression analysis, *Econometrica*, 88, 265-296.

[3] Acemoglu, D., Johnson, S. and J. A. Robinson (2001) The colonial origins of comparative development: an empirical investigation, American Economic Review, 91, 1369-1401.

[4] Duflo, E. and A. Banerjee (2017) *Handbook of field experiments*, Elsevier.

[5] Freedman, D. A. (2008a) On regression adjustments in experiments with several treatments, *Annals of Applied Statistics*, 2, 176-96.

[6] Freedman, D. A. (2008b) On regression adjustments to experimental data, *Advances in Applied Mathematics,* 40, 180-93.

[7] Hall, A. R. (2005) *Generalized Method of Moments*, Oxford University Press.

[8] Hall, A. R. and A. Inoue (2003) The large sample behaviour of the generalized method of moments estimator in misspecified models, *Journal of Econometrics*, 114, 361-394.

[9] Hansen, B. E. (2021) Criterion-based inference without the information equality: the weighted chi-square distribution, Working paper.

[10] Imbens, G. W. and T. Lancaster (1994) Combining micro and macro data in microeconometric models, *Review of Economic Studies*, 61, 655-680.

[11] Jiang, L., Linton, O. B., Tang, H., and Zhang, Y. (2024) Improving estimation efficiency via regression-adjustment in covariate-adaptive randomizations with imperfect compliance, *Review of Economics and Statistics*, 1-45.

[12] Kang, H., Peck, L. and L. Keele (2018) Inference for instrumental variables: a randomization inference approach, *Journal of the Royal Statistical Society Series A*, 181, 1231-1254.

[13] Li, X. and P. Ding (2017) General forms of finite population central limit theorems with applications to causal inference, *Journal of the American Statistical Association*, 112, 1759-1769.

[14] Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique, *Annals of Applied Statistics*, 7, 295-318.

[15] Muralidharan, K. and P. Niehaus (2017) Experimentation at scale, *Journal of Economic Perspectives*, 31, 103-124.

[16] Negi, A. and J. M. Wooldridge (2021) Revisiting regression adjustment in experiments with heterogeneous treatment effects, *Econometric Reviews,* 40, 504-534.

[17] Newey, W. K. (1991) Uniform convergence in probability and stochastic equicontinuity, *Econometrica*, 59, 1161-7.

[18] Newey, W. K. and D. L. McFadden (1994) Large sample estimation and hypothesis testing, in Engle, R. F. and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, 2111-2245, Elsevier.

[19] Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Sciences,* 10, 1–5

[20] Práŝková, Z. and P. K. Sen (2009) Asymptotics in finite population sampling, *Handbook of Statistics*, 29, 489-522, Elsevier.

[21] Rambachan, A. and Roth, J (2025) Design-based uncertainty for quasi-experiments, *Journal of the American Statistical Association*, 1-15.

[22] Ren, J. (2024) Model-assisted complier average treatment effect estimates in randomized experiments with noncompliance, *Journal of Business & Economic Statistics*, 42(2), 707-718.

[23] Rosenbaum, P. R. (2002) *Observational studies*, Springer.

[24] Sakamoto, K., and Y. Shimizu (2025) Design-based and network sampling-based uncertainties in network experiments, arXiv:2506.22989.

[25] Tsiatis, A., Davidian, M., Zhang, M. and X. Lu (2008) Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach, *Statistics in Medicine*, 27, 4658-4677.

[26] Wang, X., Wang, T. and Liu, H. (2023) Rerandomization in stratified randomized experiments, *Journal of the American Statistical Association*, 118, 1295-1304.

[27] Xu, R. (2021) Potential outcomes and finite-population inference for M-estimators, *Econometrics Journal*, 24, 162-176.

[28] Xu, R. and J. M. Wooldridge (2022) A design-based approach to spatial correlation, *arXiv preprint arXiv:2211.14354.*

Department of Economics, University of Wisconsin-Madison, 1180 Observatory Dr., Madison, 53706, WI, USA.

Email address: haruo.kakehi@wisc.edu

Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan.

Email address: matsushita.y@r.hit-u.ac.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

Email address: t.otsu@lse.ac.uk