Efficient Covariate Adjustment in Stratified Experiments

Max Cytrynbaum*

February 7, 2023

Abstract

This paper studies covariate adjusted estimation of the average treatment effect (ATE) in stratified experiments. We work in the stratified randomization framework of Cytrynbaum (2021), which includes matched tuples designs (e.g. matched pairs), coarse stratification, and complete randomization as special cases. Interestingly, we show that the Lin (2013) interacted regression is generically asymptotically inefficient, with efficiency only in the edge case of complete randomization. Motivated by this finding, we derive the optimal linear covariate adjustment for a given stratified design, constructing several new estimators that achieve the minimal variance. Conceptually, we show that optimal *linear* adjustment of a stratified design is equivalent in large samples to doubly-robust *semiparametric* adjustment of an independent design. We also develop novel asymptotically exact inference for the ATE over a general family of adjusted estimators, showing in simulations that the usual Eicker-Huber-White confidence intervals can significantly overcover. Our inference methods produce shorter confidence intervals by fully accounting for the precision gains from both covariate adjustment and stratified randomization. Simulation experiments and an empirical application to the Oregon Health Insurance Experiment data (Finkelstein et al. (2012)) demonstrate the value of our proposed methods.

*Yale Department of Economics. Correspondence: max.cytrynbaum@yale.edu

1 Introduction

This paper studies covariate adjusted estimation of the average treatment effect (ATE) in stratified experiments. Researchers often make use of both stratified randomization at the design stage and ex-post covariate adjustment to improve the precision of experimental estimates. Indeed, Lin (2013) showed in a finite population setting that the regression estimator with full treatment-covariate interactions is asymptotically weakly more efficient than difference-of-means (unadjusted) estimation for completely randomized designs. Negi and Wooldridge (2021) extended these results to estimation of the ATE in a super-population framework. However, questions remain about the consequences of combining stratification and regression adjustment for estimator efficiency and the power of inference methods. To study these questions, we use the stratified randomization framework of Cytrynbaum (2021), which includes matched tuples designs (e.g. matched pairs), coarse stratification, and complete randomization as special cases. As in Negi and Wooldridge (2021) and Bai et al. (2021), we target the ATE as this is often the most policy-relevant parameter in social science applications.

Interestingly, we show that the Lin (2013) regression is generically asymptotically inefficient in the family of linearly adjusted estimators, with efficiency only in the edge case of complete randomization. Motivated by this finding, we study efficient linear covariate adjustment for stratified designs.

Our first set of results characterizes the optimal linear adjustment coefficient for a given stratification. We show that asymptotically, the Lin (2013) regression uses the wrong objective function, minimizing a marginal variance objective that is insensitive to the stratification. By contrast, the optimal adjustment coefficient minimizes a *mean conditional variance* objective, conditional on the implemented stratification. Intuitively, the optimal adjustment is tailored to the experimental design, ignoring variance components that are predictable by the stratification variables. Section 3.2 draws an interesting connection with semiparametric regression, showing that optimal linear adjustment of a given stratified design is asymptotically equivalent to doubly-robust semiparametric adjustment of an i.i.d. design. This complements the recent findings in Cytrynbaum (2021), which showed that the difference-of-means estimator attains the Hahn (1998) semiparametric variance bound for the ATE[1] under finely stratified randomization (see Section 2 for details).

Our second set of results constructs feasible versions of the oracle linear adjustment derived in Section 3.1. First, we study adjustment under a "rich covariates" assumption, requiring the conditional expectation of the covariates given the stratification variables to be affine in a known vector of transformations of these variables. Under this assumption, adding sufficiently rich transformations of the stratification variables to the Lin (2013) estimator restores asymptotic efficiency. Next, we relax this condition, providing three different regression estimators that are asymptotically efficient under weak conditions. In particular, we show asymptotic optimality of within-stratum (inconsistently) partialled versions of the interacted and tyranny-of-the-minority estimators (Lin (2013)) as well as a generalization of an estimator proposed in Imbens and Rubin (2015) for the case of

[1]We refer to the semiparametric variance bound for i.i.d. observations $(Y, D, \psi(X))$, where $\psi(X)$ are the stratification variables, introduced in Section 2 below.

matched pairs.

Our final contribution is to develop novel asymptotically exact inference methods for covariate adjusted estimation under stratified designs. By asymptotically exact, we mean that coverage probabilities of our confidence intervals converge to the specified nominal level (no overcoverage). These methods apply to a generic family of linear covariate adjustments and randomization schemes, including the non-interacted regression estimator, Lin (2013) interacted regression, and all of the other estimators considered in this paper. Our simulations and empirical application suggest that the usual Eicker-Huber-White confidence intervals can significantly overcover in this setting, generalizing similar results in Bai et al. (2021) for the difference-of-means estimator under matched-pairs designs.

There has been significant interest in treatment effect estimation under different experimental designs in the recent literature. Some papers studying covariate adjustment under stratified randomization include Bugni et al. (2018), Fogarty (2018), Liu and Yang (2020), Lu and Liu (2022), Ma et al. (2020), Reluga et al. (2022), Wang et al. (2021), Ye et al. (2022), and Zhu et al. (2022). These works differ from our paper in at least one of the following ways: (1) studying inference on the sample average treatment effect (SATE) rather than the ATE in a super-population, (2) restricting to coarse stratification (stratum size going to infinity), or (3) proving weak efficiency gains but not optimality. In a finite population setting, Zhu et al. (2022) shows asymptotic efficiency of a projection-based estimator similar but not equivalent to the "partialled Lin" approach considered in Section 3.4.2. In the same setting, Lu and Liu (2022) prove efficiency of a tyranny-of-the-minority style regression similar to one the considered in Section 3.4.4. Both papers give asymptotically conservative inference on the SATE, while we provide asymptotically exact inference on the ATE using a generalized pairs-of-pairs (Abadie and Imbens (2008)) style approach. See the remarks in Section 3.4 below for a more detailed comparison.

Relative to the above papers, the super-population framework considered here creates new technical challenges. For example, as pointed out in Bai et al. (2021), matching units into data-dependent strata post-sampling produces a complicated dependence structure between the treatment assignments and random covariates. We deal with this using a tight-matching condition (Equation 2.1) and martingale CLT analysis from Cytrynbaum (2021). This setting also has analytical advantages that allow us to establish novel conceptual results. For example, the variational characterization of the optimal adjustment coefficent in Section 3.1 allows us to state explicit necessary and sufficient conditions for the efficiency of several commonly used regression estimators (Sections 3.4, 8.1). Similarly, the efficiency of Lin estimation under a "rich covariates" condition (Section 3.3) and the equivalence between optimal linear adjustment of stratified designs and doubly-robust semiparametric adjustment (Section 3.2) appear to be unique in this literature. To the best of our knowledge, we give the first asymptotically exact inference on the ATE for generic covariate adjusted estimators under finely stratified randomization (Section 4).

The rest of the paper is organized as follows. In Section 2 we define notation and introduce the family of stratified designs that we will consider throughout the paper. Section 3 gives our main results characterizing optimal covariate adjustment and constructing efficient estimators. Section 4 provides asymptotically exact inference on the ATE for generic linearly adjusted estimators. In Sections 5 and 6, we study the finite sample properties of

our method, including an empirical application to the Oregon Health Insurance Experiment Data (Finkelstein et al. (2012)). Section 7 concludes with some recommendations for practitioners.

2 Framework and Stratified Designs

For a binary treatment $d \in \{0, 1\}$, let $Y_i(1)$, $Y_i(0)$ denote the treated and control potential outcomes, respectively. For treatment assignment D_i, let $Y_i = Y_i(D_i) = D_i Y_i(1) + (1 - D_i) Y_i(0)$ be the observed outcome. Let X_i denote covariates. Consider data $(X_i, Y_i(1), Y_i(0))_{i=1}^n$ sampled i.i.d. from a super-population of interest. We are interested in estimating the average treatment effect in this population, $\text{ATE} = E[Y(1) - Y(0)]$.

After sampling units $i = 1, \ldots, n$, treatments $D_{1:n}$ are assigned by stratified randomization. In particular, we use the "local randomization" framework introduced in Cytrynbaum (2021).

Definition 2.1 (Local Randomization)**.** Let treatment proportions $p = a/k$ with $\gcd(a, k) = 1$. Partition the experimental units into disjoint groups $g \subset [n]$ with $[n] = \bigsqcup_g \{i \in g\}$ and $|g| = k$.[2] Let $\psi(X) \in \mathbb{R}^{d_\psi}$ denote a vector of stratification variables. Suppose that the groups that satisfy a homogeneity condition[3] with respect to $\psi(X)$

$$\frac{1}{n} \sum_g \sum_{\substack{i,j \in g \\ i<j}} |\psi(X_i) - \psi(X_j)|_2^2 = o_p(1) \tag{2.1}$$

Independently for each $|g| = k$, draw treatment variables $(D_i)_{i \in g}$ by setting $D_i = 1$ for exactly a out of k units, completely at random. For a stratification satisfying these conditions, we denote $D_{1:n} \sim \text{Loc}(\psi, p)$.

Remark 2.2 (Matched Tuples)**.** Equation 2.1 requires units in a group to have similar $\psi(X_i)$ values and can be thought of as a tight-matching condition. Cytrynbaum (2021) gives a "block path" algorithm to match units into groups that provably satisfy this condition. Drawing treatments $D_{1:n} \sim \text{Loc}(\psi, p)$ produces a "matched k-tuples" design for $p = a/k$. Matched pairs corresponds to the case $p = 1/2$.

Remark 2.3 (Coarse Stratification)**.** At first glance, Definition 2.1 produces n/k groups of units that are tightly matched in $\psi(X)$ space, suggesting a fine stratification. However, complete randomization and coarse stratification can also be obtained in this framework. For example, for complete randomization with $p = 2/3$, set $\psi_i = 1$ for all i and form groups g at random. This gives a "random matched triples" representation of complete randomization with $p = 2/3$. Similarly, coarse stratification with large fixed strata $S(X) \in \{1, \ldots, m\}$ can be obtained by setting $\psi(X) = S(X)$ and matching units with identical $S(X)$ values into groups at random.[4] Because of this, our framework enables a unified analysis for a wide range of stratifications.

[2] For notational simplicity, assume that n is divisible by k.

[3] Bai et al. (2021) introduce a related condition for super-population analysis of matched pairs, using the unsquared Euclidean norm.

[4] Formally, in the first case $g = g(\pi_n)$ for some data-independent randomness π_n and $g = g(\pi_n, S(X_i)_{i=1}^n)$ in the second case. See Cytrynbaum (2021) for details.

Remark 2.4 (Treatment Proportions)**.** Cytrynbaum (2021) extends Definition 2.1 to accommodate general propensity scores $p(x)$. We restrict to the case with $p(x) = p$ constant in this paper, as this simplifies presentation of the main issues while also encompassing the majority of use cases for experiments in the social sciences.

We suppose that the experimenter:

(1) Samples units and observes baseline covariates $(X_i)_{i=1}^n$.

(2) Forms data-dependent groups $g_i = g_i(\psi_{1:n})$ that satisfy Equation 2.1 for some stratification variables $\psi(X)$.

(3) Draws treatment assignments $D_{1:n} \sim \text{Loc}(\psi, p)$, observes outcomes $Y_i(D_i)$, and forms an estimate of ATE.

The simple difference-of-means estimator is given by the coefficient $\widehat{\theta}$ on D_i in the regression $Y_i \sim 1 + D_i$. Before proceeding, we state a helpful variance decomposition for $\widehat{\theta}$ under stratified randomization that will be used extensively below. Let $c(X) = E[Y(1) - Y(0)|X]$ denote the conditional average treatment effect (CATE) and $\sigma_d^2(X) = \text{Var}(Y(d)|X)$ the heteroskedasticity function. Define the *balance function*

$$b(X; p) = E[Y(1)|X]\left(\frac{1-p}{p}\right)^{1/2} + E[Y(0)|X]\left(\frac{p}{1-p}\right)^{1/2} \tag{2.2}$$

We often denote $b = b(X; p)$ in what follows. Cytrynbaum (2021) shows that if $D_{1:n} \sim \text{Loc}(\psi, p)$ then $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

$$V = \text{Var}(c(X)) + E[\text{Var}(b|\psi)] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right] \tag{2.3}$$

The variance V is in fact the Hahn (1998) semiparametric variance bound for the ATE (with covariates $\psi(X)$), giving a formal sense in which stratification does nonparametric regression adjustment "by design." The middle term is the most important for our analysis below. For example, the difference in efficiency between stratifications ψ_1 and ψ_2 (for fixed p) is simply $E[\text{Var}(b|\psi_1)] - E[\text{Var}(b|\psi_2)]$. Note also that $E[\text{Var}(b|\psi)] \leq \text{Var}(b)$ for any ψ, showing how stratification removes the variance due to fluctuations that are predictable by $\psi(X)$.

Moving beyond the difference-of-means estimator $\widehat{\theta}$, suppose that at the analysis stage, the experimenter has access[5] to covariates $h(X)$. One may try to further improve the efficiency of ATE estimation by regression adjustment using these covariates, as in Lin (2013). We describe the interaction between covariate adjustment and stratification in Section 3.1 below, characterizing the optimal linear adjustment.

3 Main Results

3.1 Efficient Linear Adjustment in Stratified Experiments

In this section, we begin by studying the efficiency of commonly used covariate-adjusted estimators for the ATE under stratified randomization. Lin (2013) showed that in a completely randomized experiment, $\psi = 1$ in our framework, regression adjustment with full

[5] At this stage, we are agnostic to whether $\psi(X) \subseteq h(X)$.

treatment-covariate interactions is asymptotically weakly more efficient than difference-of-means estimation.[6]

Interestingly, we show that this result is atypical. For a general stratified experiment with $\psi \neq 1$, Lin (2013) style regression adjustment may be strictly inefficient relative to difference-of-means. The problem is that the interacted regression solves the wrong optimization problem, estimating the same adjustment coefficient for any stratification in the class $D_{1:n} \sim \text{Loc}(\psi, p)$. In general, this leads to a sub-optimal variance-covariance tradeoff, potentially making the interacted regression even less efficient than difference-of-means.

Denote $h_i = h(X_i)$ and de-meaned covariates $\tilde{h}_i = h_i - E_n[h_i]$. Let the Lin estimator $\widehat{\theta}_L$ be the coefficient on D_i in the fully-interacted regression

$$Y_i \sim (1, \tilde{h}_i) + D_i(1, \tilde{h}_i) \tag{3.1}$$

The following theorem characterizes the efficiency of this estimator. The technical conditions needed for the result are given in Assumption 8.2 in the appendix.

Theorem 3.1. *Let 8.2 hold. If $D_{1:n} \sim \text{Loc}(\psi, p)$ then $\sqrt{n}(\widehat{\theta}_L - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with*

$$V = \text{Var}(c(X)) + E\left[\text{Var}(b - \gamma_L' h | \psi)\right] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

and coefficient

$$\gamma_L = \underset{\gamma \in \mathbb{R}^{d_h}}{\text{argmin}} \, \text{Var}(b - \gamma' h)$$

Note that only the middle term of this expression differs from the limiting variance of the unadjusted estimator, changing from $E[\text{Var}(b|\psi)]$ for difference-of-means to $E[\text{Var}(b - \gamma_L' h|\psi)]$ for the Lin estimator. In the case of complete randomization ($\psi = 1$), the Lin estimator is weakly more efficient that difference-of-means, since in this case

$$E[\text{Var}(b - \gamma_L' h|\psi)] = \text{Var}(b - \gamma_L' h) = \min_{\gamma} \text{Var}(b - \gamma' h) \leq \text{Var}(b)$$

For general stratification ($\psi \neq 1$), however, the Lin estimator may be strictly inefficient. This is because the Lin estimator solves the wrong optimization problem in general

$$\gamma_L = \underset{\gamma \in \mathbb{R}^{d_h}}{\text{argmin}} \, \text{Var}(b - \gamma' h) \neq \underset{\gamma \in \mathbb{R}^{d_h}}{\text{argmin}} \, E[\text{Var}(b - \gamma' h|\psi)]$$

In particular, it uses the same adjustment coefficient $\gamma_L = \text{argmin}_\gamma \, \text{Var}(b - \gamma' h)$ for any stratification variables $\psi(X)$. The following example shows how this can lead to strict inefficiency relative to difference of means estimation.

