ORIGINAL ARTICLE

Model-assisted analyses of cluster-randomized experiments

Fangzhou Su | Peng Ding [ORCID]

Department of Statistics, University of California, Berkeley, California, USA

Correspondence
Peng Ding, Department of Statistics, University of California, Berkeley, CA 94720, USA.
Email: pengdingpku@berkeley.edu

Abstract

Cluster-randomized experiments are widely used due to their logistical convenience and policy relevance. To analyse them properly, we must address the fact that the treatment is assigned at the cluster level instead of the individual level. Standard analytic strategies are regressions based on individual data, cluster averages and cluster totals, which differ when the cluster sizes vary. These methods are often motivated by models with strong and unverifiable assumptions, and the choice among them can be subjective. Without any outcome modelling assumption, we evaluate these regression estimators and the associated robust standard errors from the design-based perspective where only the treatment assignment itself is random and controlled by the experimenter. We demonstrate that regression based on cluster averages targets a weighted average treatment effect, regression based on individual data is suboptimal in terms of efficiency and regression based on cluster totals is consistent and more efficient with a large number of clusters. We highlight the critical role of covariates in improving estimation efficiency and illustrate the efficiency gain via both simulation studies and data analysis. The asymptotic analysis also reveals the efficiency-robustness trade-off by comparing the properties of various estimators using data at different levels with and without covariate adjustment. Moreover, we show that the robust standard errors are convenient approximations to the true asymptotic

---

standard errors under the design-based perspective. Our theory holds even when the outcome models are misspecified, so it is model-assisted rather than model-based. We also extend the theory to a wider class of weighted average treatment effects.

KEYWORDS

analysis of covariance, design-based inference, efficiency-robustness trade-off, group-randomized trial, potential outcomes, robust standard error

1 | INTRODUCTION

Cluster-randomized experiments, also known as group-randomized trials, are widely used in empirical research, where randomization over units within clusters is either unethical or logistically infeasible. For example, in many public health interventions, clusters are villages, units are households and randomization is implemented at the village level (Donner & Klar, 2000; Turner et al., 2017a,b); in many educational interventions, clusters are classrooms, units are students and randomization is implemented at the classroom level (Raudenbush, 1997; Raudenbush & Schwartz, 2020; Schochet, 2013; Schochet et al., 2021). In addition to their convenience, cluster-randomized experiments can circumvent the problem of interference among units and are policy relevant if the intervention of a policy is at the cluster level.

A proper analysis of a cluster-randomized experiment must first clearly specify the population and parameter of interest and then address the fact that randomization is at the cluster level. Model-based analyses address these issues simultaneously by imposing certain parametric assumptions and correlation structure on the error terms (Donner & Klar, 2000; Graubard & Korn, 1994; Green & Vavreck, 2008). However, these modelling assumptions are often too strong and can lead to bias under model misspecifications. We take an alternative perspective by first defining the parameter of interest based on potential outcomes for treatment effects and then deriving the properties of the regression estimators under the randomization of the treatment assignment. This design-based or randomization-based perspective has been exploited by Miratrix et al. (2013), Dasgupta et al. (2015), Imbens and Rubin (2015), Fogarty (2018a), Mukerjee et al. (2018) and Liu and Yang (2020) for analysing various randomized experiments, and by Imai et al. (2009), Schochet (2013), Middleton and Aronow (2015), Athey and Imbens (2017), Abadie et al. (2017) and Schochet et al. (2021) for analysing cluster-randomized experiments in particular.

We will discuss regression estimators based on individual data, cluster totals and cluster averages, with and without adjusting for covariates. We demonstrate that regression estimators based on individual data and cluster totals are consistent for the average treatment effect, with the latter giving more efficient estimators when adjusted for cluster sizes and cluster-level covariates. Our results supplement Schochet et al. (2021) by providing a unified theory for regressions using data at different levels and regressions with treatment–covariates interaction. We quantify the efficiency-robustness trade-off by comparing the design-based properties of various estimators. In particular, our theory demonstrates that the covariate-adjusted estimator based on cluster totals has higher asymptotic efficiency with predictive covariates at the cost of robustness when the

cluster sizes have large variability. In contrast, the covariate-adjusted estimator based on individual data and the unadjusted estimator based on cluster totals are suboptimal in asymptotic efficiency but generally more robust. We also show that the cluster-robust standard errors (Liang & Zeger, 1986) are conservative estimators for the true asymptotic standard errors in individual-level regressions, and heteroskedasticity-robust standard errors (Huber, 1967; White, 1980) are conservative estimators for the true asymptotic standard errors in cluster-total regressions. The former has been implemented empirically (Athey & Imbens, 2017; Green & Vavreck, 2008; Schochet, 2013). Schochet et al. (2021) provide some results for the case without covariates, and we extend their results by establishing a more unified theory for the design-based properties of the cluster-robust standard errors especially for the individual-level regression estimators adjusted for covariates. The latter extends Lin (2013)'s theory for completely randomized experiments to accommodate the cluster structure.

Middleton (2008) criticized the cluster-average regression because of its inconsistency for the average treatment effect. In contrast, we argue that cluster-average regressions view clusters themselves as the population and target a different estimand that is informative for many applications (Athey & Imbens, 2017; Donner & Klar, 2000). We define a more general estimand as the weighted average of cluster-level effects and show that weighted least squares (WLS) based on cluster averages yields consistent estimators for this estimand. We further propose a new regression estimator based on weighted cluster averages, which dominates cluster-average regressions and can leverage covariate information to improve efficiency.

Although the estimators are motivated by parametric or semi-parametric models, we extend the current literature on agnostic regressions (Fogarty, 2018b; Lin, 2013; Liu & Yang, 2020; Schochet et al., 2021), and derive their design-based properties allowing for misspecifications of these models. Our theory has clear practical implications. For the average treatment effect, we recommend using the cluster-total regression adjusted for the cluster size and cluster-level of covariates. For the weighted average effect, we recommend using the new regression estimator based on weighted cluster averages adjusted for the weight and cluster-level of covariates. The heteroskedasticity-robust standard errors for both estimators provide convenient approximations for the true asymptotic standard errors in constructing confidence intervals. The recommended estimators can be easily implemented by standard software packages with appropriately specified covariates and weights. Importantly, these simple estimators have strong theoretical guarantees.

The paper proceeds as follows. Section 2 sets up the basic notation and framework for the design-based inference in cluster-randomized experiments, and introduces four regression estimators based on individual data or cluster totals, with or without covariates adjustment. Section 3 discusses the asymptotic properties of the four estimators introduced in Section 2. Section 4 proposes a general class of weighted average treatment effects over clusters, and analyses four estimators that are analogous to those in Section 2. Section 5 unifies the theory and gives practical recommendations. Section 6 uses simulation studies to evaluate the finite-sample properties of various estimators and robust standard errors. Section 7 illustrates the theory with a recent cluster-randomized experiment. Section 8 concludes and the supplementary material contains the proofs. Replication files for this paper are publicly available at Harvard Dataverse: https://doi.org/10.7910/DVN/7FBJPU

We use the following notation throughout the paper. Let $o_{\mathbb{P}}(1)$ denote a random variable converging to zero in probability, with the probability measure $\mathbb{P}$ induced by the randomization of the treatment assignment. Let $R_M \rightsquigarrow \mathcal{N}(0, 1)$ represent that R_M converges in distribution to a standard Normal random variable as the number of clusters M goes to infinity. Let 'se' and '$\hat{se}$' denote the asymptotic standard error and the estimated asymptotic standard error of an estimator.

2 | NOTATION AND FRAMEWORK

2.1 | Potential outcomes, treatment assignment and observed data

Consider a study with N units, clustered, for example, by classrooms or villages. Cluster i has n_i units $(i = 1, \ldots, M)$, and the total number of units is $N = \sum_{i=1}^{M} n_i$. Let (i, j) index the jth unit within cluster i $(i = 1, \ldots, M; j = 1, \ldots, n_i)$. Unit (i, j) has covariates x_{ij}, and cluster i has covariates c_i. The experimenter randomly assigns eM clusters to receive the treatment and $(1 - e)M$ clusters to receive the control, where $0 < e < 1$ is a fixed number denoting the proportion of treated clusters. Let Z_i be the treatment indicator for cluster i and Z_{ij} be the treatment indicator for unit (i, j). In a cluster-randomized experiment, units within a cluster receive identical treatment levels. So if cluster i receives treatment, then $Z_{ij} = Z_i = 1$; if cluster i receives control, then $Z_{ij} = Z_i = 0$. Let $\sum_{ij} = \sum_{i=1}^{M} \sum_{j=1}^{n_i}$ denote the summation over all units. Let $\mathcal{T} = \{(i, j): Z_{ij} = 1\}$ be the indices of units under treatment and $\mathcal{C} = \{(i, j): Z_{ij} = 0\}$ be the indices of units under control. Their cardinalities $n_{\mathcal{T}} = \sum_{ij} Z_{ij}$ and $n_{\mathcal{C}} = \sum_{ij}(1 - Z_{ij})$ represent the total numbers of units under treatment and control, respectively, which are random if the cluster sizes vary.

For unit (i, j), let $Y_{ij}(1)$ and $Y_{ij}(0)$ be the potential outcomes (Neyman, 1923) under treatment and control, respectively. The average potential outcome in treatment arm z $(z = 0, 1)$ is $\overline{Y}(z) = N^{-1} \sum_{ij} Y_{ij}(z)$, and the average treatment effect is

$$\tau = \overline{Y}(1) - \overline{Y}(0) = N^{-1} \sum_{ij} \{Y_{ij}(1) - Y_{ij}(0)\}$$

over all units. The observed outcome is a function of the treatment indicator and potential outcomes: $Y_{ij} = Z_{ij} Y_{ij}(1) + (1 - Z_{ij}) Y_{ij}(0)$. A central goal in analysing a cluster-randomized experiment is to make inference about τ using the observed data $\{(Z_{ij}, Y_{ij}, x_{ij}, c_i): i = 1, \ldots, M; j = 1, \ldots, n_i\}$. Our analysis is design-based, that is, we condition on all covariates and potential outcomes, with randomness coming from $(Z_1, \ldots, Z_n)$, a random permutation of eM 1's and $(1 - e)M$ 0's. Conditioning on the potential outcomes, we have a finite population of N units and do not impose any stochastic outcome modelling assumption.

2.2 | A special case with $n_i = 1$: completely randomized experiment

To motivate model-assisted analyses of cluster-randomized experiments, we first review the existing design-based theory for completely randomized experiments. The special case has $n_i = 1$, and simplifies the double index (i, j) to the single index i. The average treatment effect reduces to $\tau = M^{-1} \sum_{i=1}^{M} \{Y_i(1) - Y_i(0)\}$. Neyman (1923) showed that the difference-in-means estimator $\hat{\tau}$ is unbiased for τ. Fisher (1935) proposed to use the coefficient of Z_i in the ordinary least squares (OLS) fit of Y_i on $(1, Z_i, c_i)$ to estimate τ, hoping to leverage the information in c_i to improve estimation efficiency. However, Freedman (2008a,b) criticized Fisher (1935)'s regression estimator because it can be even less efficient than $\hat{\tau}$. Lin (2013) proposed a slight modification to address Freedman (2008a,b)'s critiques: with centred covariates such that $\overline{c} = M^{-1} \sum_{i=1}^{M} c_i = 0$, the coefficient of Z_i in the OLS fit of Y_i on $(1, Z_i, c_i, Z_i c_i)$ is consistent and asymptotically more efficient than $\hat{\tau}$, with the true asymptotic standard error conservatively estimated by the heteroskedasticity-robust standard error.

Importantly, Freedman (2008a,b) and Lin (2013) did not assume that the linear models are correctly specified. They conditioned on the potential outcomes and evaluated the regression estimators

from the design-based perspective. So their theories are model-assisted rather than model-based. An overarching goal of this paper is to develop the design-based theory for commonly used regression estimators and robust standard errors for analysing cluster-randomized experiments.

2.3 | Some common estimators and their regression formulations

Motivated by the definition $\tau = \overline{Y}(1) - \overline{Y}(0)$, it is intuitive to use the difference-in-means $\hat{\tau}_I = \overline{Y}_\mathcal{T} - \overline{Y}_\mathcal{C}$ as an estimator, where $\overline{Y}_\mathcal{T} = n_\mathcal{T}^{-1} \sum_{ij \in \mathcal{T}} Y_{ij}$ and $\overline{Y}_\mathcal{C} = n_\mathcal{C}^{-1} \sum_{ij \in \mathcal{C}} Y_{ij}$ are the average observed outcomes under treatment and control, respectively. Numerically, $\hat{\tau}_I$ is identical to the coefficient of Z_{ij} in the OLS fit of Y_{ij} on $(1, Z_{ij})$ with all (i, j)'s. The clustering structure of the data suggests using the cluster-robust standard error $\hat{se}_{LZ}(\hat{\tau}_I)$ (Athey & Imbens, 2017; Barrios et al., 2012; Liang & Zeger, 1986; Schochet, 2013) to construct a Wald-type confidence interval.

Lin (2013)'s analysis of completely randomized experiments motivates us to use covariates to further improve the estimation efficiency. We modify the above regression as follows: first centre the unit-level covariates such that $\overline{x} = N^{-1} \sum_{ij} x_{ij} = 0$; then obtain $\hat{\tau}_I^{\mathrm{adj}}$, the coefficient of Z_{ij} in the OLS fit of Y_{ij} on $(1, Z_{ij}, x_{ij}, Z_{ij}x_{ij})$; finally report the cluster-robust standard error $\hat{se}_{LZ}(\hat{\tau}_I^{\mathrm{adj}})$.

On the other hand, the parameter τ has an alternative form:

$$\tau = M^{-1} \sum_{i=1}^{M} \{\tilde{Y}_{i\cdot}(1) - \tilde{Y}_{i\cdot}(0)\},$$

where $\tilde{Y}_{i\cdot}(z) = \sum_{j=1}^{n_i} Y_{ij}(z)M/N$ is the scaled cluster total potential outcome under treatment arm z ($z = 0, 1$). So τ is also the average treatment effect on the cluster level with outcomes defined as the scaled cluster totals. With the observed scaled cluster total $\tilde{Y}_{i\cdot} = Z_i \tilde{Y}_{i\cdot}(1) + (1 - Z_i)\tilde{Y}_{i\cdot}(0)$, we can use

$$\begin{aligned}
\hat{\tau}_T &= (eM)^{-1} \sum_{Z_i=1} \tilde{Y}_{i\cdot} - \{(1-e)M\}^{-1} \sum_{Z_i=0} \tilde{Y}_{i\cdot} \\
&= (eN)^{-1} \sum_{ij \in \mathcal{T}} Y_{ij} - \{(1-e)N\}^{-1} \sum_{ij \in \mathcal{C}} Y_{ij}
\end{aligned}$$

to estimate τ (Athey & Imbens, 2017; Li & Ding, 2017; Middleton & Aronow, 2015). Numerically, $\hat{\tau}_T$ is identical to the coefficient of Z_i in the OLS fit of $\tilde{Y}_{i\cdot}$ on $(1, Z_i)$ with the cluster total data. Since this regression uses aggregate data, we do not need to account for clustering and can simply use the heteroskedasticity-robust standard error $\hat{se}_{HW}(\hat{\tau}_T)$ (Huber, 1967; White, 1980).

Again, we can modify the above regression to leverage the cluster-level covariates to improve efficiency. Because we have completely randomized experiments on clusters, Li and Ding (2017) suggested to extend Lin (2013)'s estimator: first centre the cluster-level covariates at $\overline{c} = M^{-1} \sum_{i=1}^{M} c_i = 0$; then obtain $\hat{\tau}_T^{\mathrm{adj}}$, the coefficient of Z_i in the OLS fit of $\tilde{Y}_{i\cdot}$ on $(1, Z_i, c_i, Z_ic_i)$; finally report the heteroskedasticity-robust standard error $\hat{se}_{HW}(\hat{\tau}_T^{\mathrm{adj}})$.

We have introduced four estimators with the subscript 'I' or 'T' denoting regression based on individual data or cluster totals, and the superscript 'adj' denoting covariate adjustment. For regressions based on individual data, we use the subscript 'LZ' to denote cluster-robust standard error since the treatment assignment is at the cluster level. For regressions based on scaled cluster totals, we use the subscript 'HW' to denote heteroskedasticity-robust standard error since the non-robust standard errors can be inconsistent for the true asymptotic standard errors even in the special case with $n_i = 1$, as pointed out by Freedman (2008a, b) and Lin (2013).

The above four estimators all have reasoned justifications. However, their design-based statistical properties are not entirely clear. First, are these estimators consistent for τ when the linear models are misspecified? Analysing the regression estimators based on individual data is a novel contribution. Even though Lin (2013) and Li and Ding (2017) suggested the consistency of the regression estimators based on cluster totals, the previous theory does not directly cover the case with unequal cluster sizes, especially when some cluster sizes diverge. Second, cluster-robust standard errors are motivated by analysing correlated data under certain modelling assumptions (Cameron & Miller, 2015; Liang & Zeger, 1986), so we must demonstrate that they are also applicable in the design-based inference without those modelling assumptions. The next section will answer these questions, and the following section will extend the theory to a general estimand.

3 | ASYMPTOTICS FOR THE REGRESSION-BASED ESTIMATORS

3.1 | Basic setup for the asymptotic analysis

To derive the asymptotic properties of the four estimators under cluster-randomized experiments, we embed the N units into a sequence of finite populations with clustering structure and an increasing number of clusters $M \to \infty$. All quantities depend on M, but we suppress the dependence in notation for simplicity. Li and Ding (2017) gave a review on finite-population asymptotics, and Middleton and Aronow (2015) gave more details on asymptotics in cluster-randomized experiments. We need to impose some regularity conditions to derive limiting theorems for cluster-randomized experiments.

Assumption 1 The proportion of treated clusters e has a limit in $(0, 1)$, and the dimensions (p_x, p_c) of covariates (x_{ij}, c_i) do not depend on M.

Assumption 1 requires that the numbers of treated and control clusters are not too imbalanced, and restricts the discussion to covariates with a fixed dimension. With some technical modifications, we can extend the results to the case with diverging (p_x, p_c) using the techniques in Bloniarz et al. (2016) and Lei and Ding (2021).

Assumption 2 The potential outcomes satisfy $N^{-1} \sum_{ij} Y_{ij}(z)^4 = O(1)$ for $z = 0, 1$. The covariates satisfy $N^{-1} \sum_{ij} \|x_{ij}\|_\infty^4 = O(1)$, $M^{-1} \sum_{i=1}^M \|c_i\|_\infty^2 = O(1)$ and $M^{-1} max_{1 \le i \le M} \|c_i\|_\infty^2 = o(1)$.

Assumption 2 restricts the moments of the potential outcomes and covariates. We impose stronger moment conditions on unit-level variables, for example, $Y_{ij}(z)$, than on the cluster-level variables to ensure that the scaled cluster totals $\tilde{Y}_{i\cdot}(z) = \sum_{j=1}^{n_i} Y_{ij}(z) M/N$ are well-behaved even when some cluster sizes are much larger than others.

The asymptotic properties depend crucially on

$$\omega_i = n_i/N, \quad \Omega = \max_{1 \le i \le M} \omega_i, \quad \tilde{\omega}_i = \omega_i M = n_i/(N/M),$$

the proportion of units in cluster i, its maximum value and its scaled value measuring the size of cluster i divided by the average cluster size, respectively. When all clusters have equal sizes, $\omega_i = \Omega = 1/M$.

When some clusters are much larger than others, Ω can grow faster than $1/M$. In some extreme cases, Ω can have order $O(1)$. The relative value $\tilde{\omega}_i$ has mean $M^{-1} \sum_{i=1}^M \tilde{\omega}_i = 1$ but can have diverging higher moments when some clusters have extreme sizes. With this notation, we can write $N^{-1} \sum_{ij} Y_{ij}(z)^4 = M^{-1} \sum_{i=1}^M \tilde{\omega}_i \mu_{4i}$ where $\mu_{4i} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}(z)^4$ is the mean of $Y_{ij}(z)^4$ within cluster i. So the condition $N^{-1} \sum_{ij} Y_{ij}(z)^4 = O(1)$ in Assumption 2 becomes more interpretable. In particular, if we further assume that the $(\tilde{\omega}_i, \mu_{4i})_{i=1}^M$ are independent and identically distributed draws from a population, then Assumption 2 requires that the relative cluster size and the cluster mean of $Y_{ij}(z)^4$ has a finite covariance. Other conditions in Assumption 2 have similar interpretations.

To express the asymptotic distributions, we introduce additional notation. Define the centred potential outcome as $\varepsilon_{ij}(z) = Y_{ij}(z) - \overline{Y}(z)$. Similar to $\tilde{Y}_{i\cdot}(z)$, we can define other scaled cluster totals, for example, $\tilde{\varepsilon}_{i\cdot}(z) = \sum_{j=1}^{n_i} \varepsilon_{ij}(z) M/N = \tilde{Y}_{i\cdot}(z) - n_i \overline{Y}(z) M/N$ for the centred potential outcome and $\tilde{x}_{i\cdot} = \sum_{j=1}^{n_i} x_{ij} M/N$ for covariates. Let $\overline{Y}_{i\cdot}(z) = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}(z)$ and $\overline{x}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$ be the cluster averages of the potential outcomes and covariates, respectively. Let $\{\mathcal{E}_i(1), \mathcal{E}_i(0)\}_{i=1}^M$ be the cluster-level potential outcomes with zero averages over clusters. The $\mathcal{E}_i(z)$ can be $\tilde{\varepsilon}_{i\cdot}(z)$ or other transformed cluster-level potential outcomes. Motivated by Neyman (1923), we define

$$V_c\{\mathcal{E}_i(z)\} = \frac{1}{M} \sum_{i=1}^M \left\{ \frac{\mathcal{E}_i(1)^2}{e} + \frac{\mathcal{E}_i(0)^2}{1-e} \right\},$$

$$V\{\mathcal{E}_i(z)\} = V_c\{\mathcal{E}_i(z)\} - \frac{1}{M} \sum_{i=1}^M \left\{ \mathcal{E}_i(1) - \mathcal{E}_i(0) \right\}^2$$

as the variances of the difference in means of $\{\mathcal{E}_i(1), \mathcal{E}_i(0)\}_{i=1}^M$, multiplied by M, in completely randomized experiments with and without the constant treatment effect assumption.

3.2 | Regression estimators based on individual-level data

Define the sample versions of the centred potential outcomes as $\hat{\varepsilon}_{ij} = Y_{ij} - \overline{Y}_{\mathcal{T}}$ for treated units and $\hat{\varepsilon}_{ij} = Y_{ij} - \overline{Y}_C$ for control units. We have the following theorem.