Example 3.2 (Strict Inefficiency)**.** Suppose that $b = f_1(\psi) + e_1$ and $h = f_2(\psi) + e_2$ with $\text{Cov}(f_1, f_2) > 0$, $\text{Var}(e_2) > 0$ and (ψ, e_1, e_2) jointly independent. Then

$$\gamma_L = \text{Var}(h)^{-1} \, \text{Cov}(h, b) = \frac{\text{Cov}(f_1(\psi), f_2(\psi))}{\text{Var}(f_2(\psi)) + \text{Var}(e_2)} > 0$$

[6] Lin (2013) works in a design-based framework, with fixed potential outcomes and covariates and target parameter SATE $= E_n[Y_i(1) - Y_i(0)]$.

Then the Lin estimator is strictly less efficient than unadjusted estimation since

$$\begin{aligned} E[\mathrm{Var}(b - \gamma' h|\psi)] - E[\mathrm{Var}(b|\psi)] &= -2\gamma_L E[\mathrm{Cov}(h, b|\psi)] + \gamma_L^2 E[\mathrm{Var}(h|\psi)] \\ &= \gamma_L^2 \,\mathrm{Var}(e_2) > 0 \end{aligned}$$

The second equality since $E[\mathrm{Cov}(h, b|\psi)] = E[\mathrm{Cov}(e_1, e_2|\psi)] = \mathrm{Cov}(e_1, e_2) = 0$.

Non-interacted Regression - An analogue of Theorem 3.1 also holds for the "naive" non-interacted regression estimator

$$Y_i \sim 1 + D_i + h_i \tag{3.2}$$

under stratified designs $D_{1:n} \sim \mathrm{Loc}(\psi, p)$. For completeness, we give asymptotic theory and optimality conditions for this estimator in Section 8.1 in the appendix. Asymptotically exact inference is available in Section 4.

The proof of Theorem 3.1 shows that the Lin estimator $\widehat{\theta}_L$ can be written in the canonical form $\widehat{\theta}_L = \widehat{\theta} - \widehat{\gamma}_L'(\bar{h}_1 - \bar{h}_0)$ with treatment arm covariate means $\bar{h}_1 = \frac{E_n[h_i D_i]}{E_n[D_i]}$ and $\bar{h}_0 = \frac{E_n[h_i(1-D_i)]}{E_n[1-D_i]}$. In this case, the *adjustment coefficient* $\widehat{\gamma}_L = (1-p)(\widehat{a}_1 + \widehat{a}_0) + p\widehat{a}_0$, where $\widehat{a}_1$ and $\widehat{a}_0$ are the coefficients on D_i and $D_i h_i$ in the Lin regression. We show that most commonly used estimators can be written in this standard form for some $\widehat{\gamma}$, up to lower order factors.

Define the variance normalization constant $c_p = \sqrt{p(1-p)}$. The following theorem describes the asymptotic properties of adjusted estimators $\widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p$.

Theorem 3.3. *Let Assumption 8.2 hold. Suppose* $\widehat{\gamma} \xrightarrow{p} \gamma$ *and consider the adjusted estimator*

$$\widehat{\theta}(\widehat{\gamma}) = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p$$

If $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ *then* $\sqrt{n}(\widehat{\theta}(\widehat{\gamma}) - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V(\gamma))$

$$V(\gamma) = \mathrm{Var}(c(X)) + E\Big[\mathrm{Var}(b - \gamma' h|\psi)\Big] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right] \tag{3.3}$$

Motivated by Theorem 3.3, we define a linearly-adjusted estimator to be asymptotically efficient if it minimizes the variance expression in Equation 3.3.

Definition 3.4 (Optimal Linear Adjustment)**.** The estimator $\widehat{\theta}(\widehat{\gamma}) = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p$ is *efficient* for design $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ and covariates $h(X)$ if $\widehat{\gamma} \xrightarrow{p} \gamma^*$ for an optimal adjustment coefficient

$$\gamma^* \in \operatorname*{argmin}_{\gamma \in \mathbb{R}^{d_h}} E\Big[\mathrm{Var}(b - \gamma' h|\psi)\Big]$$

In particular, $V(\gamma^*) = \min_{\gamma \in \mathbb{R}^{d_h}} V(\gamma)$.

Note that efficiency is defined *relative* to a design $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ and covariates $h(X)$. Since setting $\gamma = 0$ gives the difference-of-means estimator $\widehat{\theta}$, any optimal estimator is in particular weakly more efficient than $\widehat{\theta}$.

Remark 3.5 (Conditional Variance Minimization)**.** Our analysis shows that treatment arm imbalances in either (1) covariates $h(X_i)$ or (2) potential outcomes $Y_i(d)$ that are

predictable by $\psi(X)$ do not contribute to first-order asymptotic variance under the stratified design $D_{1:n} \sim \text{Loc}(\psi, p)$. Because of this, the optimal covariate-adjusted estimator $\widehat{\theta} - \gamma^*(\bar{h}_1 - \bar{h}_0)c_p$ minimizes the *conditional* variance $E[\text{Var}(b - \gamma' h|\psi)]$, rather than marginal variance $\text{Var}(b - \gamma' h)$ targeted by the Lin estimator.

Remark 3.6 (Non-Uniqueness)**.** If $E[\text{Var}(h|\psi)] \succ 0$, then the optimal adjustment coefficient in Definition 3.4 is given uniquely by the "conditional" OLS coefficient

$$\gamma^* = E[\text{Var}(h|\psi)]^{-1} E[\text{Cov}(h, b|\psi)]$$

In general, γ^* may not be unique. In particular, this occurs whenever $h(x)$ includes Lipschitz functions of the stratification variables $\psi(x)$. For example, if $h(x) = (z(\psi), w(x))$ then we have

$$E[\text{Var}(b - \gamma_z' z - \gamma_w' w|\psi)] = E[\text{Var}(b - \gamma_w' w|\psi)] \qquad \forall \gamma_z \in \mathbb{R}^{d_z}$$

Indeed, our analysis shows that $\gamma_z'(\bar{z}_1 - \bar{z}_0) = o_p(n^{-1/2})$ for any coefficient γ_z and Lipschitz function $\psi \to z(\psi)$. Intuitively, since the covariate $z(\psi)$ is already finely balanced by the design, ex-post adjustment by $z(\psi)$ cannot improve first-order efficiency.

3.2 Efficient Adjustment and Semiparametric Regression

This section shows that optimal linear adjustment in stratified experiments (Definition 3.4) is asymptotically as efficient as doubly-robust *semiparametric* adjustment in experiments with iid treatments. In particular, we show first-order asymptotic equivalence of the following (design, estimator) pairs

$$(D_{1:n} \sim \text{Loc}(\psi, p), \text{efficient linear}) \iff (D_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p), \text{oracle semiparametric})$$

To define the latter, let $\mathcal{G}$ be a vector space of functions and suppose $E[Y(d)|\psi], E[h|\psi] \in \mathcal{G}$. For $d = 0, 1$, consider the oracle semiparametric regression model

$$(g_d, \gamma_d) = \underset{g \in \mathcal{G}, \gamma \in \mathbb{R}^{d_h}}{\text{argmin}} \; E[(Y(d) - g(\psi) - \gamma' h)^2]$$

Define the covariate adjustment $f_d(x) = g_d(\psi(x)) + \gamma_d' h(x)$ and consider a Robins and Rotnitzky (1995) style augmented inverse propensity weighting (AIPW) estimator

$$\widehat{\theta}_5 = E_n[f_1(X_i) - f_0(X_i)] + E_n\left[\frac{D_i(Y_i - f_1(X_i))}{p}\right] - E_n\left[\frac{(1 - D_i)(Y_i - f_0(X_i))}{1 - p}\right]$$

The next theorem shows that optimal linear adjustment of the design $D_{1:n} \sim \text{Loc}(\psi, p)$ is asymptotically equivalent to oracle semiparametric adjustment, with non-parametric control over $\psi(X)$.

Theorem 3.7. *Suppose* $D_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. *Then* $\sqrt{n}(\widehat{\theta}_5 - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

$$V^* = \text{Var}(c(X)) + \min_{\gamma \in \mathbb{R}^{d_h}} E\left[\text{Var}(b - \gamma' h|\psi)\right] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1 - p}\right]$$

The proof is given in Section 8.2 of the appendix. In particular, note that the semiparametric adjustment variance $V^* = \min_\gamma V(\gamma)$, the minimal variance in Definition 3.4. This

shows that stratification plus efficient linear adjustment can be as efficient as semiparametric AIPW estimation.

The next two sections show how to construct linearly-adjusted estimators for the design $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ that achieve the optimal variance V^*.

3.3 Efficiency by Rich Strata Controls

Example 3.2 showed that in general the Lin estimator is inefficient for non-trivial stratifications ($\psi \neq 1$). This section provides a "rich covariates" style condition on relationship between covariates and stratification variables under which a simple parametric correction of the Lin estimator recovers full efficiency. The basic idea is to include rich enough functions of the stratification variables in the adjustment set. Theorem 3.9 below shows that this forces the Lin estimator to solve the correct conditional variance minimization problem in Definition 3.4.

Consider adjusting for covariates $h(X) = (w(X), z(\psi))$. The following assumption requires that the conditional mean $E[w|\psi]$ is well-approximated by transformations $z(\psi)$ of the stratification variables.

Assumption 3.8. *There exist $c \in \mathbb{R}^{d_w}$ and $\Lambda \in \mathbb{R}^{d_w \times d_z}$ such that $E[w|\psi] = c + \Lambda z(\psi)$.*

Our next theorem shows that adding such transformations $z(\psi)$ to the adjustment set recovers full efficiency for the Lin estimator.

Theorem 3.9. *Suppose Assumptions 3.8 and 8.2 hold. Fix adjustment set $h(x) = (w(x), z(\psi))$. Then the Lin estimator $\widehat{\theta}_L$ is fully efficient for the design $D_{1:n} \sim \mathrm{Loc}(\psi, p)$. In particular, $\sqrt{n}(\widehat{\theta}_L - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$*

$$V^* = \mathrm{Var}(c(X)) + \min_{\gamma \in \mathbb{R}^{d_h}} E\left[\mathrm{Var}(b - \gamma' h|\psi)\right] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

Moreover, we have

$$\min_{\gamma \in \mathbb{R}^{d_h}} E[\mathrm{Var}(b - \gamma' h|\psi)] = \min_{\gamma \in \mathbb{R}^{d_w}} E[\mathrm{Var}(b - \gamma' w|\psi)]$$

In practice, Theorem 3.9 suggests including flexible functions $z(\psi)$ of the stratification variables in the adjustment set. The proof is given in Section 8.3.

Remark 3.10 (Indirect Efficiency Gain)**.** The second statement of the theorem shows that the Lin estimator is as efficient as optimal adjustment for the subvector $w(X) \subseteq (w(X), z(\psi))$. In this sense, the efficiency improvement due to including $z(\psi)$ is *indirect*: The efficiency improvement arises solely from the inclusion of $z(\psi)$ changing the coefficient on covariates $w(X)$. Indeed, our analysis shows that $\widehat{\theta} - \alpha'(\bar{z}_1 - \bar{z}_0) = \widehat{\theta} + o_p(n^{-1/2})$ for any $\alpha \in \mathbb{R}^{d_z}$, so adjustment by $z(\psi)$ alone cannot affect the first-order asymptotic variance. Intuitively, we are just using $z(\psi)$ as a device to "tilt" the coefficient on $w(X)$, forcing the Lin estimator to solve the correct *conditional* variance optimization problem.

Remark 3.11 (Frisch-Waugh)**.** The optimal coefficient $\gamma^* = \mathrm{argmin}_\gamma E[\mathrm{Var}(b - \gamma' w|\psi)]$ is generally not equal to the marginal variance minimizer $\gamma_L = \mathrm{argmin}_\gamma \mathrm{Var}(b - \gamma' w)$

estimated by the Lin regression. However, if $E[w|\psi] = c + \Lambda z(\psi)$ then one can show by Frisch-Waugh style reasoning that

$$\begin{aligned}\operatorname*{argmin}_{\gamma}\left[\min_{\alpha}\operatorname{Var}(b-\alpha' z-\gamma' w)\right] &= \operatorname*{argmin}_{\gamma}\operatorname{Var}\left(b-\gamma'(w-E[w|\psi])\right)\\ &= \operatorname*{argmin}_{\gamma} E[\operatorname{Var}(b-\gamma' w|\psi)] = \gamma^*\end{aligned}$$

The next example applies Theorem 3.9 to show that including a leave-one-out set of strata indicators as covariates in the Lin estimator guarantees full efficiency in the case of *coarse* stratification.

Example 3.12 (Coarse Stratification)**.** Consider stratified randomization $D_{1:n} \sim \text{Loc}(S, p)$ with fixed strata $S(x) \in \{1, \ldots, m\}$. Let $h(x) = (w(x), z(s))$ with leave-one-out strata indicators $z(S_i) = (\mathbb{1}(S_i = k))_{k=1}^{m-1}$. In this case, it's easy to see that $E[w|S] = c + \Lambda z$ for the constant $c = E[w|S = m]$ and matrix $\Lambda_{jk} = (E[w_j|S = k] - E[w_j|S = m])_{jk}$, $1 \le j \le d_w$ and $1 \le k \le m-1$. By Theorem 3.9, the coefficient $\widehat{\theta}_L$ on D_i in the Lin regression with treatment-strata interactions

$$Y_i \sim (1, w_i, z_i) + D_i(1, w_i, z_i)$$

is fully efficient. Then $\sqrt{n}(\widehat{\theta}_L - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

$$V^* = \operatorname{Var}(c(X)) + \min_{\gamma} E\left[\operatorname{Var}(b-\gamma' w|S)\right] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

In particular, Lin (2013) adjustment is weakly more efficient than difference-of-means estimation. If $E[\operatorname{Cov}(w, b|S)] \neq 0$, so that covariates $w(X)$ are predictive of potential outcomes (and thus $b(X)$) conditionally on $S(X)$, it is strictly more efficient.

Remark 3.13 (Fine Stratification)**.** Note that the Example 3.12 only applies to coarse stratification, where the strata $S(x) \in \{1, \ldots, m\}$ are data-independent and fixed as $n \to \infty$. For fine stratification $D_{1:n} \sim \text{Loc}(\psi, p)$ with continuous covariates $\psi(x)$, the strata are data-dependent and $m \asymp n$, so Theorem 3.22 does not apply. Indeed, note that for matched pairs, for instance, the Lin regression in Example 3.12 with strata-treatment interactions would have $n + 2\dim(h) > n$ covariates, producing collinearity. We give new estimators and analysis in Section 3.4 that overcome such difficulties, allowing fully efficient adjustment under weak assumptions for any design in the class $D_{1:n} \sim \text{Loc}(\psi, p)$.