Theorem 1 *Let* $\text{se}^2(\hat{\tau}_I) = V\{\tilde{\varepsilon}_{i\cdot}(z)\}/M$. *Under Assumptions 1 and 2, if* $\Omega = o(1)$, *then* $\hat{\tau}_I = \tau + o_{\mathbb{P}}(1)$; *if* $\Omega = o(M^{-2/3})$ *and* $V\{\tilde{\varepsilon}_{i\cdot}(z)\} \not\to 0$, *then* $(\hat{\tau}_I - \tau)/\text{se}(\hat{\tau}_I) \rightsquigarrow \mathcal{N}(0, 1)$ *and*

$$M \times \hat{\text{se}}_{LZ}^2(\hat{\tau}_I) = \frac{M}{n_{\mathcal{T}}^2} \sum_{i=1}^M Z_i \left(\sum_{j=1}^{n_i} \hat{\varepsilon}_{ij} \right)^2 + \frac{M}{n_C^2} \sum_{i=1}^M (1 - Z_i) \left(\sum_{j=1}^{n_i} \hat{\varepsilon}_{ij} \right)^2 = V_c\{\% \hat{\varepsilon}_{i\cdot}(z)\} + o_{\mathbb{P}}(1).$$

The condition on Ω is stronger for asymptotic Normality than for consistency, and both hold if all cluster sizes have the same order $O(N/M)$. Theorem 1 demonstrates that the cluster-robust standard error $\hat{\text{se}}_{LZ}(\hat{\tau}_I)$ is a conservative estimator for the true asymptotic standard error due to the non-negative difference between $V_c(\cdot)$ and $V(\cdot)$, which equals zero when $\tilde{\varepsilon}_{i\cdot}(1) - \tilde{\varepsilon}_{i\cdot}(0) = 0$ or $\overline{Y}_{i\cdot}(1) - \overline{Y}_{i\cdot}(0) = \tau$ for all i. In contrast, without accounting for clustering, the heteroskedasticity-robust standard error $\hat{\text{se}}_{HW}(\hat{\tau}_I)$ can be anti-conservative because Angrist and Pischke (2008 chapter 8) implies

$$N \times \widehat{\mathrm{se}}^2_{\mathrm{HW}}(\widehat{\tau}_{\mathrm{I}}) = \frac{N}{n_{\mathcal{T}}^2} \sum_{ij \in \mathcal{T}} \widehat{\bar{\varepsilon}}^2_{ij} + \frac{N}{n_C^2} \sum_{ij \in C} \widehat{\bar{\varepsilon}}^2_{ij} = \frac{1}{N} \sum_{ij} \left\{ \frac{\varepsilon_{ij}(1)^2}{e} + \frac{\varepsilon_{ij}(0)^2}{1-e} \right\} + o_{\mathbb{P}}(1).$$

If $n_i = 1$ for all clusters, then $\widehat{\mathrm{se}}^2_{\mathrm{HW}}(\widehat{\tau}_{\mathrm{I}})$ coincides with $\widehat{\mathrm{se}}^2_{\mathrm{LZ}}(\widehat{\tau}_{\mathrm{I}})$. If $n_i > 1$, then $\widehat{\mathrm{se}}^2_{\mathrm{HW}}(\widehat{\tau}_{\mathrm{I}})$ is often much smaller than $\widehat{\mathrm{se}}^2_{\mathrm{LZ}}(\widehat{\tau}_{\mathrm{I}})$. To fix ideas, we consider a simple case with constant treatment effect $Y_{ij}(1) - Y_{ij}(0) = \tau$. From Theorem 1, $\widehat{\mathrm{se}}^2_{\mathrm{LZ}}(\widehat{\tau}_{\mathrm{I}})$ is consistent for the true asymptotic standard error, but we can verify that $\widehat{\mathrm{se}}^2_{\mathrm{LZ}}(\widehat{\tau}_{\mathrm{I}})/\widehat{\mathrm{se}}^2_{\mathrm{HW}}(\widehat{\tau}_{\mathrm{I}})$ has probability limit close to

$$1 + \frac{\sum_{i=1}^M \sum_{(j,k):j\neq k} \varepsilon_{ij}(0)\varepsilon_{ik}(0)}{\sum_{ij} \varepsilon_{ij}(0)^2},$$

which is larger than 1 if the centred control potential outcomes are positively correlated within clusters. The magnitude of the ratio $\widehat{\mathrm{se}}^2_{\mathrm{LZ}}(\widehat{\tau}_{\mathrm{I}})/\widehat{\mathrm{se}}^2_{\mathrm{HW}}(\widehat{\tau}_{\mathrm{I}})$ is closely related to the intraclass correlation coefficient in cluster sampling (Lohr, 2019) and cluster-randomized experiments (Donner & Klar, 2000). Therefore, without accounting for clustering, we may have invalid inference even with large samples, which is well known since Cornfield (1978).

Schochet et al. (2021) focus on a slightly more general design and discussed the design-based properties of the cluster-robust standard error from an individual-level regression without covariates. Their result hints at Theorem 1, and we present a slightly more formal result under weaker regularity conditions on the sizes of the clusters. We further extend the theory to the regressions with covariates. The asymptotic properties of the covariate-adjusted regression estimator based on individual data rely on more regularity conditions on the covariates and potential outcomes.

Assumption 3 For $z = 0, 1$, $N^{-1} \sum_{ij} x_{ij}\varepsilon_{ij}(z) = M^{-1} \sum_{i=1}^M \tilde{\omega}_i\{n_i^{-1} \sum_{j=1}^{n_i} x_{ij}\varepsilon_{ij}(z)\}$ converges to a finite vector, and $N^{-1} \sum_{ij} x_{ij}x_{ij}^{\mathrm{T}} = M^{-1} \sum_{i=1}^M \tilde{\omega}_i\{n_i^{-1} \sum_{j=1}^{n_i} x_{ij}x_{ij}^{\mathrm{T}}\}$ converges to a finite and invertible matrix.

Let $Q_{\mathrm{I}}(z)$ be the coefficient of x_{ij} in the OLS fit of $\varepsilon_{ij}(z)$ on x_{ij} with individual data ($z = 0, 1$). This theoretical quantity depends on all potential outcomes and is useful for expressing the asymptotic distributions of the estimators based on regressions with individual data.

Theorem 2 *Define $r_i(z) = \tilde{\varepsilon}_{i\cdot}(z) - \tilde{x}_{i\cdot}^{\mathrm{T}} Q_{\mathrm{I}}(z)$ and $\mathrm{se}^2(\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}}) = V\{r_i(z)\}/M$. Under Assumptions 1–3, if $\Omega = o(1)$, then $\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}} = \tau + o_{\mathbb{P}}(1)$; if $\Omega = o(M^{-2/3})$ and $V\{r_i(z)\} \nrightarrow 0$, then $(\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}} - \tau)/\mathrm{se}(\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}}) \rightsquigarrow \mathcal{N}(0, 1)$ and $M \times \widehat{\mathrm{se}}^2_{\mathrm{LZ}}(\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}}) = V_c\{r_i(z)\} + o_{\mathbb{P}}(1).$*

Theorem 2 demonstrates that the regression estimator $\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}}$ adjusted for covariates x_{ij} is consistent, and the associated cluster-robust standard error is a conservative estimator for the true asymptotic standard error. However, the form of $r_i(z)$ in Theorem 2 is $\tilde{\varepsilon}_{i\cdot}(z) - \tilde{x}_{i\cdot}^{\mathrm{T}} Q$, which suggests that an optimal choice of Q should be the coefficient of $\tilde{x}_{i\cdot}$ in the OLS fit of $\tilde{\varepsilon}_{i\cdot}(z)$ on $\tilde{x}_{i\cdot}$ based on cluster totals, rather than $Q_{\mathrm{I}}(z)$ from the individual-level regression. Consequently, it is possible that $\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}}$ is even less efficient than $\widehat{\tau}_{\mathrm{I}}$. We give a numeric example in Section 6.3. In the special case with $n_i = N/M$ and x_{ij} not varying with j, $Q_{\mathrm{I}}(z)$ equals the coefficient of $\tilde{x}_{i\cdot}$ in the OLS fit of $\tilde{\varepsilon}_{i\cdot}(z)$ on $\tilde{x}_{i\cdot}$, so $\widehat{\tau}_{\mathrm{I}}^{\mathrm{adj}}$ has smaller asymptotic variance than $\widehat{\tau}_{\mathrm{I}}$. In particular, when $n_i = 1$ for all clusters, Theorem 2 recovers Lin (2013) and Li and Ding (2017).

Lin (2013) popularized the OLS estimator with treatment–covariates interactions, but it is common in social science to use $\widehat{\tau}_{\mathrm{I}}^{\mathrm{ancova}}$, the coefficient of Z_{ij} in the OLS fit of Y_{ij} on $(1, Z_{ij}, x_{ij})$ without

interactions (e.g. Graubard & Korn, 1994; Green & Vavreck, 2008; Guiteras et al., 2015; Schochet, 2013; Schochet et al., 2021). Its properties are similar to $\widehat{\tau}_I^{\mathrm{adj}}$ in Theorem 2.

Corollary 1 *Define* $r_i(z) = \tilde{\varepsilon}_{i\cdot}(z) - \tilde{x}_{i\cdot}^{\mathrm{T}}\{eQ_I(1) + (1-e)Q_I(0)\}$ *and* $\mathrm{se}^2(\widehat{\tau}_I^{\mathrm{ancova}}) = V\{r_i(z)\}/M$. *Under Assumptions 1–3, if* $\Omega = o(1)$ *then* $\widehat{\tau}_I^{\mathrm{ancova}} = \tau + o_{\mathbb{P}}(1)$; *if* $\Omega = o(M^{-2/3})$ *and* $V\{r_i(z)\} \nrightarrow 0$, *then* $(\widehat{\tau}_I^{\mathrm{ancova}} - \tau)/\mathrm{se}(\widehat{\tau}_I^{\mathrm{ancova}}) \rightsquigarrow \mathcal{N}(0, 1)$ *and* $M \times \widehat{\mathrm{se}}_{LZ}^2(\widehat{\tau}_I^{\mathrm{ancova}}) = V_c\{r_i(z)\} + o_{\mathbb{P}}(1)$.

Corollary 1 demonstrates that $\widehat{\tau}_I^{\mathrm{ancova}}$ is consistent and the associated cluster-robust standard error is conservative for the true asymptotic standard error. Similar to Freedman (2008b), $\widehat{\tau}_I^{\mathrm{ancova}}$ can be even less efficient than $\widehat{\tau}_I$; different from Lin (2013), $\widehat{\tau}_I^{\mathrm{adj}}$ does not necessarily improve $\widehat{\tau}_I^{\mathrm{ancova}}$ in terms of asymptotic efficiency. Because $\widehat{\tau}_I^{\mathrm{ancova}}$ is suboptimal even in the case with $n_i = 1$ (Freedman, 2008b; Lin, 2013), we do not recommend it unless we believe that there is little treatment heterogeneity explained by x_{ij}. Schochet et al. (2021) discuss a general version of $\widehat{\tau}_I^{\mathrm{adj}}$ in cluster-randomized experiments with a blocking factor. They establish the central limit theorem but do not give a proof for the validity of the cluster-robust standard error. We fill the gap by showing the asymptotic conservativeness of the cluster-robust standard error. Moreover, Schochet et al. (2021) do not discuss $\widehat{\tau}_I^{\mathrm{adj}}$ that allows for treatment effect heterogeneity with respect to the covariates.

3.3 | Regression estimators based on scaled cluster totals

Recall that $\tau = M^{-1} \sum_{i=1}^{M} \{\tilde{Y}_{i\cdot}(1) - \tilde{Y}_{i\cdot}(0)\}$ is the average treatment effect on the scaled cluster totals $\tilde{Y}_{i\cdot}$, and a cluster-randomized experiment is essentially a completely randomized experiment on clusters. So we expect previous results on the regression analysis of completely randomized experiments to be applicable if we view the scaled cluster total as a new outcome. Although conceptually straightforward, we must address the technical difficulty that the cluster sizes can vary by extending the central limit theorems under completely randomized experiments. We show the asymptotic properties of $\widehat{\tau}_T$ and $\widehat{\tau}_T^{\mathrm{adj}}$, allowing some cluster sizes to be much larger than others. The following theorems extend Lin (2013) and Li and Ding (2017).

Theorem 3 *Let* $r_i(z) = \tilde{Y}_{i\cdot}(z) - \overline{Y}(z)$ *and* $\mathrm{se}^2(\widehat{\tau}_T) = V\{r_i(z)\}/M$. *Under Assumptions 1 and 2, if* $\Omega = o(1)$, *then* $\widehat{\tau}_T = \tau + o_{\mathbb{P}}(1)$; *if* $\Omega = o(M^{-2/3})$ *and* $V\{r_i(z)\} \nrightarrow 0$, *then* $(\widehat{\tau}_T - \tau)/\mathrm{se}(\widehat{\tau}_T) \rightsquigarrow \mathcal{N}(0, 1)$ *and* $M \times \widehat{\mathrm{se}}_{HW}^2(\widehat{\tau}_T) = V_c\{r_i(z)\} + o_{\mathbb{P}}(1)$.