3.4 Generic Efficient Adjustment

In this section, we provide several covariate-adjusted estimators that are fully efficient under weak conditions (without imposing a Assumption 3.8). We begin by considering non-interacted regression adjustment with full strata fixed effects, a popular estimation strategy in applied work. We show that this estimator is fully efficient in the case of matched pairs or experiments with limited treatment effect heterogeneity, but not in general. Next, we show that the following estimators are asymptotically fully efficient under weak assumptions:

(1) A within-stratum partialled version of the Lin (2013) estimator.

(2) "Grouped OLS", generalizing a proposal of Imbens and Rubins (2015).

(3) Tyranny-of-the-minority (ToM) adjustment for stratified designs.

3.4.1 Strata Fixed Effects Estimator

Recall that for $p = a/k$, a finely stratified design $D_{1:n} \sim \text{Loc}(\psi, p)$ partitions the experimental units $\{1, \dots, n\}$ into n/k disjoint groups g. Define the fixed effects estimator $\widehat{\theta}_1$ by the least squares equation

$$Y_i = \widehat{\theta}_1 D_i + \widehat{\gamma}_1' h_i + \sum_{g=1}^{n/k} \widehat{a}_g \mathbb{1}(g_i = g) + e_i \tag{3.4}$$

This estimator is numerically equivalent to difference-of-means if we do not include covariates h_i. However, in general $\widehat{\theta}_1$ is not equivalent to the regression $Y_i \sim (1, D_i, h_i)$, and the strata fixed effects play an important role in optimizing the adjustment coefficient $\widehat{\gamma}_1$ to the experimental design. Remark 3.19 below explains this effect in the context of the partialled Lin estimator, developed below. Before stating our results, first consider the following assumption.

Assumption 3.14. *Conditional variance* $E[\text{Var}(h|\psi)] \succ 0$.

Remark 3.15 (Conditional Variance Assumption)**.** Assumption 3.14 rules out including any function $z(\psi)$ of the stratification variables in the adjustment set. To see why this is necessary, note that in Equation 3.4, such a variable $z(\psi_i)$ would be asymptotically collinear[7] with the full set of strata indicators. Because of this, adjustment by such variables needs to be handled slightly differently. Note that by the properties of fine stratification, the estimator $\widehat{\theta} - \alpha'(\bar{z}_1 - \bar{z}_0) = \widehat{\theta} - o_p(n^{-1/2})$ for any $\alpha \in \mathbb{R}^{d_z}$, so adjustment by $z(\psi)$ cannot directly affect asymptotic efficiency. Nevertheless, one may still wish to adjust by $z(\psi)$ to correct for finite sample imbalances in the stratification variables. For expositional purposes, we maintain Assumption 3.14 throughout this section but remove it in Section 3.5 below, slightly modifying the estimators.

The next theorem shows that $\widehat{\theta}_1$ is fully efficient in the case of matched pairs ($p = 1/2$) or if treatment effect heterogeneity is limited, but may be inefficient in general.

Theorem 3.16. *Suppose Assumptions 3.14 and 8.2 hold. The estimator has representation* $\widehat{\theta}_1 = \widehat{\theta} - \widehat{\gamma}_1'(\bar{h}_1 - \bar{h}_0) + O_p(n^{-1})$. *If* $D_{1:n} \sim \text{Loc}(\psi, p)$ *then* $\sqrt{n}(\widehat{\theta}_1 - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

$$V = \text{Var}(c(X)) + E[\text{Var}(b - c_p^{-1}\gamma_1' h|\psi)] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

The coefficient $\widehat{\gamma}_1 \xrightarrow{p} \gamma_1$ *with* $c_p^{-1}\gamma_1 = \text{argmin}_{\gamma \in \mathbb{R}^{d_h}} E[\text{Var}(f - \gamma' h|\psi)]$ *and target function* $f(x) = m_1(x)\sqrt{\frac{p}{1-p}} + m_0(x)\sqrt{\frac{1-p}{p}}$. *The estimator is fully efficient if one of the following*

(a) $p = 1/2$ *(matched pairs).*

(b) $E[Y_i(1) - Y_i(0)|h_i, \psi_i] = E[Y_i(1) - Y_i(0)|\psi_i]$.

(c) $E[\text{Cov}(Y_i(1) - Y_i(0), h(X)|\psi(X))] = 0$.

[7] More precisely, there exist vectors $\alpha^n \in \mathbb{R}^{n/k}$ such that $E_n[(z_i - \sum_g \alpha_g^n \mathbb{1}(g_i = g))^2] = o_p(1)$.

Asymptotically exact inference for the fixed effects estimator above under fine stratification will be given in Section 4. See Section 8.4 for the proof.

Remark 3.17 (Conditions for Full Efficiency)**.** If $p \neq 1/2$, then the target function $f(x) \neq b(x)$. Because of this, the fixed effects estimator solves the wrong variance minimization problem in general. The last statement of the theorem gives conditions such that $\operatorname{argmin}_\gamma E[\operatorname{Var}(f - \gamma' h|\psi)] = \operatorname{argmin}_\gamma E[\operatorname{Var}(b - \gamma' h|\psi)]$, in which case $\widehat{\theta}_1$ is fully efficient. If $p \neq 1/2$ then condition (c) is both necessary and sufficient for full efficiency. It requires that the covariates h_i have no (linear) predictive power for treatment effects $Y_i(1) - Y_i(0)$ conditional on the stratification variables. Condition (b) requiring that the conditional average treatment effect $c(X)$ only varies with the stratification variables $\psi(X)$ and implies condition (c). The weaker mean-independence condition $E[Y_i(1) - Y_i(0)|\psi_i, h_i] = E[Y_i(1) - Y_i(0)|\psi_i]$ is also sufficient.

In the next two subsections, we develop estimators that are fully efficient for any finely stratified design, without imposing strong assumptions on treatment effect heterogeneity or treatment proportions.

3.4.2 Partialled Lin Estimator

First, we define a within-group partialled version of the Lin (2013) estimator. Let g_i denote the group that unit i belongs to in the stratified design $D_{1:n} \sim \text{Loc}(\psi, p)$. Define the partialled covariates $\check{h}_i$ to be the residuals from a fixed effects regression

$$h_i = \sum_{g=1}^{n/k} \widehat{a}_g \mathbb{1}(g_i = g) + \check{h}_i \qquad E_n[\check{h}_i \mathbb{1}(g_i = g)] = 0 \ \ \forall g \tag{3.5}$$

This gives $\check{h}_i = h_i - k^{-1} \sum_{j \in g(i)} h_j$. For example, if $k = 2$ then this is just the within-pair covariate difference $\check{h}_i = (1/2)(h_i - h_{m(i)})$, where i is matched to $m(i)$. Next, we use these partialled covariates in the Lin regression

$$Y_i = \widehat{c} + \widehat{\theta}_2 D_i + \widehat{a}_0 \check{h}_i + \widehat{a}_1 D_i \check{h}_i + e_i \tag{3.6}$$

Define the *partialled* Lin estimator $\widehat{\theta}_2$ to be the coefficient on D_i in this regression. Theorem 3.22 below shows that both this estimator and the grouped OLS estimator are fully efficient in the sense of Definition 3.4. The proof also shows that we may write

$$\widehat{\theta}_2 = \widehat{\theta} - \widehat{\gamma}_2'(\bar{h}_1 - \bar{h}_0)c_p \qquad \widehat{\gamma}_2 = (\widehat{a}_1 + \widehat{a}_0)\sqrt{\frac{1-p}{p}} + \widehat{a}_0\sqrt{\frac{p}{1-p}}$$

This representation is used in our inference methods in Section 4.

Remark 3.18 (Treatment-Strata Interactions)**.** Alternatively, one may try to run a Lin regression that also includes the leave-one-out strata fixed effects $z_i^n = (\mathbb{1}(g(i) = g))_{i=1}^{n/k-1}$, as in Example 3.12. We would like to run the regression

$$Y_i \sim (1, h_i, z_i^n) + D_i(1, h_i, z_i^n)$$

By counting regressors, we require $\dim(h) \leq n(1/2 - k^{-1})$ for non-collinearity when $p = a/k$. This is violated for matched pairs designs, which would have $n + 2\dim(h) > n$

regressors. In fact, the regressors above are always collinear if either $a = 1$ or $a = k - 1$. To see the problem, one can show that in contrast to Equation 3.5, this estimator partials covariates h_i separately in each treatment arm, using $\check{h}_i = h_i - a^{-1} \sum_{j \in g(i)} h_j D_j$ if $D_i = 1$ and $\check{h}_i = h_i - (k-a)^{-1} \sum_{j \in g(i)} h_j (1 - D_j)$ if $D_i = 0$. If $a = 1$, for instance, then $\check{h}_i = h_i - h_i = 0$ for all i.

When feasible, our analysis shows that such an estimator is asymptotically equivalent to the partialled Lin estimator above. However, we expect worse finite sample properties because of noisier within-arm partialling.

Remark 3.19 (Intuition for Optimality)**.** Theorem 3.3 showed that an adjusted estimator $\widehat{\theta} - \widehat{\gamma}(\bar{h}_1 - \bar{h}_0)c_p$ is fully efficient if $\widehat{\gamma} \xrightarrow{p} \gamma^*$ and γ^* makes the optimal *conditional* variance vs. covariance tradeoff

$$\gamma^* \in \operatorname*{argmin}_{\gamma} E[\operatorname{Var}(b - \gamma' h | \psi)] = \operatorname*{argmin}_{\gamma} -2\gamma' E[\operatorname{Cov}(h, b | \psi)] + \gamma' E[\operatorname{Var}(h|\psi)]\gamma$$

Intuitively, by using the within-stratum partialled regressors $\check{h}_i = h_i - k^{-1} \sum_{j \in g(i)} h_j$, we force the Lin estimator to only use the covariate signal that is orthogonal to the stratification variables. More precisely, the proof of Theorem 3.22 shows that under the design $D_{1:n} \sim \text{Loc}(\psi, p)$ we have $E_n[\check{h}_i \check{h}_i'] \asymp E[\operatorname{Var}(h|\psi)] + o_p(1)$ and $E_n[\check{h}_i Y_i(d)] \asymp E[\operatorname{Cov}(h, Y(d)|\psi)] + o_p(1)$. Because of this, the partialled Lin estimator solves the optimal conditional variance problem, instead of just minimizing the marginal variance.

3.4.3 Grouped OLS Estimator

Next, we generalize a proposal of Imbens and Rubin (2015) for matched pairs experiments to general stratified designs. For each group of units g in the design $D_{1:n} \sim \text{Loc}(\psi, p)$, define the within-group difference-of-means for outcomes and covariates

$$y_g = \frac{1}{k} \sum_{i \in g} \frac{Y_i D_i}{p} - \frac{1}{k} \sum_{i \in g} \frac{Y_i (1 - D_i)}{1 - p} \qquad h_g = \frac{1}{k} \sum_{i \in g} \frac{h_i D_i}{p} - \frac{1}{k} \sum_{i \in g} \frac{h_i (1 - D_i)}{1 - p}$$

For any group-indexed array $(x_g)_g$, denote $E_g[x_g] = \frac{k}{n} \sum_g x_g$. Define the *grouped OLS* estimator $\widehat{\theta}_3$ by the regression

$$y_g = \widehat{\theta}_3 + \widehat{\gamma}_3 h_g + e_g \qquad E_g[(1, h_g) e_g] = 0 \tag{3.7}$$

For motivation, if we exclude covariates Equation 3.7 becomes $y_g = \widehat{\theta} + e_g$ with $E_g[e_g] = 0$. Then $\widehat{\theta}$ is just the difference-of-means estimator defined above since

$$\widehat{\theta} = E_g[y_g] = \frac{k}{n} \sum_g \left[\frac{1}{k} \sum_{i \in g} \frac{Y_i D_i}{p} - \frac{1}{k} \sum_{i \in g} \frac{Y_i (1 - D_i)}{1 - p} \right] = \bar{Y}_1 - \bar{Y}_0$$

By OLS properties, the adjusted version can be written

$$\widehat{\theta}_3 = E_g[y_g] - \widehat{\gamma}_3' E_g[h_g] = \widehat{\theta} - \widehat{\gamma}_3' (\bar{h}_1 - \bar{h}_0)$$

with coefficient $\widehat{\gamma}_3 = \operatorname{Var}_g(h_g)^{-1} \operatorname{Cov}_g(h_g, y_g)$.

Remark 3.20 (Intuition for Efficiency)**.** The grouped OLS estimator uses within-group differences of covariates $\bar{h}_{g1} - \bar{h}_{g0}$ to predict within-group outcome differences $\bar{Y}_{1g} - \bar{Y}_{0g}$. Intuitively, by doing this we only use the variation in covariates and potential outcomes that is orthogonal to the stratification variables. This forces least squares to compute a *conditional* variance-covariance tradeoff, solving the optimal adjustment problem in Definition 3.4. The proof of Theorem 3.22 shows that if $D_{1:n} \sim \text{Loc}(\psi, p)$ then as $n \to \infty$

$$\widehat{\gamma}_3 = \text{Var}_g(h_g)^{-1} \text{Cov}_g(h_g, y_g) \xrightarrow{p} c_p \operatorname*{argmin}_{\gamma} E[\text{Var}(b - \gamma' h | \psi)]$$

Remark 3.21. Imbens and Rubin (2015) propose a special case of this estimator for matched pairs experiments ($p = 1/2$). Their analysis uses a toy sampling model where the pairs themselves are drawn "pre-matched" from a super-population. By contrast, we model the experimental units as being sampled from a super-population, matching units into data-dependent groups post-sampling. As noted above, this more realistic sampling experiment complicates the analysis, producing different limiting variances and requiring different inference procedures. In a design-based setting, with non-random covariates and potential outcomes, Fogarty (2018) shows for the case of matched pairs that grouped OLS is weakly more efficient than difference-of-means. By contrast, we show asymptotic optimality in a superpopulation framework, over a much larger family of stratifications.

3.4.4 Tyranny-of-the-Minority (ToM) Regression

Finally, we define tyranny-of-the-minority (ToM) regression, extending Lin (2013). Let $\widehat{\gamma}_4$ be the coefficient in the regression $Y_i^w = \widehat{\gamma}_4' \check{h}_i + e_i$ with weighted outcomes

$$Y_i^w = D_i Y_i \frac{(1-p)^{1/2}}{p^{3/2}} + (1 - D_i) Y_i \frac{p^{1/2}}{(1-p)^{3/2}} \tag{3.8}$$

Define the ToM estimator $\widehat{\theta}_4 = \widehat{\theta} - \widehat{\gamma}_4'(\bar{h}_1 - \bar{h}_0)c_p$.

Our main theorem shows that all three estimators are asymptotically fully efficient (Definition 3.4) for any stratified design $D_{1:n} \sim \text{Loc}(\psi, p)$.

Theorem 3.22. *Suppose Assumptions 3.14 and 8.2 hold. If* $D_{1:n} \sim \text{Loc}(\psi, p)$*, then* $\widehat{\theta}_2 - \widehat{\theta}_k = o_p(n^{-1/2})$ *for* $k = 3, 4$*. We have* $\sqrt{n}(\widehat{\theta}_2 - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$

$$V^* = \text{Var}(c(X)) + \min_{\gamma} E[\text{Var}(b - \gamma' h | \psi)] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

Methods for asymptotically exact inference on the ATE using these estimators are discussed in Section 4 below. The proof is given in Section 8.4 of the appendix.

3.5 Further Adjustment for Stratification Variables

In this section give versions of the fixed effects, partialled Lin, and grouped OLS estimators that allow for further adjustment by covariates $z(\psi)$ that are functions of the stratification variables. As noted in Remark 3.15 above, this cannot affect first-order efficiency but may improve finite sample performance by correcting for any remaining imbalances in $\psi(X)$.

Denote $z_i = z(\psi_i)$. We replace the fixed effects estimator in Equation 3.4 by the coefficient $\widehat{\tau}_1$ on D_i in the regression

$$Y_i \sim (1, D_i, \check{h}_i, z_i) \tag{3.9}$$

Define the partialled Lin estimator $\widehat{\tau}_2$ to be the coefficient on D_i in the regression

$$Y_i \sim (1, \check{h}_i, z_i) + D_i(1, \check{h}_i, z_i) \tag{3.10}$$

Let $\widehat{\theta}_3$ be the coefficient from the grouped OLS regression $y_g = \widehat{\theta}_3 + \widehat{\gamma}_3' w_g + e_g$ with $E_g[e_g(1, w_g)] = 0$ as above. Let $\widehat{\alpha}_3$ be the coefficient from the OLS equation

$$c_p(Y_i^w - \widehat{\gamma}_3' h_i) = \widehat{c} + \widehat{\alpha}_3' z_i + e_i \qquad E_n[e_i(1, z_i)] = 0$$

with weighted outcomes Y_i^w defined above. Define the modified grouped OLS estimator

$$\widehat{\tau}_3 = \widehat{\theta}_3 - \widehat{\alpha}_3'(\bar{z}_1 - \bar{z}_0)c_p \tag{3.11}$$

For intuition, we show in the appendix that $E[Y_i^w | X_i] = b(X_i)c_p^{-1}$, so Y_i^w is a noisy signal for $b(X_i)$. Then the OLS equation defining $\widehat{\alpha}_3$ behaves like the population regression $b(X) - w(X)'\gamma_3 \sim (1, z(\psi))$, where γ_3 is the fully efficient adjustment coefficient for $w(X)$.

Our next theorem shows that these estimators are asymptotically equivalent to the original versions above.

Theorem 3.23. *Suppose Assumptions 3.14 and 8.2 hold, as well as* $\mathrm{Var}(z) \succ 0$ *and* $E[|z|_2^2] < \infty$. *Then if* $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ *we have*

$$\widehat{\tau}_1 = \widehat{\theta}_1 + o_p(n^{-1/2}) \qquad \widehat{\tau}_2 = \widehat{\theta}_2 + o_p(n^{-1/2}) \qquad \widehat{\tau}_3 = \widehat{\theta}_3 + o_p(n^{-1/2})$$

Each estimator has the form $\widehat{\tau}_k = \widehat{\theta}_k - \widehat{\alpha}_k'(\bar{z}_1 - \bar{z}_0)c_p$ *with* $\widehat{\alpha}_1 \xrightarrow{p} \mathrm{argmin}_\alpha \mathrm{Var}(f - \alpha' z)$, $\widehat{\alpha}_2 \xrightarrow{p} \mathrm{argmin}_\alpha \mathrm{Var}(b - \alpha' z)$ *and* $\widehat{\alpha}_3 \xrightarrow{p} \mathrm{argmin}_\alpha \mathrm{Var}(b - c_p^{-1}\gamma_3' h - \alpha' z)$ *and* f *as in 3.16.*

The first statement shows that the stratification variable adjusted versions $\widehat{\tau}_k$ have the same first-order asymptotic properties as $\widehat{\theta}_k$ for $k = 1, 2, 3$, described in Theorems 3.16 and 3.22. From the second statement, we can interpret the adjustment coefficients $\widehat{\alpha}_k$ as taking a conservative approach where we ignore the stratification and adjust for covariates $z(\psi)$ as if the experiment were completely randomized.

4 Inference

This section introduces new methods for inference on the ATE in stratified experiments. We show how to use any linearly-adjusted estimator to construct confidence intervals for the ATE that are asymptotically exact in the sense that coverage probabilities converge to the specified nominal level, without overcoverage.

Overcoverage is known to be a problem in stratified experiments when inference is based on the usual Eicker-Huber-White (EHW) variance estimator. In the unadjusted case, for example, EHW variance estimation based on the regression $Y \sim 1 + D$ is generically conservative unless $\psi = 1$, even after including a full set of strata fixed effects (Bai et al. (2021)). To the best of our knowledge, we give the first asymptotically non-conservative

variance estimators for covariate-adjusted ATE estimation.

All the estimators considered in this paper have a representation of the form

$$\widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p + O_p(n^{-1})$$

for some adjustment coefficient $\widehat{\gamma} \xrightarrow{p} \gamma$. Our main inference result in Theorem 4.2 applies to any covariate adjustment of this form, enabling asymptotically exact inference using any of the estimators mentioned in this paper.

Before continuing, we discuss asymptotically exact inference for unadjsuted estimation. The procedure introduced in Cytrynbaum (2021) extends the "pairs-of-pairs" idea from Abadie and Imbens (2008). In particular, we (1) form centroids $\frac{1}{k}\sum_{i\in g}\psi_i$ for each stratum, (2) match stratum centroids into pairs and (3) match treated units to treated units and control units to control units between paired strata. This results in a bijective matching function $\mu : [n] \to [n]$ with $D_i = D_{\mu(i)}$ for all i. Crucially, the centroid-pairing step (2) ensures a tight matching, in the sense that provably $E_n[|\psi_i - \psi_{\mu(i)}|_2^2] = o_p(1)$ as $n \to \infty$.[8] We use the matching $\mu(\cdot)$ to define variance estimators in Definition 4.1 below.

Definition 4.1 (Unadjusted Variance Estimation)**.** Define the following variance estimation components

$$\widehat{v}_1 = E_n\left[\frac{D_i(1-p)}{p^2}Y_iY_{\mu(i)}\right] \qquad \widehat{v}_0 = E_n\left[\frac{(1-D_i)p}{(1-p)^2}Y_iY_{\mu(i)}\right]$$
$$\widehat{v}_{10} = 2n^{-1}\sum_{1\le i<j\le n}\frac{\mathbb{1}(g(i)=g(j))}{k}\frac{D_i(1-D_j)Y_iY_j}{p(1-p)}$$

Let the unadjusted variance estimator $\widehat{V}$ be given by

$$\widehat{V} = \mathrm{Var}_n\left(\frac{(D_i-p)Y_i}{p-p^2}\right) - \widehat{v}_1 - \widehat{v}_0 - 2\widehat{v}_{10}$$

Theorem 5.3 of Cytrynbaum (2021) shows that $\widehat{V} = V + o_p(1)$, where V is the unadjusted asymptotic variance

$$V = \mathrm{Var}(c(X)) + E[\mathrm{Var}(b|\psi)] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

Our next theorem modifies $\widehat{V}$ to account for the efficiency gains (or losses) of linear covariate adjustment.

Theorem 4.2. *Suppose Assumptions 3.14 and 8.2 hold. Let $\widehat{V}$ be as in Definition 4.1 and suppose $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ with $p = a/k$. Consider an estimator $\widehat{\theta}(\widehat{\gamma}) = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p$ with $\widehat{\gamma} \xrightarrow{p} \gamma$ and asymptotic variance*

$$V(\gamma) = \mathrm{Var}(c(X)) + E[\mathrm{Var}(b - \gamma' h|\psi)] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

[8] See the explicit construction in Cytrynbaum (2021) for details

Define $\widehat{V}_h = \frac{k}{k-1} E_n[\check{h}_i \check{h}_i']$ *and* $\widehat{V}_b = \frac{k}{k-1} E_n[\check{h}_i Y_i^w]$*, with* Y_i^w *the weighted outcomes given in Equation 3.8. Construct the adjusted variance estimator*

$$\widehat{V}(\gamma) = \widehat{V} - 2\widehat{\gamma}'\widehat{V}_b + \widehat{\gamma}'\widehat{V}_h\widehat{\gamma} \tag{4.1}$$

Then $\widehat{V}(\gamma) \xrightarrow{p} V(\gamma)$*.*

See Section 8.5 for the proof.[9]

By Theorem 4.2, the confidence interval

$$\widehat{C} = \left[\widehat{\theta}(\widehat{\gamma}) - \frac{\widehat{V}(\gamma)^{1/2}}{\sqrt{n}} c_{1-\alpha/2},\ \widehat{\theta}(\widehat{\gamma}) + \frac{\widehat{V}(\gamma)^{1/2}}{\sqrt{n}} c_{1-\alpha/2}\right] \tag{4.2}$$

is asymptotically exact in the sense that $P(\text{ATE} \in \widehat{C}) = 1 - \alpha + o(1)$. We document the finite sample properties of this confidence interval in the simulations in Section 5 and empirical application in Section 6 below.

5 Simulations

In this section, we use simulations to examine the finite sample performance of the estimators studied above. We consider quadratic outcome models of the form

$$Y_i(d) = \psi_i' Q_d \psi_i + \psi_i' L_d + c_d \cdot u_i + \epsilon_i^d \qquad E[\epsilon_i^d | X_i] = 0$$

for $d \in \{0, 1\}$. The component $u_i = u(X_i)$ represents covariate signal that is independent of the stratification variables $\psi(X_i)$. After implementing the design $D_{1:n} \sim \text{Loc}(\psi, p)$, we receive access to scalar covariates h_i that are correlated with both ψ_i and $Y_i(d)$. In particular, suppose that

$$h_i = \psi_i' Q_h \psi_i + \psi_i' L_h + u_i \qquad E[u_i | \psi_i] = 0$$

In the following simulations, we let $\psi_i \sim N(0, I_m)$, $u_i \sim N(0, 1)$, and $\epsilon_i^d \sim N(0, 1/10)$ with $(\psi_i, u_i, \epsilon_i^d)$ jointly independent. We use treatment proportions $p = 2/3$ unless otherwise specified. With $m \equiv \dim(\psi)$, let $A \in \mathbb{R}^{m \times m}$ have $A_{ij} = 1$ for $i \neq j$ and $A_{ii} = 0$. We simulate the following DGP's:

> **Model** 1: Quadratic coefficients $Q_h = (1/m^2)A$ and $Q_0 = Q_1 = (1/m)A$. Linear coefficients $L_0 = \mathbb{1}_m$, $L_1 = 2\mathbb{1}_m$, $L_h = \mathbb{1}_m$. Regressor signal $c_1 = c_0 = -3$.
>
> **Model** 2: As in Model 1 but $c_0 = -4$ and $c_1 = -1$.
>
> **Model** 3: As in Model 2 but $p = 1/2$.
>
> **Model** 4: As in Model 1 but $c_0 = 2$ and $c_1 = 4$.
>
> **Model** 5: As in Model 1 but $c_0 = 2$ and $c_1 = 4$ and $p = 1/2$.
>
> **Model** 6: As in Model 1 but $Q_h = (1/100)A$.

[9] The proof shows in particular that $\widehat{V}_b \xrightarrow{p} E[\text{Cov}(h, b|\psi)]$ and $\widehat{V}_h \xrightarrow{p} E[\text{Var}(h|\psi)]$.

We begin by comparing the efficiency properties of different linearly adjusted estimators. **Unadj** refers to simple difference-of-means (unadjusted). The **Lin** estimator is studied in Theorem 3.1. **Naive** refers to the non-interacted regression $Y \sim (1, D, h)$, (Theorem 8.1). **FE** refers to the fixed effects estimator (Theorem 3.16) and **Plin** the partialled Lin estimator (Theorem 3.22). **GO** refers to Grouped OLS and **ToM** refers to Tyranny-of-the-Minority estimation (Theorem 3.22). **Strata Controls** refer to alternative versions of each of the previous estimators that further adjust for $z(\psi)$, as discussed in Section 3.5. Here, we set $z(\psi) = \psi$. All results are calculated using 1000 Monte Carlo repetitions.

Table 1: Ratio of MSE's (%) for adjusted vs. unadjusted estimation.

		No Strata Controls							Strata Controls					
$(n, \dim(\psi))$	Model	Unadj	Naive	Lin	FE	Plin	GO	ToM	Naive	Lin	FE	Plin	GO	ToM
(600, 2)	1	100	115.0	103.0	27.3	27.3	27.2	102.8	29.5	29.4	19.9	21.1	21.2	21.1
	2	100	126.8	102.3	48.4	39.7	39.7	101.8	58.4	41.6	42.1	34.4	34.7	34.6
	3	100	114.5	114.5	46.5	46.5	46.5	114.3	50.6	50.7	42.1	42.1	42.2	41.9
	4	100	21.2	27.8	21.8	18.1	18.1	27.9	17.6	20.3	25.7	20.8	20.9	21.0
	5	100	28.2	28.2	17.7	17.7	17.7	28.1	20.6	20.6	20.1	20.1	20.1	19.9
	6	100	100.1	100.2	16.4	16.4	16.4	99.8	6.9	6.8	11.1	11.0	11.2	11.0
(1200, 2)	1	100	114.7	102.9	20.8	20.5	20.7	102.7	28.5	28.5	16.2	16.7	16.8	16.7
	2	100	128.9	102.4	41.3	34.0	34.1	102.2	57.5	40.0	37.7	31.1	31.1	30.9
	3	100	115.9	116.0	41.3	41.3	41.3	115.8	49.7	49.7	38.9	38.9	38.9	38.8
	4	100	22.8	29.1	26.0	20.3	20.4	29.1	19.4	21.7	29.2	22.7	22.9	22.8
	5	100	27.0	26.9	16.1	16.1	16.1	26.9	19.1	19.0	17.6	17.7	17.7	17.6
	6	100	100.1	100.1	12.4	12.4	12.4	99.9	6.8	6.8	9.3	9.3	9.3	9.3
(1200, 5)	1	100	144.5	129.1	77.6	77.5	77.5	128.9	22.4	21.8	39.1	44.4	49.7	44.4
	2	100	146.0	124.0	86.9	79.1	79.0	124.0	45.9	33.7	55.4	51.6	56.6	51.3
	3	100	137.7	137.6	80.4	80.4	80.4	137.6	41.2	41.1	54.2	54.1	59.0	54.0
	4	100	28.2	32.6	33.3	28.0	28.0	32.6	27.3	21.4	57.6	45.8	48.8	46.0
	5	100	31.9	32.0	25.0	25.0	25.0	31.9	18.2	18.1	41.7	41.6	43.0	41.5
	6	100	135.5	135.5	74.5	74.5	74.5	135.4	17.3	17.3	40.8	40.9	43.5	40.7

Table 1 studies estimator efficiency, showing the mean squared error (MSE) ratio, relative to difference-of-means, for each of the adjusted estimators above. In models 1, 2, and 3, both Naive and Lin style linear adjustment are strictly inefficient relative to unadjusted estimation. Following the discussion in Theorem 3.1, these models have marginal covariance $\text{Cov}(Y(d), h) > 0$ but $E[\text{Cov}(Y(d), h|\psi)] < 0$ conditional on the stratification variables. Because of this, the optimal adjustment coefficient $\gamma^* < 0$, while Naive and Lin regression estimate positive adjustment coefficients $\gamma_N, \gamma_L > 0$. **Plin**, **GO**, and **FE** perform well across specifications for Models 1, 2, and 3.[10] While asymptotically efficient, **ToM** performs poorly in finite samples when $p \neq 1/2$, since dividing by small propensity weights results in highly variable estimates of the optimal adjustment coefficient γ^*.

For Model 6, **Lin** with $z(\psi) = \psi$ controls is optimal. This is suggested by Theorem 3.9: since $E[w|\psi]$ is (approximately) linear in ψ, including covariates $z(\psi) = \psi$ in the Lin estimator should give (approximate) efficiency. For Models 4 and 5, the (generally inefficient) **Naive** and **Lin** methods are competitive with the generic efficient methods from Section

[10]Theorem 3.16 can be used to show that **FE** is asymptotically efficient for Models 1 and 3, but not for Model 2.