Assumption 4 *For* $z = 0, 1$, $M^{-1} \sum_{i=1}^{M} c_i \tilde{Y}_{i\cdot}(z) = M^{-1} \sum_{i=1}^{M} \tilde{\omega}_i c_i \tilde{Y}_{i\cdot}(z)$ *converges to a finite vector, and* $M^{-1} \sum_{i=1}^{M} c_i c_i^{\mathrm{T}}$ *converges to a finite and invertible matrix.*

Theorem 4 *Let* $r_i(z)$ *be the residual from the OLS fit of* $\tilde{Y}_{i\cdot}(z)$ *on* $(1, c_i)$ *and define* $\mathrm{se}^2(\widehat{\tau}_T^{\mathrm{adj}}) = V\{r_i(z)\}/M$. *Under Assumptions 1, 2 and 4, if , then* $\widehat{\tau}_T^{\mathrm{adj}} = \tau + o_{\mathbb{P}}(1)$; *if* $\Omega = o(M^{-1})$ *and* $V\{r_i(z)\} \nrightarrow 0$, *then* $(\widehat{\tau}_T^{\mathrm{adj}} - \tau)/\mathrm{se}(\widehat{\tau}_T^{\mathrm{adj}}) \rightsquigarrow \mathcal{N}(0, 1)$ *and* $M \times \widehat{\mathrm{se}}_{HW}^2(\widehat{\tau}_T^{\mathrm{adj}}) = V_c\{r_i(z)\} + o_{\mathbb{P}}(1)$.

Comparing Theorems 3 and 4, covariate adjustment always improves the asymptotic efficiency because $\mathrm{se}^2(\widehat{\tau}_T^{\mathrm{adj}}) \leq \mathrm{se}^2(\widehat{\tau}_T)$ from the regressions based on cluster totals. We omit the detailed discussion of $\widehat{\tau}_T^{\mathrm{ancova}}$, the coefficient of Z_i in the OLS fit of $\tilde{Y}_{i\cdot}$ on $(1, Z_i, c_i)$. Similar to $\widehat{\tau}_T$ and $\widehat{\tau}_T^{\mathrm{adj}}$, it is also consistent and asymptotically Normal for τ, and the associated heteroskedasticity-robust standard error is conservative for the true asymptotic standard error. However, it may not even improve the efficiency of $\widehat{\tau}_T$, and it is suboptimal compared to $\widehat{\tau}_T^{\mathrm{adj}}$.

3.4 | Comparison

We first compare $\hat{\tau}_I$ and $\hat{\tau}_T$: $\hat{\tau}_T$ is the Horvitz–Thompson estimator and thus it is unbiased (Middleton & Aronow, 2015); $\hat{\tau}_I$ is the Hájek estimator and thus it has finite-sample bias unless the clusters have equal sizes. In some extreme cases, $\hat{\tau}_T$ has smaller mean squared error than $\hat{\tau}_I$. For example, we have the following finite-sample result. The conditions are very strong and unlikely to hold, but we keep it here for mathematical interest.

Proposition 1 *If $e = 1/2$ and $\overline{Y}(z) = 0$ for $z = 0, 1$, then $E(\hat{\tau}_T - \tau)^2 \le E(\hat{\tau}_I - \tau)^2$.*

Nevertheless, $\hat{\tau}_I$ is consistent and its bias diminishes with large M. In most cases, $\hat{\tau}_T$ has larger asymptotic variance than $\hat{\tau}_I$ because the residual $\tilde{Y}_{i.}(z) - n_i \overline{Y}(z)M/N$ in Theorem 1 adjusts for the cluster size while the residual $\tilde{Y}_{i.}(z) - \overline{Y}(z)$ in Theorem 3 does not. Moreover, it is well known (Fuller, 2009; Middleton & Aronow, 2015) that $\hat{\tau}_T$ is not invariant to the location shift of the outcomes, that is, if we transform the potential outcomes to $b + Y_{ij}(z)$ for $z = 0, 1$, then $\hat{\tau}_T$ will change but $\hat{\tau}_I$ will not. We do not recommend using $\hat{\tau}_T$.

We can fix the problems of $\hat{\tau}_T$ by covariate adjustment with c_i including n_i. Below we compare various regression-adjusted estimators. We write particular covariates in parentheses after an estimator to emphasize the change of the default covariates. For instance, $\hat{\tau}_T^{\mathrm{adj}}(n_i, \tilde{x}_{i.})$ denotes the regression estimator based on cluster totals adjusted for centred c_i with $c_i = (n_i, \tilde{x}_{i.})$.

Corollary 2 *The asymptotic variances satisfy*

$$\mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i, \tilde{x}_{i.})\} \le \mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i)\} \le \mathrm{se}^2(\hat{\tau}_I), \quad \mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i, \tilde{x}_{i.})\} \le \mathrm{se}^2(\hat{\tau}_I^{\mathrm{adj}}).$$

These inequalities hold asymptotically if we replace the standard errors by the estimated ones.

Inequality $\mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i, \tilde{x}_{i.})\} \le \mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i)\}$ follows because Lin (2013)'s regression adjustment will always improve the asymptotic variance. Inequality $\mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i)\} \le \mathrm{se}^2(\hat{\tau}_I)$ follows from the fact that the residual sum of squares from the OLS fit of $\sum_{j=1}^{n_i} Y_{ij}(z)M/N - \overline{Y}(z)$ on $n_i M/N - 1$ is smaller than $\sum_{i=1}^{M} \tilde{\varepsilon}_{i.}(z)^2$. So adjusting for n_i in the regression based on cluster totals improves the simple difference-in-means, and adjusting for cluster totals of additional covariates leads to more improvement asymptotically. Middleton and Aronow (2015) discussed the role of n_i and $\tilde{x}_{i.}$ in improving efficiency, and our estimator has the smallest asymptotic variance within the class of linearly adjusted estimators which follows from the optimality result in completely randomized experiments (Li & Ding, 2017).

Inequality $\mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i, \tilde{x}_{i.})\} \le \mathrm{se}^2(\hat{\tau}_I^{\mathrm{adj}})$ follows from the fact that $\sum_{i=1}^{M} \{\tilde{r}_{i.}(z)\}^2 = \sum_{i=1}^{M} \{\tilde{\varepsilon}_{i.}(z) - \tilde{x}_{i.}^{\mathrm{T}} Q_I(z)\}^2$ in Theorem 2 is larger than the residual sum of squares from the OLS fit of $\tilde{\varepsilon}_{i.}(z)$ on $\tilde{x}_{i.}$, which is larger than that from the OLS fit of $\sum_{j=1}^{n_i} Y_{ij}(z)M/N - \overline{Y}(z)$ on $(n_i M/N - 1, \tilde{x}_{i.})$. So adjusting for $(n_i, \tilde{x}_{i.})$ in the regression based on cluster totals improves the regression adjustment of x_{ij} based on the individual data. It demonstrates that regressions based on cluster totals dominate regressions based on individual data asymptotically. With a special choice of covariates $x_{ij} = \overline{x}_{i.}$ in individual-level regression, Corollary 2 implies $\mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i, \tilde{x}_{i.})\} \le \mathrm{se}^2\{\hat{\tau}_I^{\mathrm{adj}}(\overline{x}_{i.})\}$. One may wonder whether regressions based on cluster total can still dominate regressions based on individual data if the latter already includes (n_i, x_{ij}) as covariates. The answer is yes because $\mathrm{se}^2\{\hat{\tau}_T^{\mathrm{adj}}(n_i, n_i^2, \tilde{x}_{i.})\} \le \mathrm{se}^2\{\hat{\tau}_I^{\mathrm{adj}}(n_i, x_{ij})\}$ follows from the second inequality in Corollary 2.

However, we must address a technical issue related to the regression estimators adjusted for n_i. When c_i includes n_i, Assumption 4 imposes moment conditions on n_i, but n_i itself can diverge.

Fortunately, by the property of OLS, $\hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(n_i) = \hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(\tilde{\omega}_i)$ and $\hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(n_i, \tilde{x}_{i\cdot}) = \hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(\tilde{\omega}_i, \tilde{x}_{i\cdot})$, which remain the same up to a scale transformation of the covariates. Therefore, Assumption 4 imposes moment conditions on $\tilde{\omega}_i$, for example, $M^{-1} \sum_{i=1}^{M} \tilde{\omega}_i(\tilde{\omega}_i - 1)\overline{Y}_{i\cdot}(z)$ converges to a finite vector. This technical issue has important implications in practice. It requires at least two moments of the $\tilde{\omega}_i$'s, which is stronger than the conditions required for the estimators based on the individual-level regressions. Therefore, the asymptotic efficiency of the estimators based on cluster totals comes at a price of the robustness with respect to the heavy-tailedness of the cluster sizes. In some extreme cases, the cluster sizes vary with some outliers, which can violate the regularity conditions for the estimators based on cluster totals. We observe this phenomenon in simulation studies in Section 6.4.

4 | A GENERAL WEIGHTED ESTIMAND AND ITS ESTIMATORS

4.1 | Choice of the estimand

In addition to the regression estimators based on individual data and cluster totals, it is also common to use regression estimators based on cluster averages (e.g. Donner & Klar, 2000; Green & Vavreck, 2008; Imai et al., 2009). However, Middleton (2008) argued against the use of them due to their biases for estimating τ even with large M. In contrast, our view differs from Middleton (2008). First, we show that we can easily remove the biases by using WLS with weights proportional to cluster sizes. Second, we argue that regression estimators based on cluster averages typically target different estimands beyond τ, which are also of interest in practice. Third, we demonstrate that the heteroskedasticity-robust standard errors from these regressions are conservative estimators for the asymptotic standard errors.

To unify the discussion, we introduce a weighted estimand. For cluster i, the average treatment effect is $\tau_i = \overline{Y}_{i\cdot}(1) - \overline{Y}_{i\cdot}(0)$, and the average observed outcome is $\overline{Y}_{i\cdot} = Z_i\overline{Y}_{i\cdot}(1) + (1 - Z_i)\overline{Y}_{i\cdot}(0)$. For weights $(\pi_i)_{i=1}^{M}$ such that $\pi_i > 0$ and $\sum_{i=1}^{M} \pi_i = 1$, define

$$\tau_\pi = \sum_{i=1}^{M} \pi_i\tau_i = \sum_{i=1}^{M} \pi_i\{\overline{Y}_{i\cdot}(1) - \overline{Y}_{i\cdot}(0)\} = \overline{Y}_\pi(1) - \overline{Y}_\pi(0),$$

as a weighted average treatment effect where $\overline{Y}_\pi(z) = \sum_{i=1}^{M} \pi_i\overline{Y}_{i\cdot}(z)$ for $z = 0, 1$. If $\pi_i = \omega_i$, then τ_π recovers τ; if $\pi_i = 1/M$, then $\tau_\pi = \overline{\tau} = M^{-1} \sum_{i=1}^{M} \tau_i$ imposes equal weights on the average effects within clusters regardless of their sizes. The estimand τ views N units as the population, and the estimand $\overline{\tau}$ views M clusters as the population. Both can be of policy relevance. For example, Donner and Klar (2000 chapter 1.5) discussed a smoking cessation trial with randomization over communities, in which τ is of interest if the aim is to influence individuals to cease smoking, and $\overline{\tau}$ is of interest if the aim is to lower the rate of smoking in a community. In some applications, the units within clusters are sampled from larger clusters, and we can use π_i to reflect the sampling weights to recover the effect on a larger population. In general, τ_π includes a class of estimands that are of interest in different scenarios.

4.2 | Four estimators

Because $\tau_\pi = \sum_{i=1}^{M} \pi_i\tau_i$ is a weighted estimand, it is natural to use weighted regressions to estimate it. We can use $\hat{\tau}_{\mathrm{A}\pi}$, the coefficient of Z_i in the WLS fit of $\overline{Y}_{i\cdot}$ on $(1, Z_i)$ with weights π_i, and report

$\widehat{se}_{HW}(\widehat{\tau}_{A\pi})$. With cluster-level covariates c_i, we can first centre them at $\overline{c}_\pi = \sum_{i=1}^{M} \pi_i c_i = 0$, then obtain $\widehat{\tau}_{A\pi}^{\text{adj}}$, the coefficient of Z_i in the WLS fit of $\overline{Y}_{i\cdot}$ on $\{1, Z_i, c_i - \overline{c}_\pi, Z_i(c_i - \overline{c}_\pi)\}$ with weights π_i, and finally report $\widehat{se}_{HW}(\widehat{\tau}_{A\pi}^{\text{adj}})$.