3.4. This is because in these cases both $\text{Cov}(Y(d), h) \approx E[\text{Cov}(Y(d), h|\psi)]$, so that "by chance" the optimal adjustment coefficient γ^* is close to the Naive and Lin adjustment coefficients γ_N and γ_L. However, the Naive and Lin coefficients are estimated much more precisely[11] than the optimal coefficient $\gamma^* = E[\text{Var}(h|\psi)]^{-1}E[\text{Cov}(h, b|\psi)]$. Summarizing our findings, the generic efficient methods perform well in finite samples when the Lin coefficient bias $\gamma_L - \gamma^*$ dominates the additional variance $\text{Var}(\widehat{\gamma}^*) > \text{Var}(\widehat{\gamma}_L)$ required to consistently estimate γ^*.

Table 2: Inference Methods - Power and Coverage.

		No Strata Controls							Strata Controls					
	Model	Unadj	Naive	Lin	FE	Plin	GO	TOM	Naive	Lin	FE	Plin	GO	TOM
%Δ CI Length vs. Unadj	1	0.0	19.3	12.9	-11.1	-11.3	-11.1	12.9	-48.8	-49.4	-35.1	-31.4	-28.0	-31.5
	2	0.0	20.5	11.2	-6.3	-9.7	-9.6	11.2	-31.2	-40.0	-24.6	-26.3	-23.5	-26.4
	3	0.0	16.0	16.0	-8.6	-8.6	-8.6	16.0	-33.8	-33.8	-24.5	-24.4	-22.3	-24.5
	4	0.0	-43.7	-39.8	-40.8	-44.4	-44.2	-39.9	-45.8	-50.4	-24.9	-31.6	-30.0	-31.7
	5	0.0	-41.4	-41.3	-47.2	-47.2	-47.2	-41.4	-52.7	-52.6	-34.5	-34.5	-33.1	-34.5
	6	0.0	16.1	16.1	-15.2	-15.2	-15.2	16.1	-55.6	-55.6	-35.8	-35.8	-33.1	-35.8
Coverage (Exact)	1	95.6	95	95.3	96	95.8	96	95.3	96.6	96.2	96.3	96.2	95.8	96.2
	2	96.3	95.6	96.1	96.5	96.1	96.1	96.1	95.5	96.2	95.7	96.2	95.7	96.1
	3	94.8	94	94.1	95.4	95.4	95.4	94	95.3	95.3	95	95.1	94.8	95
	4	96.5	97.5	98.1	95.9	96.4	96.9	97.9	96.8	96.9	95.1	96	95.4	96
	5	96.4	95.4	95.4	97	97	97	95.4	97.5	97.7	95.8	95.9	95.9	95.8
	6	95.7	95.4	95.4	96.1	96.1	96.1	95.4	96.7	96.6	96.3	96.3	96.1	96.4
Coverage (HC2)	1	98.6	95.5	95.8	99.1	99.1			99.9	98.4	99.3	98.5		
	2	98.8	95.7	95.9	99.4	99.5			97.9	94.6	98.3	97.1		
	3	99.1	95.8	93.7	99.5	99.5			97.6	91.2	97.8	96.3		
	4	99.5	99.3	93.1	100	100			97.4	69.7	97.6	97.3		
	5	99.2	96.5	90	100	100			96.7	65.4	98.9	98.1		
	6	99.5	97.1	96.4	99.9	99.9			99	96.3	99.3	98.9		

Table 2 reports finite sample efficiency and coverage properties of the asymptotically exact inference methods developed in Section 4. We let $n = 1200$ and $\dim(\psi) = 5$. Results for varying n are given in Section 6. The first panel shows % reduction in confidence interval length relative to unadjusted estimation. All confidence intervals are computed using Equation 4.2. We see that the relative efficiency of different estimators are reflected by our inference methods. In particular, asymptotically exact inference allows researchers to report shorter confidence intervals when a more efficient adjustment method is used. In the second panel, we show coverage probabilities for our asymptotically exact confidence interval (Equation 4.2), across a range of linearly adjusted estimators. The final panel shows coverage probabilities for confidence intervals based on the usual HC2 variance estimator, where applicable. The HC2-based confidence intervals significantly overcover.

6 Empirical Results

This section applies our methods to the 2008 Oregon Health Insurance Experiment, a large-scale public health intervention that randomized medicaid eligibility to low-income,

[11] More formally, the Lin coefficient $\widehat{\gamma}_L - \gamma_L = O_p(n^{-1/2})$ with $\gamma_L \neq \gamma^*$ in general. However, the proof of Theorem 3.22 shows that PLin, GO, and ToM are consistent for γ^* at the slower rate $\widehat{\gamma}^* - \gamma^* = O_p(n^{-2/(d_\psi+1)})$.

uninsured adults. Finkelstein et al. (2012) provide intent-to-treat and IV estimates of the effect of medicaid eligibility on a range of health and financial outcomes, including emergency department visits and welfare program (SNAP, TANF) enrollment in the followup period. We restrict our attention to single-member households participating in the March 2008 lottery drawing. The authors observe a number of baseline characteristics, including age, gender, previous emergency department visits, previous visits for a chronic condition, whether the individual participated in SNAP or TANF welfare programs and so on. Approximately 1/3 of the individuals in this wave were randomized to treatment, without stratification.

We perform an intent-to-treat analysis of the effect of medicaid eligibility on number of emergency department (ED) visits in the follow-up period. We want to compare the performance of different covariate adjusted estimators under a finely stratified design. To do this, we first impute[12] a full panel of potential outcomes $(\widehat{Y}_i(0), \widehat{Y}_i(1))_{i=1}^N$, letting $\widehat{Y}_i(d) = Y_i(d)$ if $D_i = d$ and $\widehat{Y}_i(d) = \max(0, \widehat{f}_d(X_i) + \epsilon_i^d)$ if $D_i \neq d$. For each Monte Carlo iteration we do the following: (1) draw $(\widehat{Y}_i(0), \widehat{Y}_i(1))_{i=1}^n$ with replacement from $(\widehat{Y}_i(0), \widehat{Y}_i(1))_{i=1}^N$ (2) draw treatment assignments $D_{1:n} \sim \text{Loc}(\psi, 1/3)$, (3) reveal $\widehat{Y}_i = \widehat{Y}_i(D_i)$, (4) estimate ATE using the covariate adjustment methods above, and (5) form confidence intervals using Equation 4.2. We test the following stratifications and adjustment sets:

> **Model** 1: ψ = (age, gender, any ED visit pre-lottery) and covariates $h(X)$ = (number of ED visits pre-lottery).
>
> **Model** 2: As in Model 1 but $h(X)$ = (number of ED visits pre-lottery, ever on SNAP, any ED visit for chronic condition,).
>
> **Model** 3: As in Model 1 but $h(X)$ = (number of ED visits pre-lottery, ever on TANF, amount of pre-lottery SNAP benefits).

Table 3: Ratio of MSE's (%) for adjusted vs. unadjusted estimation.

		No Strata Controls							Strata Controls					
n	Model	Unadj	Naive	Lin	FE	Plin	GO	ToM	Naive	Lin	FE	Plin	GO	ToM
600	1	100	85.0	85.9	84.8	85.5	85.5	84.8	84.9	86.9	85.1	86.2	85.8	86.4
	2	100	79.9	81.7	79.6	81.1	79.9	80.9	79.7	82.4	79.6	81.3	79.9	83.0
	3	100	81.7	82.8	81.9	83.4	82.9	82.1	81.8	83.6	82.3	83.9	83.3	85.7
900	1	100	85.0	85.9	84.8	85.5	85.5	84.8	84.9	86.9	85.1	86.2	85.8	86.4
	2	100	79.9	81.7	79.6	81.1	79.9	80.9	79.7	82.4	79.6	81.3	79.9	83.0
	3	100	81.7	82.8	81.9	83.4	82.9	82.1	81.8	83.6	82.3	83.9	83.3	85.7
1200	1	100	88.4	89.1	88.6	88.8	88.7	88.3	88.1	88.9	88.5	88.9	88.6	89.2
	2	100	82.0	82.8	81.6	82.4	81.9	82.5	82.0	83.2	81.4	82.3	81.7	83.9
	3	100	82.4	83.7	82.0	83.4	82.9	83.7	82.2	83.8	81.7	83.0	82.6	84.8

Table 3 shows the efficiency of various estimators (as in Section 5). The efficiency differences between adjustment strategies are minor. To see why, one can compute that for

[12] We fit $\widehat{f}_d(x)$ with LASSO separately in each arm, using the baseline covariates above and all of their two-way interactions. We draw $\epsilon_i^d \sim \mathcal{N}(0, \widehat{\sigma}_d^2)$, with $\widehat{\sigma}_d^2$ a residual variance estimate.

this DGP the optimal adjustment coefficient $\gamma^* \approx 2.2$, while the (sub-optimal) population Lin coefficient $\gamma_L \approx 2.4$. However, our estimate of the optimal coefficient is less precise, $\text{Var}(\widehat{\gamma}^*) > \text{Var}(\widehat{\gamma}_L)$. Since the bias $\gamma^* - \gamma_L$ is small, in finite samples a precise estimate of the sub-optimal coefficient is more efficient than an imprecise estimate of the optimal coefficient.

Table 4: Inference Methods - Power and Coverage.

		No Strata Controls							Strata Controls					
	n	Unadj	Naive	Lin	FE	Plin	GO	ToM	Naive	Lin	FE	Plin	GO	ToM
%Δ CI Length vs. Unadj	300	0.0	-8.9	-9.2	-9.3	-11.1	-9.4	-11.1	0.3	0.1	-9.2	-10.6	-9.4	-12.2
	450	0.0	-8.9	-9.2	-9.2	-10.5	-9.4	-10.4	0.3	-0.1	-9.2	-10.1	-9.3	-11.0
	600	0.0	-8.3	-8.5	-8.6	-9.5	-8.7	-9.4	0.0	-0.5	-8.5	-9.3	-8.7	-9.9
	900	0.0	-7.8	-8.0	-8.0	-8.7	-8.1	-8.6	-0.2	-0.8	-8.0	-8.5	-8.1	-9.0
	1200	0.0	-8.1	-8.3	-8.3	-8.9	-8.5	-8.7	-0.3	-1.0	-8.3	-8.8	-8.4	-9.0
Coverage (Exact)	300	94.0	94.6	93.8	94.5	93.8	94.5	93.1	96.3	95.4	94.5	93.9	94.4	92.2
	450	94.7	94.7	94.1	94.1	93.3	93.3	93.4	96.2	95.6	94.3	93.5	93.4	92.4
	600	94.7	94.4	94.3	94.4	93.8	94.6	93.8	96.2	95.9	94.4	94.0	94.4	92.4
	900	94.4	94.5	93.9	94.6	94.3	94.6	94.9	96.7	96.4	94.6	94.3	94.5	94.1
	1200	95.7	95.7	95.5	95.5	95.4	95.3	95.5	97.0	96.9	95.6	95.5	95.4	94.9
Coverage (HC2)	300	95.4	94.9	94.4	99.1	97.0			94.7	94.1	98.9	95.1		
	450	96.1	94.8	94.3	99.4	96.3			94.4	93.6	98.9	95.1		
	600	96.7	95.2	94.7	98.8	97.0			95.1	94.8	98.5	95.3		
	900	96.6	94.9	94.2	99.3	96.9			94.6	94.2	99.0	96.0		
	1200	97.6	96.3	95.8	99.3	97.6			96.2	95.8	99.0	96.4		

Table 4 shows power and coverage properties of the inference methods from Section 4. We fix the specification in Model 3, varying experiment size n. The asymptotically exact variance estimator $\widehat{V}(\gamma)$ (Theorem 4.2) has large variance (e.g. larger than the inconsistent HC2-style variance estimator), leading to slight undercoverage for small n. In this example, the HC2 variance estimator happens to perform well for Naive and Lin, but significantly overcovers for other estimators.

7 Discussion and Recommendations for Practice

We conclude by giving some recommendations for empirical practice based on the theory and simulations above. As discussed in Sections 5 and 6, estimates of γ^* are typically noisier than estimates of the Lin or Naive adjustment coefficients γ_L or γ_N. If $\gamma^* - \gamma_L$ is small, then adjustment using a precise estimate of γ_L may be better than an imprecise estimate of γ^* for small enough experiment sizes. On the other hand, if marginal covariance between covariates and outcomes is very different from the conditional covariance (Section 3.1) then $\gamma^* - \gamma_L$ may be large, and the Lin estimator can perform much worse than the asymptotically optimal estimators in Section 3.4.

Due to these considerations, we recommend different estimators for "small" and "large" experiments. For stratified designs with $p = 1/2$ and small n (e.g. $n < 1000$), we recommend the non-interacted regression with parametric strata controls, as in Section 3.3 and 3.5. For large n, we recommend the strata fixed effects estimator, which is asymptotically

optimal in this case. For $p \neq 1/2$ and small n, we recommend the Lin estimator with strata controls, and for large n the partialled Lin estimator with strata controls.

Regardless of the adjustment strategy, we recommend using the asymptotically exact confidence intervals developed in Section 4. Our simulations showed close to nominal coverage for these confidence intervals across all considered estimators. By contrast, confidence intervals based on the HC2 robust variance estimator exhibited significant overcoverage, failing to exploit the efficiency gains due to both covariate adjustment and stratified randomization.

References

Abadie, A. and G. W. Imbens (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, 175–187.

Bai, Y., J. P. Romano, and A. M. Shaikh (2021). Inference in experiments with matched pairs. *Journal of the American Statistical Association*.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*.

Cytrynbaum, M. (2021). Designing representative and balanced experiments by local randomization. Working Paper.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, and K. Baicker (2012). The oregon health insurance experiment: Evidence from the first year. *QJE*, 1057–1106.

Fogarty, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika 105*(4).

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*.

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

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics 7*(1), 295–318.

Liu, H. and Y. Yang (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*.

Lu, X. and H. Liu (2022). Tyranny-of-the-minority regression adjustment in randomized experiments.

Ma, W., F. Tu, and H. Liu (2020). Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective.

Negi, A. and J. M. Wooldridge (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*.

Reluga, K., T. Ye, and Q. Zhao (2022). A unified analysis of regression adjustment in randomized experiments.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–129.

Wang, B., R. Susukida, R. Mojtabai, M. Amin-Esmaeili, and M. Rosenblum (2021). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*.

Ye, T., J. Shao, Y. Yi, and Q. Zhao (2022). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association 00*(0).

Zhu, K., H. Liu, and Y. Yang (2022). Design-based theory for lasso adjustment in randomized block experiments with a general blocking scheme.

8 Appendix

8.1 Naive Regression Adjustment

For completeness, before continuing we describe the asymptotic behavior of the commonly used naive regression estimator

$$Y_i = \widehat{c} + \widehat{\theta}_N D_i + \widehat{\gamma}_N' h_i + e_i \tag{8.1}$$

under stratified designs.

Theorem 8.1. *Let 8.2 hold. If* $D_{1:n} \sim \text{Loc}(\psi, p)$ *then* $\sqrt{n}(\widehat{\theta}_N - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

$$V = \text{Var}(c(X)) + E\left[\text{Var}(b - c_p^{-1}\gamma_N' h|\psi)\right] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$$

The coefficient $\widehat{\gamma}_N \xrightarrow{p} \gamma_N$ *with* $c_p^{-1}\gamma_N = \text{argmin}_{\gamma \in \mathbb{R}^{d_h}} \text{Var}(f - \gamma' h)$ *and target function* $f(x) = m_1(x)\sqrt{\frac{p}{1-p}} + m_0(x)\sqrt{\frac{1-p}{p}}$. *The estimator* $\widehat{\theta}_N$ *is fully efficient if* $\psi = 1$ *(complete randomization) and at least one of the following*

(a) $p = 1/2$.

(b) $E[Y_i(1) - Y_i(0)|X_i] = \text{ATE}$.

(c) $\text{Cov}(Y_i(1) - Y_i(0), h(X)) = 0$.

Theorem 8.1 shows that $\widehat{\theta}_N$ is generally inefficient since it uses the wrong objective function. In particular, (1) the target function $f(x) \neq b(x)$ unless $p = 1/2$. Also (2) the limiting coefficient γ_N minimizes marginal instead of conditional variance. When $p \neq 1/2$, condition (c) is both necessary and sufficient for full efficiency. It requires that the covariates $h(X)$ have no (linear) predictive power for treatment effects. Condition (b) is a stronger restriction on treatment effect heterogeneity implying (c).

The appendix shows that this estimator can be written in the standard form studied in Theorem 3.3 up to lower order factors

$$\widehat{\theta}_N = \widehat{\theta} - \widehat{\gamma}_N'(\bar{h}_1 - \bar{h}_0) + O_p(n^{-1})$$

Using this decomposition, the results in Section 4 show how to construct asymptotically exact confidence intervals for the ATE using $\widehat{\theta}_N$.

8.2 Proofs for Section 3.1

Assumption 8.2 (Moment Conditions)**.** *Consider the following assumptions*

(i) $E[Y(d)^4] < \infty$, $E[\sigma_d^2(X)^2] < \infty$ *for* $d \in \{0, 1\}$. $E[h_{it}^4] < \infty$ *for all* $t \in [d_h]$.

(ii) The functions $E[h_i|\psi_i = \psi]$, $E[z_i|\psi_i = \psi]$ *and* $E[Y_i(d)|\psi_i = \psi]$ *for* $d \in \{0, 1\}$ *are Lipschitz continuous and* $|\psi(X)|_2 < K$ *a.s.*

(iii) $\text{Var}(h) \succ 0$.

Definition 8.3 (Conditional Weak Convergence)**.** For random variables $A_n \in \mathbb{R}^d$ and σ-algebras $(\mathcal{F}_n)_n$, define conditional weak convergence

$$A_n|\mathcal{F}_n \Rightarrow A \quad \iff \quad E[e^{it'A_n}|\mathcal{F}_n] = E[e^{it'A}] + o_p(1) \quad \forall t \in \mathbb{R}^d$$

Proof of Theorem 3.3. First, note that we have

$$\begin{aligned}\widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p &= \widehat{\gamma}' E_n\left[\frac{(D_i - p)}{\sqrt{p - p^2}} h_i\right] = (\widehat{\gamma} - \gamma)' E_n\left[\frac{(D_i - p)}{\sqrt{p - p^2}} h_i\right] + \gamma' E_n\left[\frac{(D_i - p)}{\sqrt{p - p^2}} h_i\right] \\ &= \gamma' E_n\left[\frac{(D_i - p)}{\sqrt{p - p^2}} h_i\right] + o_p(n^{-1/2})\end{aligned}$$

Denote $b_i = b(X_i)$. Then, using the fundamental expansion of the difference-of-means estimator from Theorem 3.17 of Cytrynbaum (2021), we have

$$\widehat{\theta}(\widehat{\gamma}) = E_n\left[c(X_i)\right] + E_n\left[\frac{D_i - p}{\sqrt{p - p^2}}(b_i - h_i'\gamma)\right] + E_n\left[\frac{D_i \epsilon_i^1}{p} - \frac{(1 - D_i)\epsilon_i^0}{1 - p}\right] + o_p(n^{-1/2})$$

Consider the middle term and define the residual $v_i = b_i - \gamma' h_i - E[b_i - \gamma' h_i|\psi_i]$. By Lemma 9.4 of Cytrynbaum (2021) $E_n[(D_i - p)(b_i - \gamma' h_i)] = E_n[(D_i - p)v_i] + o_p(n^{-1/2})$. Let $\mathcal{F}_{x,n} = \sigma(X_{1:n}, \pi_n)$. By Theorem 9.5 of the same paper, we have the weak limit

$$\sqrt{n} E_n\left[\frac{D_i - p}{\sqrt{p - p^2}} v_i\right] \Big| \mathcal{F}_{x,n} \Rightarrow \mathcal{N}(0, \mathrm{Var}(v))$$

in the sense of Definition 8.3. Note that $\mathrm{Var}(v) = E[v^2] = E[\mathrm{Var}(b - \gamma' h|\psi)]$. Then we have shown the decomposition

$$\begin{aligned}\sqrt{n}(\widehat{\theta}(\widehat{\gamma}) - \mathrm{ATE}) &= \sqrt{n} E_n\left[c(X_i) - \mathrm{ATE}\right] + \sqrt{n} E_n\left[\frac{D_i - p}{\sqrt{p - p^2}}(b_i - h_i'\gamma)\right] \\ &\quad + \sqrt{n} E_n\left[\frac{D_i \epsilon_i^1}{p} - \frac{(1 - D_i)\epsilon_i^0}{1 - p}\right] + o_p(1)\end{aligned}$$

Then the exact argument of Theorem 3.17 of Cytrynbaum (2021) shows that $\sqrt{n}(\widehat{\theta}(\widehat{\gamma}) - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$

$$V(\gamma) = \mathrm{Var}(c(X)) + E\Big[\mathrm{Var}(b - \gamma' h|\psi)\Big] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1 - p}\right]$$

This finishes the proof. □

Proof of Theorem 3.1. Define $W_i = (1, \tilde{h}_i)$. First consider the regression $Y_i \sim D_i W_i + (1 - D_i)W_i$, with coefficients $(\widehat{\gamma}_1, \widehat{\gamma}_0)$. By Frisch-Waugh and orthogonality of regressors, $\widehat{\gamma}_1$ is numerically equivalent to the regression coefficient $Y_i \sim D_i W_i$ and similarly for $\widehat{\gamma}_0$. Then consider $Y_i = D_i W_i' \widehat{\gamma}_1 + e_i$ with $E_n[e_i(D_i W_i)] = 0$. Then $D_i Y_i = D_i W_i' \widehat{\gamma}_1 + D_i e_i$ and $E_n[D_i e_i (D_i W_i)] = E_n[e_i (D_i W_i)] = 0$. Then $\widehat{\gamma}_1$ can be identified with the regression coefficient of $Y_i \sim W_i$ in the set $\{i : D_i = 1\}$. Let $\widehat{\gamma}_1 = (\widehat{c}_1, \widehat{\alpha}_1)$. By the usual OLS

formula

$$\widehat{c}_1 = E_n[Y_i|D_i=1] - \widehat{\alpha}_1' E_n[\tilde{h}_i|D_i=1] \qquad \widehat{\alpha}_1 = \mathrm{Var}_n(\tilde{h}_i|D_i=1)^{-1}\,\mathrm{Cov}_n(\tilde{h}_i, Y_i|D_i=1)$$

Similar formulas hold for $D_i = 0$ by symmetry. Next, note that for $m = d_h + 1$ the original regressors can be written as a linear transformation

$$\begin{pmatrix} D_i W_i \\ W_i \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ I_m & I_m \end{pmatrix} \begin{pmatrix} D_i W_i \\ (1-D_i) W_i \end{pmatrix}$$

Then the OLS coefficients for the original regression $Y_i \sim D_i W_i + W_i$ are given by the change of variables formula

$$\left(\begin{pmatrix} I_k & 0 \\ I_k & I_k \end{pmatrix}' \right)^{-1} \begin{pmatrix} \widehat{\gamma}_1 \\ \widehat{\gamma}_0 \end{pmatrix} = \begin{pmatrix} I_k & -I_k \\ 0 & I_k \end{pmatrix} \begin{pmatrix} \widehat{\gamma}_1 \\ \widehat{\gamma}_0 \end{pmatrix} = \begin{pmatrix} \widehat{\gamma}_1 - \widehat{\gamma}_0 \\ \widehat{\gamma}_0 \end{pmatrix}$$

In particular, the coefficient on D_i in the original regression is

$$\begin{aligned}
\widehat{\theta}_L = \widehat{c}_1 - \widehat{c}_0 &= E_n[Y_i - \widehat{\alpha}_1' \tilde{h}_i | D_i = 1] - E_n[Y_i - \widehat{\alpha}_0' \tilde{h}_i | D_i = 0] \\
&= \widehat{\theta} - E_n\left[\frac{\widehat{\alpha}_1' \tilde{h}_i D_i}{p}\right] + E_n\left[\frac{\widehat{\alpha}_0' \tilde{h}_i (1-D_i)}{1-p}\right] \\
&= \widehat{\theta} - E_n\left[\frac{\widehat{\alpha}_1' h_i (D_i - p)}{p}\right] - E_n\left[\frac{\widehat{\alpha}_0' h_i (D_i - p)}{1-p}\right] \\
&= \widehat{\theta} - (\widehat{\alpha}_1(1-p) + \widehat{\alpha}_0 p)' E_n\left[\frac{h_i (D_i - p)}{p(1-p)}\right] \\
&= \widehat{\theta} - \left(\widehat{\alpha}_1 \sqrt{\frac{1-p}{p}} + \widehat{\alpha}_0 \sqrt{\frac{p}{1-p}}\right)' (\bar{h}_1 - \bar{h}_0) c_p
\end{aligned}$$

The first equality since $E_n[D_i] = p$ identically. The second equality by expanding $D_i = D_i - p + p$ and using $E_n[\tilde{h}_i] = 0$ and $E_n[(D_i - p)E_n[h_i]] = 0$.

Next, consider the coefficient $\widehat{\alpha}_1 = \mathrm{Var}_n(\tilde{h}_i|D_i=1)^{-1}\,\mathrm{Cov}_n(\tilde{h}_i, Y_i|D_i=1)$. We have $\mathrm{Var}_n(\tilde{h}_i|D_i=1) = p^{-1}E_n[D_i\tilde{h}_i\tilde{h}_i'] - p^{-2}E_n[D_i\tilde{h}_i]E_n[D_i\tilde{h}_i']$. Let $1 \le t, t' \le d_h$. Then we may compute $E_n[D_i\tilde{h}_{it}\tilde{h}_{it'}] = E_n[(D_i - p)\tilde{h}_{it}\tilde{h}_{it'}] + pE_n[\tilde{h}_{it}\tilde{h}_{it'}]$. For the first term, by Lemma 9.20 of C21 Young's inequality, and Jensen

$$\begin{aligned}
&\mathrm{Var}(\sqrt{n}E_n[(D_i - p)\tilde{h}_{it}\tilde{h}_{it'}]|h_{1:n}) \le 2E_n[\tilde{h}_{it}^2\tilde{h}_{it'}^2] \le E_n[\tilde{h}_{it}^4 + \tilde{h}_{it'}^4] \\
&\le 8(E_n[h_{it}^4] + E_n[h_{it'}^4] + E_n[h_{it}]^4 + E_n[h_{it'}]^4) \le 16(E_n[h_{it}^4] + E_n[h_{it'}^4]) = O_p(1)
\end{aligned}$$

The final line is by Markov inequality and since $E[h_{it}^4] < \infty$ by assumption. Then $E_n[D_i\tilde{h}_{it}\tilde{h}_{it'}] = pE_n[\tilde{h}_{it}\tilde{h}_{it'}] + O_p(n^{-1/2})$. We also have $E_n[D_i\tilde{h}_i] = E_n[(D_i - p)\tilde{h}_i] + pE_n[\tilde{h}_i] = E_n[(D_i - p)h_i] = O_p(n^{-1/2})$. Then $\mathrm{Var}_n(\tilde{h}_i|D_i=1)^{-1} = \mathrm{Var}(h)^{-1} + O_p(n^{-1/2})$. Similar reasoning shows that $\mathrm{Cov}_n(\tilde{h}_i, Y_i|D_i=1) = \mathrm{Cov}(h_i, Y_i(1)) + O_p(n^{-1/2})$.

Then we have shown $\widehat{\alpha}_1 = \mathrm{Var}(h)^{-1}\,\mathrm{Cov}(h, Y(1)) + O_p(n^{-1/2}) = \mathrm{Var}(h)^{-1}\,\mathrm{Cov}(h, m_1) + O_p(n^{-1/2})$. By symmetry, we also have $\widehat{\alpha}_0 = \mathrm{Var}(h)^{-1}\,\mathrm{Cov}(h, m_0) + O_p(n^{-1/2})$. Putting

this all together, we have

$$\widehat{\alpha}_1\sqrt{\frac{1-p}{p}}+\widehat{\alpha}_0\sqrt{\frac{p}{1-p}}=\mathrm{Var}(h)^{-1}\,\mathrm{Cov}(h,b)+o_p(1)=\gamma_L+o_p(1)$$

Then by Theorem 3.3, $\sqrt{n}(\widehat{\theta}_L-\mathrm{ATE})\Rightarrow\mathcal{N}(0,V)$ with

$$V=V(\gamma_L)=\mathrm{Var}(c(X))+E\Big[\mathrm{Var}(b-\gamma_L'h|\psi)\Big]+E\left[\frac{\sigma_1^2(X)}{p}+\frac{\sigma_0^2(X)}{1-p}\right]$$

as claimed. The claimed representation follows from the change of variables formula above, since $\widehat{\alpha}_1=\widehat{a}_1+\widehat{a}_0$ and $\widehat{\alpha}_0=\widehat{a}_0$. This finishes the proof. □

Proof of Theorem 8.1. We have $Y_i=\widehat{c}+\widehat{\theta}_N D_i+\widehat{\gamma}_N'h_i+e_i$ with $E_n[e_i(1,D_i,h_i)]=0$. By applying Frisch-Waugh twice, we have $\check{Y}_i=\widehat{\theta}_N(D_i-p)+\widehat{\gamma}_N'\tilde{h}_i+e_i$ and

$$\widehat{\theta}_N=E_n[(\check{D}_i)^2]^{-1}E_n[\check{D}_iY_i]$$

with partialled treatment $\check{D}_i=(D_i-p)-(E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_i(D_i-p)])'\tilde{h}_i$. Then

$$\begin{aligned}(\check{D}_i)^2&=(D_i-p)^2-2(D_i-p)(E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_i(D_i-p)])'\tilde{h}_i\\&\quad+((E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_i(D_i-p)])'\tilde{h}_i)^2\equiv\eta_{i1}+\eta_{i2}+\eta_{i3}\end{aligned}$$

Using $E_n[\tilde{h}_i(D_i-p)]=O_p(n^{-1/2})$, we see that $E_n[\eta_{i2}]=O_p(n^{-1})$ and $E_n[\eta_{i3}]=O_p(n^{-1})$. Then we have $E_n[(\check{D}_i)^2]=E_n[(D_i-p)^2]+O_p(n^{-1})=p-p^2+O_p(n^{-1})$. Then apparently $\widehat{\theta}_N=(p-p^2)^{-1}E_n[\check{D}_iY_i]+O_p(n^{-1})$. Now note that

$$\begin{aligned}E_n[\check{D}_iY_i]&=E_n[(D_i-p)Y_i]-E_n[(E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_i(D_i-p)])'\tilde{h}_iY_i]\\&=E_n[(D_i-p)Y_i]-E_n[(D_i-p)\tilde{h}_i]'(E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_iY_i])\end{aligned}$$