On the other hand,

$$\tau_\pi = M^{-1} \sum_{i=1}^{M} \{M\pi_i \overline{Y}_{i\cdot}(1) - M\pi_i \overline{Y}_{i\cdot}(0)\}$$

equals the average treatment effect on the weighted outcome $M\pi_i \overline{Y}_{i\cdot}$ on the cluster level. This representation motivates to estimate τ_π using OLS fit based on the weighted outcome. Basse and Feller (2018) suggested this approach in a related setting. We can use $\widehat{\tau}_{\pi A}$, the coefficient of Z_i in the OLS fit of $M\pi_i \overline{Y}_{i\cdot}$ on $(1, Z_i)$, and report $\widehat{se}_{HW}(\widehat{\tau}_{\pi A})$. With cluster-level covariates c_i, we can first centre them at $\overline{c} = M^{-1} \sum_{i=1}^{M} c_i = 0$, then obtain $\widehat{\tau}_{\pi A}^{\text{adj}}$, the coefficient of Z_i in the OLS fit of $M\pi_i \overline{Y}_{i\cdot}$ on $(1, Z_i, c_i, Z_i c_i)$, and finally report $\widehat{se}_{HW}(\widehat{\tau}_{\pi A}^{\text{adj}})$.

Analogous to Section 2.3, we have introduced four estimators for τ_π, all associated with heteroskedasticity-robust standard errors. The orders of 'A' and 'π' in the subscript reflect the computational processes, for example, to compute $\widehat{\tau}_{A\pi}^{\text{adj}}$, we first compute the cluster averages and then run WLS, while to compute $\widehat{\tau}_{\pi A}^{\text{adj}}$, we first weight the cluster averages and then run OLS. Moreover, we centre covariates c_i at different means in these two types of estimators. In the special case with $\pi_i = 1/M$, they are identical with $\widehat{\tau}_{A\pi} = \widehat{\tau}_{\pi A}$ and $\widehat{\tau}_{A\pi}^{\text{adj}} = \widehat{\tau}_{\pi A}^{\text{adj}}$, so we use $\widehat{\tau}_A$ and $\widehat{\tau}_A^{\text{adj}}$ to denote both, omitting π.

The remaining parts of this section will derive the asymptotic properties of the above four estimators and compare their efficiencies. These asymptotic analyses will help to unify Sections 3 and 4.

4.3 | Weighted regression estimators based on cluster averages

Assumption 5 The weights satisfy $max_{1 \le i \le M} \pi_i = o(1)$, and the potential outcomes satisfy $\sum_{i=1}^{M} \pi_i \overline{Y}_{i\cdot}(z)^4 = O(1)$ for $z = 0, 1$.

Theorem 5 Let $r_i(z) = M\pi_i\{\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)\}$ and $se^2(\widehat{\tau}_{A\pi}) = V\{r_i(z)\}/M$. Under Assumptions 1 and 5, $\widehat{\tau}_{A\pi} = \tau_\pi + o_{\mathbb{P}}(1)$; if further $max_{1 \le i \le M} \pi_i = o(M^{-2/3})$ and $V\{r_i(z)\} \nrightarrow 0$, then $(\widehat{\tau}_{A\pi} - \tau_\pi)/se(\widehat{\tau}_{A\pi}) \rightsquigarrow \mathcal{N}(0, 1)$ and $M \times \widehat{se}_{HW}^2(\widehat{\tau}_{A\pi}) = V_c\{r_i(z)\} + o_{\mathbb{P}}(1)$.

If $\pi_i = 1/M$, then $\widehat{\tau}_{A\pi} = \widehat{\tau}_A$ reduces to the regression estimator based on cluster averages. Middleton (2008) showed that it is inconsistent for estimating τ, but Theorem 5 shows that it is consistent for estimating $\overline{\tau} = M^{-1} \sum_{i=1}^{M} \tau_i$. We have $\overline{\tau} = \tau + o(1)$ if and only if the finite-sample covariance between $\tilde{\omega}_i$ and τ_i has order $o(1)$. There are at least two special cases in which $\overline{\tau} = \tau + o(1)$ holds. First, if $n_i = N/M$, then $\overline{\tau} = \tau$ and $\widehat{\tau}_A = \widehat{\tau}_T$. Second, if $max_{1 \le i \le M} |\tau_i - \tau| = o(1)$, then $\widehat{\tau}_A = \tau + o_{\mathbb{P}}(1)$, but this condition rules out treatment effect heterogeneity across clusters.

Another important special case is $\pi_i = \omega_i = n_i/N$, in which we can consistently estimate τ by $\widehat{\tau}_{A\omega}$. In fact, in this case, $\widehat{\tau}_{A\omega} = \widehat{\tau}_I$. We will revisit this case in the next section.

Assumption 6 For $z = 0, 1$, $\sum_{i=1}^{M} \pi_i(c_i - \overline{c}_\pi)\overline{Y}_{i\cdot}(z)$ converges to a finite vector, $\sum_{i=1}^{M} \pi_i(c_i - \overline{c}_\pi)(c_i - \overline{c}_\pi)^{\text{T}}$ converges to a finite and invertible matrix, and $\sum_{i=1}^{M} \pi_i \|c_i\|_\infty^4 = O(1)$.

Let $Q_{A\pi}(z)$ be the coefficient of c_i in the WLS fit of $\overline{Y}_{i\cdot}(z)$ on $(1, c_i)$ with weights π_i ($z = 0, 1$). It depends on all potential outcomes and simplifies the discussion below.

Theorem 6 *Define* $r_i(z) = M\pi_i\{\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z) - (c_i - \overline{c}_\pi)^{\mathrm{T}} Q_{A\pi}(z)\}$ *and* $\mathrm{se}^2(\widehat{\tau}_{A\pi}^{\mathrm{adj}}) = V\{r_i(z)\}/M$. *Under Assumptions 1, 5 and 6,* $\widehat{\tau}_{A\pi}^{\mathrm{adj}} = \tau_\pi + o_{\mathbb{P}}(1)$; *if further* $\max_{1 \leq i \leq M} \pi_i = o(M^{-2/3})$ *and* $V\{r_i(z)\} \not\to 0$, *then* $(\widehat{\tau}_{A\pi}^{\mathrm{adj}} - \tau_\pi)/\mathrm{se}(\widehat{\tau}_{A\pi}^{\mathrm{adj}}) \rightsquigarrow \mathcal{N}(0, 1)$ *and* $M \times \widehat{\mathrm{se}}_{\mathrm{HW}}^2(\widehat{\tau}_{A\pi}^{\mathrm{adj}}) = V_{\mathrm{c}}\{r_i(z)\} + o_{\mathbb{P}}(1)$.

The form of $r_i(z)$ in Theorem 6 is $M\pi_i\{\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)\} - M\pi_i(c_i - \overline{c}_\pi)^{\mathrm{T}}Q$, which suggests that an optimal choice of Q is the coefficient of $\pi_i(c_i - \overline{c}_\pi)$ in the OLS fit of $\pi_i\{\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)\}$ on $\pi_i(c_i - \overline{c}_\pi)$, rather than $Q_{A\pi}(z)$ from the WLS based on cluster averages. Similar to the discussion of Theorem 2, $\widehat{\tau}_{A\pi}^{\mathrm{adj}}$ may not have better efficiency compared to $\widehat{\tau}_{A\pi}$. We will give an example in Section 6.3.

4.4 | Regression estimators based on weighted cluster averages

In the estimators $\widehat{\tau}_{\pi A}$ and $\widehat{\tau}_{\pi A}^{\mathrm{adj}}$, we view $M\pi_i\overline{Y}_{i\cdot}$ as a new outcome and apply Lin (2013)'s results for completely randomized experiments. Again the following theorems extend Lin (2013) and Li and Ding (2017), which include Theorems 3 and 4 as special cases with $\pi_i = \omega_i$.

Theorem 7 *Let* $r_i(z) = M\pi_i\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)$ *and* $\mathrm{se}^2(\widehat{\tau}_{\pi A}) = V\{r_i(z)\}/M$. *Under Assumption 1, if* $\sum_{i=1}^{M} \pi_i^2\overline{Y}_{i\cdot}(z)^2 = o(1)$, *then* $\widehat{\tau}_{\pi A} = \tau_\pi + o_{\mathbb{P}}(1)$; *if further* $\sum_{i=1}^{M} M^2\pi_i^4\overline{Y}_{i\cdot}(z)^4 = o(1)$ *and* $V\{r_i(z)\} \not\to 0$, *then* $(\widehat{\tau}_{\pi A} - \tau_\pi)/\mathrm{se}(\widehat{\tau}_{\pi A}) \rightsquigarrow \mathcal{N}(0, 1)$ *and* $M \times \widehat{\mathrm{se}}_{\mathrm{HW}}^2(\widehat{\tau}_{\pi A}) = V_{\mathrm{c}}\{r_i(z)\} + o_{\mathbb{P}}(1)$.

Assumption 7 *For* $z = 0, 1$, $\sum_{i=1}^{M} \pi_i c_i\overline{Y}_{i\cdot}(z)$ *converges to a finite vector, and* $M^{-1} \sum_{i=1}^{M} c_i c_i^{\mathrm{T}}$ *converges to a finite and invertible matrix.*

Theorem 8 *Let* $r_i(z)$ *be the residual from the OLS fit of* $M\pi_i\overline{Y}_{i\cdot}(z)$ *on* $(1, c_i)$ *and* $\mathrm{se}^2(\widehat{\tau}_{\pi A}^{\mathrm{adj}}) = V\{r_i(z)\}/M$. *Under Assumptions 1, 2 and 7, if* $\sum_{i=1}^{M} \pi_i^2\overline{Y}_{i\cdot}(z)^2 = o(1)$, *then* $\widehat{\tau}_{\pi A}^{\mathrm{adj}} = \tau_\pi + o_{\mathbb{P}}(1)$; *if further* $\sum_{i=1}^{M} M^3\pi_i^4\overline{Y}_{i\cdot}(z)^4 = O(1)$ *and* $V\{r_i(z)\} \not\to 0$, *then* $(\widehat{\tau}_{\pi A}^{\mathrm{adj}} - \tau_\pi)/\mathrm{se}(\widehat{\tau}_{\pi A}^{\mathrm{adj}}) \rightsquigarrow \mathcal{N}(0, 1)$ *and* $M \times \widehat{\mathrm{se}}_{\mathrm{HW}}^2(\widehat{\tau}_{\pi A}^{\mathrm{adj}}) = V_{\mathrm{c}}\{r_i(z)\} + o_{\mathbb{P}}(1)$.

Comparing Theorems 7 and 8, $\mathrm{se}^2(\widehat{\tau}_{\pi A}^{\mathrm{adj}}) \leq \mathrm{se}^2(\widehat{\tau}_{\pi A})$, so covariate adjustment always improves the asymptotic efficiency for estimating τ_π in the regression estimators based on weighted cluster averages.

4.5 | Comparison

Similar to Corollary 2, we can reduce the asymptotic variances of the WLS estimators by the OLS estimators based on weighted cluster averages with appropriately chosen covariates. Again, we write covariates in parentheses to emphasize them if they are different from the default choices.

Corollary 3 *The asymptotic variances satisfy*

$$\mathrm{se}^2\{\widehat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i, \pi_i c_i)\} \leq \mathrm{se}^2\{\widehat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i)\} \leq \mathrm{se}^2(\widehat{\tau}_{A\pi}), \quad \mathrm{se}^2\{\widehat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i, \pi_i c_i)\} \leq \mathrm{se}^2(\widehat{\tau}_{A\pi}^{\mathrm{adj}}).$$

These inequalities hold asymptotically if we replace the standard errors by the estimated ones.