By using Frisch-Waugh to partial out D_i-p from the original regression, we have $\widehat{\gamma}_N=E_n[\bar{h}_i\bar{h}_i']^{-1}E_n[\bar{h}_iY_i]$ with $\bar{h}_i=\tilde{h}_i-(E_n[(D_i-p)^2]^{-1}E_n[\tilde{h}_i(D_i-p)])(D_i-p)$. Then using $E_n[\tilde{h}_i(D_i-p)]=O_p(n^{-1/2})$ again, we have $E_n[\bar{h}_i\bar{h}_i']=E_n[\tilde{h}_i\tilde{h}_i']+O_p(n^{-1})$. Similarly, $E_n[\bar{h}_iY_i]=E_n[\tilde{h}_iY_i]-\widehat{\theta}E_n[\tilde{h}_i(D_i-p)]=E_n[\tilde{h}_iY_i]+O_p(n^{-1/2})$. Then the coefficient $\widehat{\gamma}_N=E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_iY_i]+O_p(n^{-1/2})$. Then we have shown that

$$\begin{aligned}\widehat{\theta}_N&=\widehat{\theta}-E_n\left[\frac{(D_i-p)\tilde{h}_i}{\sqrt{p-p^2}}\right]'(E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_iY_i])(p-p^2)^{-1/2}+O_p(n^{-1})\\&=\widehat{\theta}-E_n\left[\frac{(D_i-p)h_i}{\sqrt{p-p^2}}\right]'\widehat{\gamma}_N(p-p^2)^{-1/2}+O_p(n^{-1})\\&=\widehat{\theta}-(\widehat{\gamma}_N/c_p)'(\bar{h}_1-\bar{h}_0)c_p+O_p(n^{-1})\end{aligned}$$

The second line uses that $E_n[(D_i-p)c]=0$ for any constant. This shows the claimed representation. We have $E_n[\tilde{h}_i\tilde{h}_i']=\mathrm{Var}(h)+o_p(1)$. Note also that $E_n[\tilde{h}_iY_i(1)D_i]=p\,\mathrm{Cov}(h,Y(1))+o_p(1)$ and $E_n[\tilde{h}_iY_i(0)(1-D_i)]=(1-p)\,\mathrm{Cov}(h,Y(0))+o_p(1)$. Putting

this together, we have

$$\begin{aligned}\widehat{\gamma}_N/c_p &= \operatorname{Var}(h)^{-1}\operatorname{Cov}\left(h, m_1\sqrt{\frac{p}{1-p}} + m_0\sqrt{\frac{1-p}{p}}\right) + o_p(1)\\ &= \operatorname*{argmin}_{\gamma}\operatorname{Var}(f-\gamma' h) + o_p(1) = \gamma_N + o_p(1)\end{aligned}$$

Then the first claim follows from Theorem 3.3. For the efficiency claims, (a) if $p = 1/2$ and $\psi = 1$, then $f = b$ and $\gamma_N = \operatorname{argmin}_\gamma \operatorname{Var}(f - \gamma' h) = \operatorname{argmin}_\gamma E[\operatorname{Var}(b - \gamma' h|\psi)]$. For (c), if $\psi = 1$ and $\operatorname{Cov}(h, m_1 - m_0) = 0$, then we have

$$\operatorname{Cov}(h, f) - \operatorname{Cov}(h, b) = \operatorname{Cov}\left(h, (m_1 - m_0)\frac{2p-1}{\sqrt{p(1-p)}}\right) = 0$$

By expanding the variance, we have $\operatorname{argmin}_\gamma \operatorname{Var}(f - \gamma' h) = \operatorname{argmin}_\gamma \operatorname{Var}(b - \gamma' h)$. If (b) holds, then $m_1 - m_0 = 0$ and the same conclusion follows. This finishes the proof. □

Proof of Theorem 3.7. For any $\gamma \in \mathbb{R}^{d_h}$, we have $\operatorname{argmin}_{g\in\mathcal{G}} E[(Y(d) - g(\psi) - \gamma' h)^2] = E[Y(d) - \gamma' h|\psi]$ by standard arguments. Then the coefficients

$$\gamma_d = \operatorname*{argmin}_{\gamma\in\mathbb{R}^{d_h}} E[(Y(d) - \gamma' h - E[Y(d) - \gamma' h|\psi])^2] = \operatorname*{argmin}_{\gamma\in\mathbb{R}^{d_h}} E[\operatorname{Var}(Y(d) - \gamma' h|\psi)]$$

and $g_d(\psi) = E[Y(d) - \gamma_d' h|\psi]$. Define $f_d(x) = g_d(\psi) + \gamma_d' h$. Then the AIPW estimator

$$\begin{aligned}\widehat{\theta}_5 &= E_n[f_1(X_i) - f_0(X_i)] + E_n\left[\frac{D_i(Y_i - f_1(X_i))}{p}\right] - E_n\left[\frac{(1-D_i)(Y_i - f_0(X_i))}{1-p}\right]\\ &= \widehat{\theta} - E_n\left[f_1(X_i)\frac{(D_i - p)}{p}\right] - E_n\left[f_0(X_i)\frac{(D_i - p)}{1-p}\right]\\ &= \widehat{\theta} - E_n\left[(D_i - p)\left(\frac{f_1(X_i)}{p} + \frac{f_0(X_i)}{1-p}\right)\right]\\ &= E_n\left[\frac{D_i - p}{p - p^2}\left(Y_i - (1-p)f_1(X_i) - pf_0(X_i)\right)\right]\end{aligned}$$

Let $F(x) = (1+p)f_1(x) + pf_0(x)$. Then by vanilla CLT we have $\sqrt{n}(\widehat{\theta}_5 - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with $V = \operatorname{Var}\left(\frac{D_i - p}{p - p^2}(Y_i - F(X_i))\right) \equiv \operatorname{Var}(W_i)$. By fundamental expansion of the IPW estimator from Cytrynbaum (2021)

$$\begin{aligned}W_i = \frac{D_i - p}{p - p^2}(Y_i - F(X_i)) - \text{ATE} &= \left[\frac{D_i\epsilon_i^1}{p} - \frac{(1-D_i)\epsilon_i^0}{1-p}\right]\\ + [c(X_i) - \text{ATE}] + \left[\frac{D_i - p}{\sqrt{p - p^2}}\left((m_1 - f_1)\sqrt{\frac{1-p}{p}} + (m_0 - f_0)\sqrt{\frac{p}{1-p}}\right)\right]\end{aligned}$$

By the law of total variance and tower law

$$\begin{aligned}\operatorname{Var}(W) &= \operatorname{Var}(E[W|X]) + E[\operatorname{Var}(W|X)]\\ &= \operatorname{Var}(E[W|X]) + E[\operatorname{Var}(E[W|X, D]|X)] + E[\operatorname{Var}(W|X, D)]\end{aligned}$$

From the expansion above, $\text{Var}(E[W|X]) = \text{Var}(c(X) - \text{ATE}) = \text{Var}(c(X))$. Next

$$E[W|X,D] = [c(X_i) - \text{ATE}] + \left[\frac{D_i - p}{\sqrt{p - p^2}}\left((m_1 - f_1)\sqrt{\frac{1-p}{p}} + (m_0 - f_0)\sqrt{\frac{p}{1-p}}\right)\right]$$

$$E[\text{Var}(E[W|X,D]|X)] = E\left[\left((m_1 - f_1)\sqrt{\frac{1-p}{p}} + (m_0 - f_0)\sqrt{\frac{p}{1-p}}\right)^2\right]$$

$$= E\left[\left((m_1 - \gamma_1'h - E[m_1 - \gamma_1'h|\psi])\sqrt{\frac{1-p}{p}} + (m_0 - \gamma_0'h - E[Y(0) - \gamma_0'h|\psi])\sqrt{\frac{p}{1-p}}\right)^2\right]$$

$$= E\left[\text{Var}\left((m_1 - \gamma_1'h)\sqrt{\frac{1-p}{p}} + (m_0 - \gamma_0'h)\sqrt{\frac{p}{1-p}}\middle|\psi\right)\right]$$

$$= E\left[\text{Var}\left(b - \left(\gamma_1\sqrt{\frac{1-p}{p}} + \gamma_0\sqrt{\frac{p}{1-p}}\right)'h\middle|\psi\right)\right] = \underset{\gamma\in\mathbb{R}^{d_h}}{\text{argmin}}\, E[\text{Var}(b - \gamma'h|\psi)]$$

The final line by characterization of γ_d above and linearity of $Z \to \text{argmin}_\gamma E[\text{Var}(Z - \gamma'h|\psi)]$. Finally note that

$$\text{Var}(W|X,D) = E\left[\left(\frac{D_i\epsilon_i^1}{p} - \frac{(1-D_i)\epsilon_i^0}{1-p}\right)^2\middle|X,D\right] = E\left[\frac{D_i(\epsilon_i^1)^2}{p^2} + \frac{(1-D_i)(\epsilon_i^0)^2}{(1-p)^2}\middle|X_i,D_i\right]$$
$$= \frac{D_i\sigma_1^2(X_i)}{p^2} + \frac{(1-D_i)\sigma_0^2(X_i)}{(1-p)^2}$$

Then $E[\text{Var}(W|X,D)] = E\left[\frac{\sigma_1^2(X_i)}{p} + \frac{\sigma_0^2(X_i)}{1-p}\right]$. Comparing with Equation 3.3 finishes the proof. □

Proof of Theorem 3.7. For any $\gamma \in \mathbb{R}^{d_h}$, we have $\text{argmin}_{g\in\mathcal{G}} E[(Y(d) - g(\psi) - \gamma'h)^2] = E[Y(d) - \gamma'h|\psi]$ by standard arguments. Then the coefficients

$$\gamma_d = \underset{\gamma\in\mathbb{R}^{d_h}}{\text{argmin}}\, E[(Y(d) - \gamma'h - E[Y(d) - \gamma'h|\psi])^2] = \underset{\gamma\in\mathbb{R}^{d_h}}{\text{argmin}}\, E[\text{Var}(Y(d) - \gamma'h|\psi)]$$

and $g_d(\psi) = E[Y(d) - \gamma_d'h|\psi]$. Define $f_d(x) = g_d(\psi) + \gamma_d'h$. Then the AIPW estimator

$$\widehat{\theta}_5 = E_n[f_1(X_i) - f_0(X_i)] + E_n\left[\frac{D_i(Y_i - f_1(X_i))}{p}\right] - E_n\left[\frac{(1-D_i)(Y_i - f_0(X_i))}{1-p}\right]$$
$$= \widehat{\theta} - E_n\left[f_1(X_i)\frac{(D_i - p)}{p}\right] - E_n\left[f_0(X_i)\frac{(D_i - p)}{1-p}\right]$$
$$= \widehat{\theta} - E_n\left[(D_i - p)\left(\frac{f_1(X_i)}{p} + \frac{f_0(X_i)}{1-p}\right)\right]$$
$$= E_n\left[\frac{D_i - p}{p - p^2}\left(Y_i - (1-p)f_1(X_i) - pf_0(X_i)\right)\right]$$

Let $F(x) = (1+p)f_1(x) + pf_0(x)$. Then by vanilla CLT we have $\sqrt{n}(\widehat{\theta}_5 - \text{ATE}) \Rightarrow \mathcal{N}(0,V)$ with $V = \text{Var}\left(\frac{D_i - p}{p - p^2}(Y_i - F(X_i))\right) \equiv \text{Var}(W_i)$. By fundamental expansion of the IPW

estimator from Cytrynbaum (2021)

$$W_i = \frac{D_i - p}{p - p^2}\left(Y_i - F(X_i)\right) - \text{ATE} = \left[\frac{D_i\epsilon_i^1}{p} - \frac{(1-D_i)\epsilon_i^0}{1-p}\right]$$
$$+ \left[c(X_i) - \text{ATE}\right] + \left[\frac{D_i - p}{\sqrt{p-p^2}}\left((m_1 - f_1)\sqrt{\frac{1-p}{p}} + (m_0 - f_0)\sqrt{\frac{p}{1-p}}\right)\right]$$

By the law of total variance and tower law

$$\begin{aligned}\text{Var}(W) &= \text{Var}(E[W|X]) + E[\text{Var}(W|X)] \\ &= \text{Var}(E[W|X]) + E[\text{Var}(E[W|X,D]|X)] + E[\text{Var}(W|X,D)]\end{aligned}$$

From the expansion above, $\text{Var}(E[W|X]) = \text{Var}(c(X) - \text{ATE}) = \text{Var}(c(X))$. Next

$$E[W|X,D] = [c(X_i) - \text{ATE}] + \left[\frac{D_i - p}{\sqrt{p-p^2}}\left((m_1 - f_1)\sqrt{\frac{1-p}{p}} + (m_0 - f_0)\sqrt{\frac{p}{1-p}}\right)\right]$$
$$E[\text{Var}(E[W|X,D]|X)] = E\left[\left((m_1 - f_1)\sqrt{\frac{1-p}{p}} + (m_0 - f_0)\sqrt{\frac{p}{1-p}}\right)^2\right]$$

Using the definition of $f_d(x)$ gives

$$E\left[\left((m_1 - \gamma_1'h - E[m_1 - \gamma_1'h|\psi])\sqrt{\frac{1-p}{p}} + (m_0 - \gamma_0'h - E[Y(0) - \gamma_0'h|\psi])\sqrt{\frac{p}{1-p}}\right)^2\right]$$
$$= E\left[\text{Var}\left((m_1 - \gamma_1'h)\sqrt{\frac{1-p}{p}} + (m_0 - \gamma_0'h)\sqrt{\frac{p}{1-p}}\middle|\psi\right)\right]$$
$$= E\left[\text{Var}\left(b - \left(\gamma_1\sqrt{\frac{1-p}{p}} + \gamma_0\sqrt{\frac{p}{1-p}}\right)'h\middle|\psi\right)\right] = \underset{\gamma\in\mathbb{R}^{d_h}}{\text{argmin}}\, E[\text{Var}(b - \gamma'h|\psi)]$$

The final line by characterization of γ_d above and linearity of $Z \to \text{argmin}_\gamma E[\text{Var}(Z - \gamma'h|\psi)]$. Finally note that

$$\text{Var}(W|X,D) = E\left[\left(\frac{D_i\epsilon_i^1}{p} - \frac{(1-D_i)\epsilon_i^0}{1-p}\right)^2\middle|X,D\right] = E\left[\frac{D_i(\epsilon_i^1)^2}{p^2} + \frac{(1-D_i)(\epsilon_i^0)^2}{(1-p)^2}\middle|X_i,D_i\right]$$
$$= \frac{D_i\sigma_1^2(X_i)}{p^2} + \frac{(1-D_i)\sigma_0^2(X_i)}{(1-p)^2}$$

Then $E[\text{Var}(W|X,D)] = E\left[\frac{\sigma_1^2(X_i)}{p} + \frac{\sigma_0^2(X_i)}{1-p}\right]$. Comparing with Equation 3.3 finishes the proof. □

8.3 Proofs for Section 3.3

Proof of Theorem 3.9. By Theorem 3.1, the middle term of the asymptotic variance is $E[\text{Var}(b - \beta'h|\psi)]$ with $\beta = \text{Var}(h)^{-1}\,\text{Cov}(h,b)$. This is the OLS coefficient from the regression $b = a + \beta'h + e = a + \alpha'z + \gamma'w + e$ with $E[e(1,w,z)] = 0$ and $h = (w,z)$. Denote $\tilde{b} = b - E[b]$ and similarly for $\tilde{w}, \tilde{z}$. By Frisch-Waugh we have $\tilde{b} = \alpha'\tilde{z} + \gamma'\tilde{w} + e$.

Let $\check{w} = \tilde{w} - (E[\tilde{z}\tilde{z}']^{-1}E[\tilde{z}\tilde{w}'])'\tilde{z}$. Then again by Frisch-Waugh the coefficient of interest is $\gamma = E[\check{w}\check{w}']^{-1}E[\check{w}b]$. Next, we characterize this coefficient.