Inequality $\mathrm{se}^2\{\widehat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i, \pi_i c_i)\} \leq \mathrm{se}^2\{\widehat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i)\}$ follows because Lin (2013)'s regression adjustment will always improve the asymptotic variance. Inequality $\mathrm{se}^2\{\widehat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i)\} \leq \mathrm{se}^2(\widehat{\tau}_{A\pi})$ follows from the

fact that the residual sum of squares from the OLS fit of $\pi_i \overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)/M$ on $\pi_i - 1/M$ is smaller than $\sum_{i=1}^{M} [\pi_i \{\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)\}]^2$. Inequality $\mathrm{se}^2\{\hat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i, \pi_i c_i)\} \leq \mathrm{se}^2(\hat{\tau}_{A\pi}^{\mathrm{adj}})$ follows from the fact that $\sum_{i=1}^{M} \{r_i(z)/M\}^2 = \sum_{i=1}^{M} [\pi_i \{\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)\} - \pi_i(c_i - \overline{c}_\pi)^{\mathrm{T}} Q_{A\pi}(z)]^2$ in Theorem 6 is larger than the residual sum of squares from the OLS fit of $\pi_i \{\overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)\}$ on $\pi_i(c_i - \overline{c}_\pi)$, which is larger than that from the OLS fit of $\pi_i \overline{Y}_{i\cdot}(z) - \overline{Y}_\pi(z)/M$ on $(\pi_i - 1/M, \pi_i c_i - \overline{c}_\pi/M)$.

Corollary 3 highlights the importance of including π_i as a cluster-level covariate to improve the asymptotic efficiency. Overall, given estimators $\hat{\tau}_{A\pi}$ and $\hat{\tau}_{A\pi}^{\mathrm{adj}}$, we can always improve their efficiency by using $\hat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i, \pi_i c_i)$.

Schochet et al. (2021) propose an even more general weighted estimand, motivated by the need to adjust for nonresponses in cluster-randomized experiments. Their weights can vary with individuals but our weights only vary with clusters. By allowing for general weights, we can also discuss individual-level regression with weights. This is beyond the scope of this paper.

5 | UNIFICATION AND RECOMMENDATION

So far, we have introduced four estimators for each of τ and τ_π. Table 1 summarizes them and the corresponding standard errors. Table 1 presents them in parallel, and the ones in the each row are equivalent when $\pi_i = \omega_i$, as demonstrated by the proposition below.

Proposition 2 *If $\pi_i = \omega_i$ for all i = 1, ... , M, then*

$$
\begin{aligned}
\hat{\tau}_{A\omega} &= \hat{\tau}_{I}, & \hat{\mathrm{se}}_{\mathrm{HW}}^2(\hat{\tau}_{A\omega}) &= \hat{\mathrm{se}}_{\mathrm{LZ}}^2(\hat{\tau}_{I}), \\
\hat{\tau}_{A\omega}^{\mathrm{adj}} &= \hat{\tau}_{I}^{\mathrm{adj}}(c_i), & \hat{\mathrm{se}}_{\mathrm{HW}}^2(\hat{\tau}_{A\omega}^{\mathrm{adj}}) &= \hat{\mathrm{se}}_{\mathrm{LZ}}^2\{\hat{\tau}_{I}^{\mathrm{adj}}(c_i)\}, \\
\hat{\tau}_{\omega A} &= \hat{\tau}_{T}, & \hat{\mathrm{se}}_{\mathrm{HW}}^2(\hat{\tau}_{\omega A}) &= \hat{\mathrm{se}}_{\mathrm{HW}}^2(\hat{\tau}_{T}), \\
\hat{\tau}_{\omega A}^{\mathrm{adj}} &= \hat{\tau}_{T}^{\mathrm{adj}}, & \hat{\mathrm{se}}_{\mathrm{HW}}^2(\hat{\tau}_{\omega A}^{\mathrm{adj}}) &= \hat{\mathrm{se}}_{\mathrm{HW}}^2(\hat{\tau}_{T}^{\mathrm{adj}}).
\end{aligned}
$$

The equivalence relationships in Proposition 2 suggest that, in general, $(\hat{\tau}_{A\pi}, \hat{\tau}_{A\pi}^{\mathrm{adj}})$ are similar to $(\hat{\tau}_{I}, \hat{\tau}_{I}^{\mathrm{adj}})$, and $(\hat{\tau}_{\pi A}, \hat{\tau}_{\pi A}^{\mathrm{adj}})$ are similar to $(\hat{\tau}_{T}, \hat{\tau}_{T}^{\mathrm{adj}})$. Proposition 2 also explains some discussions before: $\hat{\tau}_{A\pi}^{\mathrm{adj}}$ may not gain efficiency over $\hat{\tau}_{A\pi}$ for the same reason as that $\hat{\tau}_{I}^{\mathrm{adj}}$ may not gain efficiency over $\hat{\tau}_{I}$.

Based on the comparison of asymptotic efficiency in Corollaries 2 and 3, we finally recommend the use of $\hat{\tau}_{T}^{\mathrm{adj}}(n_i, \tilde{x}_{i\cdot})$ for estimating τ and $\hat{\tau}_{\pi A}^{\mathrm{adj}}(\pi_i, \pi_i c_i)$ for estimating τ_π. Importantly, the former should include n_i and $\tilde{x}_{i\cdot}$ as cluster-level covariates and the latter should include π_i and $\pi_i c_i$ as cluster-level

TABLE 1 Estimands, estimators, standard errors and corresponding theorems in this paper

Estimand τ			Estimand τ_π		
Estimator	**Standard error**	**Theorem**	**Estimator**	**Standard error**	**Theorem**
$\hat{\tau}_{I}$	$\hat{\mathrm{se}}_{\mathrm{LZ}}(\hat{\tau}_{I})$	1	$\hat{\tau}_{A\pi}$	$\hat{\mathrm{se}}_{\mathrm{HW}}(\hat{\tau}_{A\pi})$	5
$\hat{\tau}_{I}^{\mathrm{adj}}$	$\hat{\mathrm{se}}_{\mathrm{LZ}}(\hat{\tau}_{I}^{\mathrm{adj}})$	2	$\hat{\tau}_{A\pi}^{\mathrm{adj}}$	$\hat{\mathrm{se}}_{\mathrm{HW}}(\hat{\tau}_{A\pi}^{\mathrm{adj}})$	6
$\hat{\tau}_{T}$	$\hat{\mathrm{se}}_{\mathrm{HW}}(\hat{\tau}_{T})$	3	$\hat{\tau}_{\pi A}$	$\hat{\mathrm{se}}_{\mathrm{HW}}(\hat{\tau}_{\pi A})$	7
$\hat{\tau}_{T}^{\mathrm{adj}}$	$\hat{\mathrm{se}}_{\mathrm{HW}}(\hat{\tau}_{T}^{\mathrm{adj}})$	4	$\hat{\tau}_{\pi A}^{\mathrm{adj}}$	$\hat{\mathrm{se}}_{\mathrm{HW}}(\hat{\tau}_{\pi A}^{\mathrm{adj}})$	8

covariates. For both estimators, the heteroskedasticity-robust standard errors are convenient approximations to the true asymptotic standard errors.

The regression estimators based on cluster totals and averages are more efficient although they use only summary statistics of the individual data. This feature allows for using them when only clustered or grouped administrative data are available and the individual-level data are difficult to obtain due to privacy concerns. Schochet (2020) discusses the efficiency issue in this setting, comparing some estimators based on aggregate data and individual data.

Our theory deals with general individual-level covariates x_{ij} and cluster-level covariates c_i, but we must address the issue of defining them in practice. Assume that x_{ij}^* are the basic covariates varying with both clusters and individuals, c_i^* are the basic covariates varying only with clusters, and n_i is the cluster size. For instance, in an experiment with classroom-level intervention, x_{ij}^* can be the pre-test score of student j in classroom i, and c_i^* can be the years of education of the instructor for classroom i. To conduct individual-level regressions, we can choose x_{ij} to include (x_{ij}^*, c_i^*, n_i) and their functions, and to conduct cluster-level regressions, we can choose c_i to include $(\tilde{x}_{i\cdot}, c_i^*, n_i)$ and their functions. The final covariates x_{ij} or c_i must be properly centred for different estimators discussed above. Since our theory is model-assisted, the consistency and asymptotic Normality hold with any choice of the covariates as long as the regularity conditions hold. However, the number of clusters is not extremely large in most cluster-randomized experiments, so we must specify a parsimonious regression model to avoid the bias and efficiency loss due to overfitting.

6 | SIMULATION

This section uses simulation to evaluate the finite-sample properties of various estimators under cluster-randomized experiments, complementing the asymptotic theory in previous sections that assumes $M \rightarrow \infty$. We compare various estimators and standard errors under different data generating processes. We present the simulation results using Figure 1 in the main paper but relegate more detailed information to Tables A1 and A2 in the appendix.

6.1 | Comparing estimators and standard errors

Assume $M = 160$ and $e = 0.3$. Cluster i has $n_i \sim \text{round}\{\text{Unif}(3000/M \times 0.6, 3000/M \times 1.4)\}$ units. Unit (i, j) has covariate $x_{ij} \sim i/M + \text{Unif}(-1, 1)$, and potential outcomes $Y_{ij}(1) \sim \mathcal{N}\{2n_i M/N + (x_{ij} - \overline{x})^3, 1\}$, $Y_{ij}(0) \sim \mathcal{N}\{i/M + (x_{ij} - \overline{x})^2, 1\}$, which are nonlinear in x_{ij}. We draw independent samples and fix them in simulation. Over 1000 replications, we draw cluster-level treatment indicators, obtain observed individual outcomes, and calculate various estimators and standard errors. Figure 1a presents the boxplots of the estimators and coverage rates of the 95% Wald-type confidence intervals based on various variance estimators.

In Figure 1a, the horizontal line in the boxplots shows the true τ and that in the barplots shows the 95% nominal level, respectively. Numerically, $\widehat{\tau}_{\omega A}$ and $\widehat{\tau}_{\omega A}^{adj}$ are the same as $\widehat{\tau}_T$ and $\widehat{\tau}_T^{adj}$, so we omit them. From Figure 1a, $\widehat{\tau}_T^{adj}(n_i)$ performs better than $\widehat{\tau}_I$, and $\widehat{\tau}_T^{adj}(n_i, \tilde{x}_{i\cdot})$ performs the best among all the estimators, confirming Corollary 2; $\widehat{\tau}_A$ and $\widehat{\tau}_A^{adj}(\overline{x}_{i\cdot})$ are inconsistent for τ, confirming Theorems 5 and 6; $\widehat{\tau}_{A\omega}$ and $\widehat{\tau}_{A\omega}^{adj}(\overline{x}_{i\cdot})$ perform the same as $\widehat{\tau}_I$ and $\widehat{\tau}_I^{adj}(\overline{x}_{i\cdot})$, confirming Proposition 2. The coverage rates are often lower than the 95% nominal level when using $\widehat{se}_{OLS}$ and $\widehat{se}_{HW}$ for regression estimators based on individual data, and $\widehat{se}_{OLS}$ for regression estimators based on cluster totals and averages. Therefore,

(a) Simulation in Section 6.1: comparing the estimators

(b) Simulation in Section 6.2: noise covariate

(c) Simulation in Section 6.3: a counter-example

(d) Simulation in Section 6.4: with a large cluster

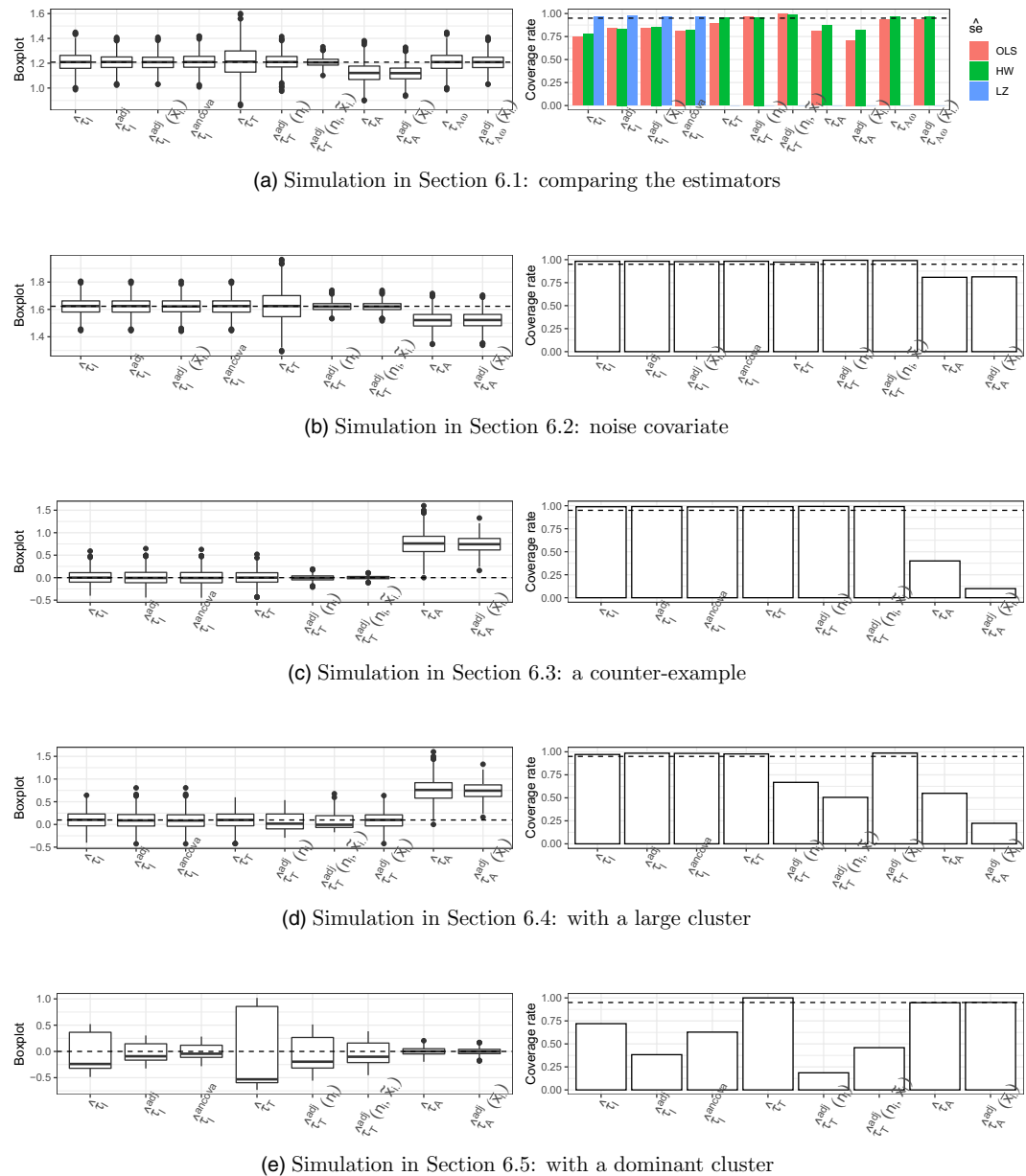(e) Simulation in Section 6.5: with a dominant cluster

FIGURE 1 Simulation [Colour figure can be viewed at wileyonlinelibrary.com]

we do not recommend those standard errors, and do not use them in other simulation studies and the application in the next section.

We end this subsection with a note on the improvement of efficiency of $\hat{\tau}_{\mathrm{T}}(n_i)$ over $\hat{\tau}_{\mathrm{I}}$ asymptotically. The key reason for this result is that the residual sum of squares from the OLS fit of $\sum_{j=1}^{n_i} Y_{ij}(z) M / N - \overline{Y}(z)$ on $n_i M / N - 1$ equals to the residual sum of squares from the OLS fit of $\tilde{\varepsilon}_{i\cdot}(z)$ on $n_i M / N - 1$, which is smaller than $\sum_{i=1}^{M} \{\tilde{\varepsilon}_{i\cdot}(z)\}^2$. In the above data generating process, we include n_i for generating $Y_{ij}(1)$ so that the improvement of $\hat{\tau}_{\mathrm{T}}(n_i)$ over $\hat{\tau}_{\mathrm{I}}$ is significant. However, if n_i is not predictive to $\tilde{\varepsilon}_{i\cdot}(z)$ at all, then the difference between $\hat{\tau}_{\mathrm{T}}(n_i)$ and $\hat{\tau}_{\mathrm{I}}$ is small. Section 7.2 will give simulation results to show this.

6.2 | Unrelated covariates

The last subsection shows a case in which covariates are predictive to the potential outcomes and thus covariate adjustment greatly improves the estimation efficiency. However, covariates may not be predictive in some applications. Although covariate adjustment cannot hurt the asymptotic efficiency based on our theoretical analysis, it is important to see whether the finite-sample results match the asymptotic analysis. We modify the data generating process to $Y_{ij}(1) \sim \mathcal{N}(2n_i M/N, 1)$ and $Y_{ij}(0) \sim \mathcal{N}(i/M, 1)$, removing all the x_{ij} term so that x_{ij} is completely unrelated to the potential outcomes.

Figure 1b shows the boxplots and coverage rates. Covariate adjustment based on x_{ij} does not improve the efficiency although the finite-sample cost is negligible in our setting. It is important to include the cluster size n_i in the cluster total regressions because $\hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(n_i)$ still improves both $\hat{\tau}_{\mathrm{T}}$ and $\hat{\tau}_{\mathrm{I}}$. The cluster size can be an important predictor of the cluster totals, which is a distinct feature of cluster-randomized experiments compared to completely randomized experiments on individuals.

Nevertheless, our results are fundamentally asymptotic with $M \to \infty$. With unrelated covariates and small M, we should apply covariate adjustment with caution since it might result in undesirable finite-sample bias without any efficiency gain.

6.3 | Counter-example for efficiency loss in covariate adjustment

We give an example where the asymptotic variance of $\hat{\tau}_{\mathrm{I}}^{\mathrm{adj}}$ is larger than that of $\hat{\tau}_{\mathrm{I}}$, so covariate adjustment in individual-level regression may cause efficiency loss. By Proposition 2, this example also apply to $\mathrm{se}^2(\hat{\tau}_{\mathrm{A}\pi}^{\mathrm{adj}}) \geq \mathrm{se}^2(\hat{\tau}_{\mathrm{A}\pi})$ by choosing $\pi_i = \omega_i$ and $c_i = \bar{x}_i$. Assume $M = 100$ and $e = 1/2$. We summarize the data generating process below, where $k = 1, \ldots, M/4$ and $j = 1, \ldots, n_i$:

	$i = 4k - 3$	$i = 4k - 2$	$i = 4k - 1$	$i = 4k$
n_i	20	20	30	10
x_{ij}	−5	−5	4	8
$Y_{ij}(1)$	−1	0	−1	5
$Y_{ij}(0)$	0	0	0	0

The finite population consists of $M/4$ replications of each of the four types of clusters. For instance, the first cluster above has size $n_i = 20$, and the covariates $x_{ij} = -5$ and potential outcomes $\{Y_{ij}(1), Y_{ij}(0)\} = (-1, 0)$ do not vary across units. Other clusters are similar. Because x_{ij} do not vary within cluster i, and we have shown that $\hat{\tau}_{\mathrm{A}\omega}$ and $\hat{\tau}_{\mathrm{A}\omega}^{\mathrm{adj}}(\bar{x}_i)$ perform the same as $\hat{\tau}_{\mathrm{I}}$ and $\hat{\tau}_{\mathrm{I}}^{\mathrm{adj}}(\bar{x}_i)$. So we omit them. Figure 1c shows that $\hat{\tau}_{\mathrm{I}}^{\mathrm{adj}}$ is slightly more variable than $\hat{\tau}_{\mathrm{I}}$, and the former indeed has larger variance than the latter.

6.4 | Cluster-randomized experiment with a large but not dominant cluster

In this subsection, we show the finite-sample properties of various estimators under a case with a large but not dominant cluster. The data generating process is the same as Section 6.3 except that the last cluster has $n_M = 50$ units. Figure 1d shows the boxplots and coverage rates. The estimators based on regressions with scaled cluster totals are biased when the cluster-level covariates include the cluster size. To demonstrate that it is the cluster size n_i causing the inconsistency, we also report $\hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(\bar{x}_i)$ and

find that it is nearly unbiased. In contrast, all the estimators based on regressions with individual-level data have small biases, illustrating the trade-off of efficiency and robustness issue we discussed in Section 3.4.

6.5 | Cluster-randomized experiment with a dominant cluster

We further investigate a case with a dominant cluster. Assume $M = 100$ and $e = 0.4$. For $i = 1, \ldots, M-1$, cluster i has $n_i \sim \text{round}\{\text{Unif}(2000/M \times 0.6, 2000/M \times 1.4)\}$ units. Cluster M has $n_M = 800$ units. Unit (i, j) has covariate $x_{ij} = i/M$ and potential outcomes $Y_{ij}(1) = Y_{ij}(0) \sim \mathcal{N}\{(i/M)^2 + (x_{ij} - \bar{x})^2, 1\}$ so that the sharp null hypothesis of no treatment effect holds. Figure 1e shows the boxplots and coverage rates. In this case, the regularity conditions on the cluster sizes are violated rendering estimators based on regression with individual data and cluster totals not asymptotically Normal. Ironically, the unbiased estimator $\hat{\tau}_T$ is the worst among all estimators. In contrast, the estimators based on regression with cluster averages $\hat{\tau}_A$ and $\hat{\tau}_A^{\text{adj}}$ are well-behaved: they are centring at the truth zero with coverage rates close to the nominal level as shown in Figure 1e.

This example shows that changing the estimand from τ to τ_π, we can use regressions based on cluster averages to reduce bias and increase efficiency when the cluster sizes vary dramatically. They can be particularly useful as test statistics for testing the sharp null hypothesis that $Y_{ij}(1) = Y_{ij}(0)$ $(i = 1, \ldots, M; j = 1, \ldots, n_i)$ because $\tau = \tau_\pi = 0$ under this null hypothesis.

7 | APPLICATION

7.1 | Data analysis

Poor sanitation leads to morbidity and mortality in developing countries. In 2012, Guiteras et al. (2015) conducted a cluster-randomized experiment in rural Bangladesh to evaluate the effectiveness of different policies on the use of hygienic latrines. To illustrate our theory, we use a subset of their original data: 63 villages were randomly assigned to the treatment arm receiving subsidies and 22 villages were randomly assigned to the control arm without subsidies. We exclude the households not eligible for subsidies or with missing outcomes, resulting in 10,125 households in total. The median, mean and maximum village sizes are 83, 119 and 500, respectively. We choose the outcome Y_{ij} as the binary indicator for whether the household (i, j) had access to a hygienic latrine or not, measured in June 2013, and covariate x_{ij} as the access rate to hygienic latrines in the community that household (i, j) belonged to, measured in January 2012 before the experiment.

We report various estimators in Figure 2a and present more details in Table A3a in the appendix. We omit $\hat{\tau}_{A\omega}$ and $\hat{\tau}_{A\omega}^{\text{adj}}(\bar{x}_{i\cdot})$ because they are identical to $\hat{\tau}_I$ and $\hat{\tau}_I^{\text{adj}}(\bar{x}_{i\cdot})$. Figure 2a shows that $\hat{\tau}_T^{\text{adj}}(n_i, \tilde{x}_{i\cdot})$ has the smallest standard error, and $\hat{\tau}_T$ has the largest standard error. Comparing these two estimators, covariate adjustment in the cluster total regression reduces the standard error to about 25%, which is equivalent to increase the number of clusters about 16 times. This dramatic improvement is not surprising given the poor performance of $\hat{\tau}_T$ in our simulation studies. Compared to the regression estimators with individual data, $\hat{\tau}_T^{\text{adj}}(n_i, \tilde{x}_{i\cdot})$ has modest improvement in the estimated standard error. Our theory suggests the use of $\hat{\tau}_T^{\text{adj}}(n_i, \tilde{x}_{i\cdot})$, based on which we conclude that averaged over all households, subsidies increased the probability of getting access to hygienic latrines by 15.2% with statistical significance at level 95%, corroborating the analysis of Guiteras et al. (2015).

(a) Point estimates with confidence intervals, and standard errors based on Guiteras et al. (2015)
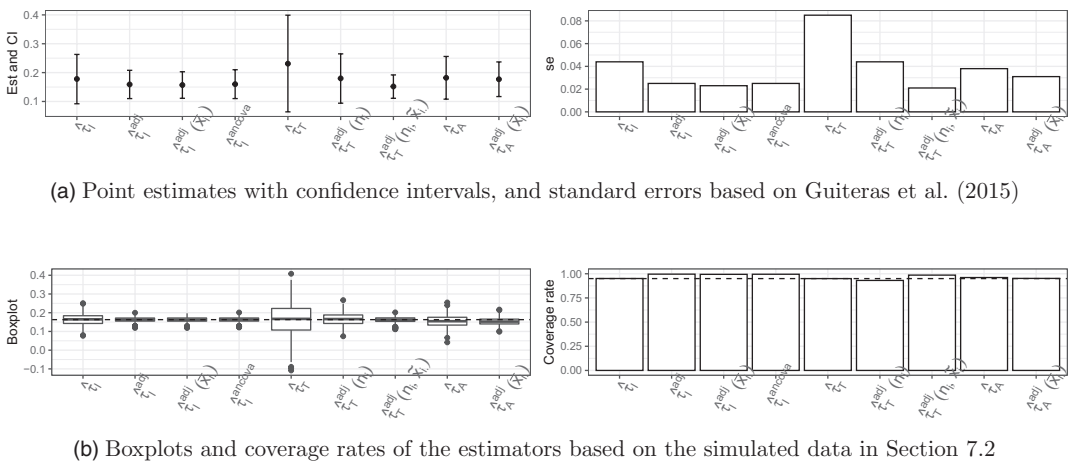
(b) Boxplots and coverage rates of the estimators based on the simulated data in Section 7.2

FIGURE 2 Application

7.2 | Simulation based on the data

We only have 22 villages under control, making the asymptotic approximations suspicious. To evaluate the estimators further, we conduct additional simulation with missing potential outcomes in Guiteras et al. (2015) imputed based on model fitting. Although not all estimators perform well, $\hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(n_i, \tilde{x}_{i\cdot})$ has small bias and the associated Wald-type confidence interval achieves the nominal coverage rate, reassuring our choice of the estimator and confidence interval.

We fit a linear probability model of Y_{ij} on $(1, Z_{ij}, x_{ij}, Z_{ij}x_{ij})$, and then impute the missing potential outcomes as Bernoulli random variables. Pretending that these are the true potential outcomes, we can compute τ, various estimators and the associated standard errors. Figure 2b shows the boxplots and coverage rates of estimators. In this case, we have a relatively small number of clusters in the control group, and several estimators perform poorly: $\hat{\tau}_{\mathrm{A}}$ and $\hat{\tau}_{\mathrm{A}}^{\mathrm{adj}}(\bar{x}_{i\cdot})$ are biased because $\bar{\tau}$ differs from τ for our chosen seed for simulation; $\hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(n_i)$ is biased although the asymptotic theory suggests that it is consistent. Fortunately, $\hat{\tau}_{\mathrm{I}}, \hat{\tau}_{\mathrm{I}}^{\mathrm{adj}}, \hat{\tau}_{\mathrm{I}}^{\mathrm{adj}}(\bar{x}_{i\cdot}), \hat{\tau}_{\mathrm{I}}^{\mathrm{ancova}}, \hat{\tau}_{\mathrm{T}}$ and $\hat{\tau}_{\mathrm{T}}^{\mathrm{adj}}(n_i, \tilde{x}_{i\cdot})$ have small biases and the associated Wald-type confidence intervals are conservative. We relegate more detailed results to Table A3b in the appendix.

8 | DISCUSSION

Cluster-randomized experiments are widely used in social sciences and public health. Currently, model-based inference is the dominant framework for analysing them (Donner & Klar, 2000; Turner et al., 2017b). Following a growing literature that focuses on the experimental design itself, we re-investigate the theoretical properties of several regression estimators under the design-based perspective that does not require correctly specified models. We found that the commonly used OLS estimators with individual data are suboptimal, and a simpler regression estimator based on scaled cluster totals is consistent and efficient with a Normal limiting distribution. We also proposed a regression estimator for a general weighted average treatment effect over clusters.

With clustered individual data, we have discussed regression estimators based on OLS, which are special cases of the generalized estimating equation with the independent working correlation matrix (Liang & Zeger, 1986). Surprisingly, the simplest version of the generalized estimating equation

suffices for generating a consistent and asymptotically Normal estimator for the average treatment effect. It is interesting to extend the discussion to other distributions, link functions and working correlation matrices. Another widely used method to analyse cluster-randomized experiments is the linear mixed effects model (Donner & Klar, 2000; Green & Vavreck, 2008). However, the causal interpretation of the estimator from this model is unknown under the design-based perspective that allows the model to be misspecified. This is an intriguing problem that is worth further investigations.

We focused on the asymptotic theory requiring a large number of clusters, but many cluster-randomized experiments have only a small number of clusters. In those cases, we can use bias-corrected robust standard errors (Angrist & Pischke, 2008) to improve the finite-sample coverage rates. More fundamentally, the asymptotic theory can break down with a small fixed number of clusters. Alternative asymptotic regimes are possible, but they must further restrict the outcome-generating process. One alternative approach is to apply the Bayesian machinery by imposing some parametric assumptions on the data generating process and prior distributions on the unknown parameters. Feller and Gelman (2015) reviewed the standard approach of using hierarchical models for causal inference with cluster-randomized experiments. This approach requires strong modelling assumptions, and it is a curious research topic to study its design-based properties. Another alternative approach is to use the Fisher randomization tests that yield finite-sample exact p-values (Ding & Keele, 2018; Gail et al., 1996; Zhang et al., 2012). This approach often restricts the model of the individual treatment effect, for example, testing the sharp null hypothesis. Our asymptotic analysis, coupled with Zhao and Ding (2021), suggests that Fisher randomization tests with the estimators studentized by the appropriate standard errors deliver p-values that are exact under the sharp null hypothesis and asymptotically valid under the weak null hypothesis of $\tau = 0$. This is our final recommendation for analysing cluster-randomized experiments with small M.

We ignored noncompliance, a common issue in cluster-randomized experiments (Frangakis et al., 2002; Jo et al., 2008; Kang & Keele, 2018a,b; Small et al., 2008). Our analysis covers the case with complete randomization on clusters, but many clustered experiments are matched or stratified based on pretreatment covariates (Donner & Klar, 2000; Imai et al., 2009; Schochet et al., 2021). We will investigate these directions in future research.

ORCID

Peng Ding http://orcid.org/0000-0002-2704-2353

REFERENCES

Abadie, A., Athey, S., Imbens, G.W. & Wooldridge, J. (2017) When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Angrist, J.D. & Pischke, J.-S. (2008) *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Athey, S. & Imbens, G.W. (2017) The econometrics of randomized experiments. In: Banerjee, A. & Duflo, E. (Eds.) *Handbook of economic field experiments*, vol. 1, chapter 3. Amsterdam: North-Holland, pp. 73–140.

Barrios, T., Diamond, R., Imbens, G.W. & Kolesár, M. (2012) Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, 107, 578–591.

Basse, G. & Feller, A. (2018) Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113, 41–55.

Bloniarz, A., Liu, H., Zhang, C.H., Sekhon, J. & Yu, B. (2016) Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7383–7390.

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

Cornfield, J. (1978) Randomization by group: A formal analysis. *American Journal of Epidemiology*, 108, 100–102.

Dasgupta, T., Pillai, N.S. & Rubin, D.B. (2015) Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 727–753.

Ding, P. & Keele, L. (2018) Rank tests in unmatched clustered randomized trials applied to a study of teacher training. *Annals of Applied Statistics*, 12, 2151–2174.

Donner, A. & Klar, N. (2000) *Design and analysis of cluster randomization trials in health research*. London, UK: Arnold Publishers Limited.

Feller, A. & Gelman, A. (2015) Hierarchical models for causal effects. In: Scott, R. & Kosslyn, S. (Eds.) *Emerging trends in the social and behavioral sciences*. New York: Wiley.

Fisher, R.A. (1935) *The design of experiments*, 1st edn. Edinburgh, London: Oliver and Boyd.

Fogarty, C.B. (2018a) On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 1035–1056.

Fogarty, C.B. (2018b) Regression assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105, 994–1000.

Frangakis, C.E., Rubin, D.B. & Zhou, X.-H. (2002) Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics*, 3, 147–164.

Freedman, D.A. (2008a) On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, 2, 176–196.

Freedman, D.A. (2008b) On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40, 180–193.

Fuller, W.A. (2009) *Sampling statistics*. New York: John Wiley & Sons.

Gail, M.H., Mark, S.D., Carroll, R.J., Green, S.B. & Pee, D. (1996) On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15, 1069–1092.

Graubard, B.I. & Korn, E.L. (1994) Regression analysis with clustered data. *Statistics in Medicine*, 13, 509–522.

Green, D.P. & Vavreck, L. (2008) Analysis of cluster-randomized experiments: A comparison of alternative estimation approaches. *Political Analysis*, 16, 138–152.

Guiteras, R., Levinsohn, J. & Mobarak, A.M. (2015) Encouraging sanitation investment in the developing world: A cluster-randomized trial. *Science*, 348, 903–906.

Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Le Cam, L.M. & Neyman, J. (Eds.) *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. Berkeley, California: University of California Press, pp. 221–233.

Imai, K., King, G. & Nall, C. (2009) The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science*, 24, 29–53.

Imbens, G.W. & Rubin, D.B. (2015) *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge: Cambridge University Press.

Jo, B., Asparouhov, T., Muthén, B.O., Ialongo, N.S. & Brown, C.H. (2008) Cluster randomized trials with treatment noncompliance. *Psychological Methods*, 13, 1–18.

Kang, H. & Keele, L. (2018a) Estimation methods for cluster randomized trials with noncompliance: A study of a biometric smartcard payment system in India. *arXiv preprint* arXiv:1805.03744.

Kang, H. & Keele, L. (2018b) Spillover effects in cluster randomized trials with noncompliance. *arXiv preprint* arXiv:1808.06418.

Lei, L. & Ding, P. (2021) Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, page in press.

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

Liang, K.-Y. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

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

Liu, H. & Yang, Y. (2020) Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107, 935–948.

Lohr, S.L. (2019) *Sampling: Design and analysis*, 2nd edn. New York: Chapman and Hall/CRC.

Middleton, J.A. (2008) Bias of the regression estimator for experiments using clustered random assignment. *Statistics and Probability Letters*, 78, 2654–2659.

Middleton, J.A. & Aronow, P.M. (2015) Unbiased estimation of the average treatment effect in clusterrandomized experiments. *Statistics, Politics and Policy*, 6, 39–75.

Miratrix, L.W., Sekhon, J.S. & Yu, B. (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 369–396.

Mukerjee, R., Dasgupta, T. & Rubin, D.B. (2018) Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113, 868–881.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). Reprinted ed. *Statistical Science*, 5, 465–472.

Raudenbush, S.W. (1997) Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173–185.

Raudenbush, S.W. & Schwartz, D. (2020) Randomized experiments in education, with implications for multilevel causal inference. *Annual Review of Statistics and its Application*, 7, 177–208.

Schochet, P.Z. (2013) Estimators for clustered education RCTs using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics*, 38, 219–238.

Schochet, P.Z. (2020) Analyzing grouped administrative data for RCTs using design-based methods. *Journal of Educational and Behavioral Statistics*, 45, 32–57.

Schochet, P.Z., Pashley, N.E., Miratrix, L.W. & Kautz, T. (2021) Design-based ratio estimators and central limit theorems for clustered, blocked RCTs. *Journal of the American Statistical Association*, page in press.

Small, D.S., Ten Have, T.R. & Rosenbaum, P.R. (2008) Randomization inference in a group–randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. *Journal of the American Statistical Association*, 103, 271–279.

Turner, E.L., Li, F., Gallis, J.A., Prague, M. & Murray, D.M. (2017a) Review of recent methodological developments in group-randomized trials: Part 1—design. *American Journal of Public Health*, 107, 907–915.

Turner, E.L., Prague, M., Gallis, J.A., Li, F. & Murray, D.M. (2017b) Review of recent methodological developments in group-randomized trials: Part 2—analysis. *American Journal of Public Health*, 107, 1078–1086.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.

Zhang, K., Traskin, M. & Small, D.S. (2012) A powerful and robust test statistic for randomization inference in group-randomized trials with matched pairs of groups. *Biometrics*, 68, 75–84.

Zhao, A. & Ding, P. (2021) Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics*, page in press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Su F, Ding P. Model-assisted analyses of cluster-randomized experiments. *J R Stat Soc Series B*. 2021;83:994–1015. https://doi.org/10.1111/rssb.12468