By assumption, $E[w|\psi] = c + \Lambda z$. De-meaning both sides gives $E[\tilde{w}|\psi] = \Lambda\tilde{z}$. Write $\tilde{u} = \tilde{w} - E[\tilde{w}|\psi] = \tilde{w} - \Lambda\tilde{z}$ with $E[\tilde{u}|\psi] = 0$. Then we have

$$E[\tilde{z}\tilde{w}'] = E[\tilde{z}(\tilde{w} - E[\tilde{w}|\psi] + E[\tilde{w}|\psi])'] = E[\tilde{z}\tilde{u}'] + E[\tilde{z}\tilde{z}'\Lambda'] = E[\tilde{z}\tilde{z}']\Lambda'$$

Then $\check{w} = \tilde{w} - (E[\tilde{z}\tilde{z}']^{-1}E[\tilde{z}\tilde{z}']\Lambda')'\tilde{z} = \tilde{w} - \Lambda\tilde{z} = \tilde{u}$. We have now shown that

$$\gamma = E[\tilde{u}\tilde{u}']^{-1}E[\tilde{u}b] = E[\mathrm{Var}(\tilde{w}|\psi)]^{-1}E[\mathrm{Cov}(\tilde{w}, b|\psi)] = E[\mathrm{Var}(w|\psi)]^{-1}E[\mathrm{Cov}(w, b|\psi)]$$

In particular, the coefficient $\beta = (\alpha, \gamma)$ is optimal

$$\begin{aligned} E[\mathrm{Var}(b - \beta'h|\psi)] &= E[\mathrm{Var}(b - \gamma'w|\psi)] = \min_{\tilde{\gamma}} E[\mathrm{Var}(b - \tilde{\gamma}'w|\psi)] \\ &= \min_{\tilde{\alpha},\tilde{\gamma}} E[\mathrm{Var}(b - \tilde{\alpha}'z - \tilde{\gamma}'w|\psi)] = \min_{\beta} E[\mathrm{Var}(b - \beta'h|\psi)] \end{aligned}$$

The second equality since $z = z(\psi)$. This completes the proof. □

8.4 Proofs for Section 3.4

Proof of Theorem 3.16. By Frisch-Waugh $\check{Y}_i = \widehat{\theta}_1\check{D}_i + \widehat{\gamma}_1'\check{h}_i + e_i$ with $\check{D}_i = D_i - k^{-1}\sum_{j\in g(i)} D_j = D_i - p$ and $\check{h}_i = h_i - k^{-1}\sum_{j\in g(i)} h_j$. Applying Frisch-Waugh again, the estimator is $\widehat{\theta}_1 = E_n[(\bar{D}_i)^2]^{-1}E_n[\bar{D}_iY_i]$ with

$$\bar{D}_i = (D_i - p) - (E_n[\check{h}_i\check{h}_i{}']^{-1}E_n[\check{h}_i(D_i - p)])'\check{h}_i$$

By Lemma 8.9 we have $E_n[\check{h}_i\check{h}_i{}'] \xrightarrow{p} \frac{k-1}{k}E[\mathrm{Var}(h|\psi)] \succ 0$, so that $E_n[\check{h}_i\check{h}_i{}']^{-1} = O_p(1)$. By the definition of stratification, $E_n[(D_i - p)\mathbb{1}(g(i) = g)] = 0$ for all g. Then defining $\bar{h}_g \equiv k^{-1}\sum_{j\in g} h_j$ we may write

$$\begin{aligned} E_n[(D_i - p)\check{h}_i] &= E_n\left[(D_i - p)\left(h_i - \sum_g \mathbb{1}(g(i) = g)\bar{h}_g\right)\right] \\ &= E_n[(D_i - p)h_i] = O_p(n^{-1/2}) \end{aligned}$$

The final equality since $E[|h|_2^2] < \infty$ and by Lemma 9.20 of Cytrynbaum (2021). Then apparently $E_n[(\tilde{D}_i)^2] = E_n[(D_i - p)^2] + O_p(n^{-1})$ so that $E_n[(\tilde{D}_i)^2]^{-1} = (p - p^2)^{-1} + O_p(n^{-1})$. Then we have shown that

$$\begin{aligned} \widehat{\theta}_1 &= \frac{E_n[(D_i - p)Y_i]}{p - p^2} - \frac{E_n[\check{h}_i(D_i - p)]'E_n[\check{h}_i\check{h}_i{}']^{-1}E_n[\check{h}_iY_i]}{p - p^2} + O_p(n^{-1}) \\ \widehat{\theta}_1 &= \widehat{\theta} - (\bar{h}_1 - \bar{h}_0)'E_n[\check{h}_i\check{h}_i{}']^{-1}E_n[\check{h}_iY_i] + O_p(n^{-1}) \end{aligned}$$

By Lemma 8.9 we have

$$\begin{aligned} E_n[\check{h}_i Y_i] &= E_n[\check{h}_i D_i Y_i(1)] + E_n[\check{h}_i (1-D_i) Y_i(0)] \\ &= \frac{p(k-1)}{k} E[\mathrm{Cov}(h, Y(1)|\psi)] + \frac{(1-p)(k-1)}{k} E[\mathrm{Cov}(h, Y(0)|\psi)] + o_p(1) \\ &= \frac{(k-1)}{k} E\left[\mathrm{Cov}\left(h, p \cdot m_1(X) + (1-p) \cdot m_0(X)|\psi\right)\right] + o_p(1) \end{aligned}$$

Putting this together, we have

$$c_p^{-1} E_n[\check{h}_i \check{h}_i']^{-1} E_n[\check{h}_i Y_i] \xrightarrow{p} E[\mathrm{Var}(h|\psi)]^{-1} E[\mathrm{Cov}(h, f|\psi)] = \underset{\gamma}{\mathrm{argmin}}\, E[\mathrm{Var}(f - \gamma' h|\psi)]$$

Similar reasoning as above shows that $\widehat{\gamma}_1 = E_n[\check{h}_i \check{h}_i']^{-1} E_n[\check{h}_i Y_i] + O_p(n^{-1/2})$. Then we have representation $\widehat{\theta}_1 = \widehat{\theta} - (c_p^{-1}\widehat{\gamma}_1)'(\bar{h}_1 - \bar{h}_0)c_p + o_p(n^{-1/2})$. The efficiency claims follow identically to the reasoning in Theorem 8.1. This finishes the proof. □

Proof of Theorem 3.22 (Part I). Consider the regression $Y_i \sim D_i(1, \check{h}_i) + (1 - D_i)(1, \check{h}_i)$ with $\check{h}_i = h_i - k^{-1} \sum_{j \in g(i)} h_j$. Denote the OLS coefficients by $(\widehat{c}_1, \widehat{\alpha}_1)$ and $(\widehat{c}_0, \widehat{\alpha}_0)$ respectively. By Frisch-Waugh, the coefficient $(\widehat{c}_1, \widehat{\alpha}_1)$ is given by the equation

$$Y_i = \widehat{c}_1 + \widehat{\alpha}_1' \check{h}_i + e_i \qquad E_n[e_i(1, \check{h}_i)|D_i = 1] = 0$$

By the usual OLS formula $\widehat{\alpha}_1 = \mathrm{Var}_n(\check{h}_i|D_i = 1)^{-1} \mathrm{Cov}_n(\check{h}_i, Y_i|D_i = 1)$. Observe that by definition of stratification

$$P_n(g(i) = g|D_i = 1) = \frac{P_n(D_i = 1|g(i) = g) P_n(g(i) = g)}{P_n(D_i = 1)} = P_n(g(i) = g)$$

This shows that $E_n[E_n[h_i|g(i)]|D_i = 1] = E_n[E_n[h_i|g(i)]] = E_n[h_i]$, so that $E_n[\check{h}_i|D_i = 1] = E_n[h_i|D_i = 1] - E_n[h_i] = E_n[p^{-1}(D_i - p)h_i] = O_p(n^{-1/2})$ as above. Then we have

$$\begin{aligned} \mathrm{Var}_n(\check{h}_i|D_i = 1) &= E_n[\check{h}_i \check{h}_i'|D_i = 1] - E_n[\check{h}_i|D_i = 1] E_n[\check{h}_i|D_i = 1]' \\ &= E_n[\check{h}_i \check{h}_i'|D_i = 1] + O_p(n^{-1}) \end{aligned}$$

Similarly, $\mathrm{Cov}_n(\check{h}_i, Y_i|D_i = 1) = E_n[\check{h}_i Y_i|D_i = 1] + O_p(n^{-1/2})$. Then we have

$$\begin{aligned} \widehat{\alpha}_1 &= E_n[\check{h}_i \check{h}_i'|D_i = 1]^{-1} E_n[\check{h}_i Y_i|D_i = 1] + O_p(n^{-1/2}) \\ &= \frac{k-1}{k} \frac{k}{k-1} E[\mathrm{Var}(h|\psi)]^{-1} E[\mathrm{Cov}(h, Y(1)|\psi)] + o_p(1) \end{aligned}$$

by Lemma 8.9. Similarly, $\widehat{\alpha}_0 = E[\mathrm{Var}(h|\psi)]^{-1} E[\mathrm{Cov}(h, Y(0)|\psi)] + o_p(1)$. By the usual OLS formula, the constant term $\widehat{c}_1$ has form $\widehat{c}_1 = E_n[Y_i|D_i = 1] - \widehat{\alpha}_1' E_n[\check{h}_i|D_i = 1]$ and

similarly for $\widehat{c}_0$. By change of variables used in the proof of Theorem 3.1, our estimator

$$\begin{aligned}\tilde{\theta} = \widehat{c}_1 - \widehat{c}_0 &= E_n[Y_i|D_i=1] - E_n[Y_i|D_i=0] - \left[\widehat{\alpha}_1' E_n[\check{h}_i|D_i=1] - \widehat{\alpha}_0' E_n[\check{h}_i|D_i=0]\right] \\ &= \widehat{\theta} - E_n\left[\frac{\widehat{\alpha}_1' h_i(D_i-p)}{p} + \frac{\widehat{\alpha}_0' h_i(D_i-p)}{1-p}\right] \\ &= \widehat{\theta} - \left[\widehat{\alpha}_1\sqrt{\frac{1-p}{p}} + \widehat{\alpha}_0\sqrt{\frac{p}{1-p}}\right]' E_n\left[\frac{h_i(D_i-p)}{\sqrt{p-p^2}}\right]\end{aligned}$$

Define $\widehat{\gamma} = \widehat{\alpha}_1\sqrt{\frac{1-p}{p}} + \widehat{\alpha}_0\sqrt{\frac{p}{1-p}}$. Then by work above

$$\begin{aligned}\widehat{\gamma} &= E[\mathrm{Var}(h|\psi)]^{-1} E\left[\mathrm{Cov}\left(h, \sqrt{\frac{1-p}{p}}Y(1) + \sqrt{\frac{p}{1-p}}Y(0)|\psi\right)\right] + o_p(1) \\ &= E[\mathrm{Var}(h|\psi)]^{-1} E\left[\mathrm{Cov}\left(h, b|\psi\right)\right] + o_p(1) = \operatorname*{argmin}_{\gamma} E[\mathrm{Var}(b - \gamma' h|\psi)] + o_p(1)\end{aligned}$$

Then applying Theorem 3.3 completes the proof. As before, $\widehat{\alpha}_1 = \widehat{a}_1 + \widehat{a}_0$ and $\widehat{\alpha}_0 = \widehat{a}_0$ by change of variables. □

Proof of Theorem 3.22 (Part II). Next, we analyze the group OLS estimator. By Theorem 3.3, it suffices to show that

$$\widehat{\gamma}_3 = \mathrm{Var}_g(h_g)^{-1}\,\mathrm{Cov}_g(h_g, y_g) = c_p \cdot E[\mathrm{Var}(h|\psi)]^{-1} E[\mathrm{Cov}(h, b|\psi)] + o_p(1)$$

For the first term, note that $E_g[h_g] = O_p(n^{-1/2})$ as above, so that $\mathrm{Var}(h_g) = E_g[h_g h_g'] - E_g[h_g]E_g[h_g]' = E_g[h_g h_g'] + O_p(n^{-1})$. Similarly, $\mathrm{Cov}_g(h_g, y_g) = E_g[h_g y_g] + O_p(n^{-1/2})$. Applying Lemma 8.7 to each component of $h_i h_i'$ shows that

$$E_g[h_g h_g'] = \frac{k}{n}\sum_g \left(k^{-1}\sum_{i\in g}\frac{h_i(D_i-p)}{p-p^2}\right)\left(k^{-1}\sum_{i\in g}\frac{h_i'(D_i-p)}{p-p^2}\right) = \frac{kE[\mathrm{Var}(h|\psi)]}{a(k-a)} + o_p(1)$$

Using the fundamental expansion of the IPW estimator, we have

$$\begin{aligned}E_g[y_g h_g] &= \frac{k}{n}\sum_g \left(k^{-1}\sum_{i\in g}\frac{h_i(D_i-p)}{p-p^2}\right)\left(k^{-1}\sum_{i\in g}\frac{Y_i(D_i-p)}{p-p^2}\right) \\ &= \frac{k}{n}\sum_g \left(k^{-1}\sum_{i\in g}\frac{h_i(D_i-p)}{p-p^2}\right)\left(k^{-1}\sum_{i\in g} c(X_i) + \frac{b_i(D_i-p)}{\sqrt{p-p^2}} + \frac{D_i\epsilon_i^1}{p} - \frac{(1-D_i)\epsilon_i^0}{1-p}\right) \\ &\equiv A_n + B_n + C_n\end{aligned}$$

First, note that $A_n = O_p(n^{-1/2})$ and $C_n = O_p(n^{-1/2})$ by Lemma 8.7. Moreover, we have

$$\begin{aligned}B_n &= \frac{k}{n}\sum_g \left(k^{-1}\sum_{i\in g}\frac{h_i(D_i-p)}{p-p^2}\right)\left(k^{-1}\sum_{i\in g}\frac{b_i(D_i-p)}{\sqrt{p-p^2}}\right) \\ &= \frac{k\sqrt{p-p^2}}{a(k-a)}E[\mathrm{Cov}(h, b|\psi)] + o_p(1) = \frac{E[\mathrm{Cov}(h, b|\psi)]}{\sqrt{a(k-a)}} + o_p(1)\end{aligned}$$

Putting this together, by continuous mapping we have

$$\begin{aligned}\widehat{\gamma}_3 = \mathrm{Var}_g(h_g)^{-1}\,\mathrm{Cov}_g(h_g, y_g) &= \frac{a(k-a)}{k}\frac{1}{\sqrt{a(k-a)}}E[\mathrm{Var}(h|\psi)]^{-1}E[\mathrm{Cov}(h,b|\psi)] + o_p(1) \\ &= \sqrt{p-p^2}E[\mathrm{Var}(h|\psi)]^{-1}E[\mathrm{Cov}(h,b|\psi)] + o_p(1)\end{aligned}$$

Applying Theorem 3.3 completes the proof. □

Proof of Theorem 3.22 (Part III). Finally, we analyze the ToM estimator. Consider the sample moment $E_n[\check{h}_i Y_i^w]$. Applying Lemma 8.9 gives

$$\begin{aligned}E_n[\check{h}_i Y_i^w] &= E_n[\check{h}_i D_i Y_i]\frac{(1-p)^{1/2}}{p^{3/2}} + E_n[\check{h}_i(1-D_i)Y_i]\frac{p^{1/2}}{(1-p)^{3/2}} \\ &= \frac{(1-p)^{1/2}}{p^{3/2}}\frac{p(k-1)}{k}E[\mathrm{Cov}(h,m_1|\psi)] \\ &+ \frac{p^{1/2}}{(1-p)^{3/2}}\frac{(1-p)(k-1)}{k}E[\mathrm{Cov}(h,m_0|\psi)] + o_p(1) \\ &= \frac{k-1}{k}\frac{(1-p)^{1/2}}{p^{1/2}}E[\mathrm{Cov}(h,m_1|\psi)] + \frac{p^{1/2}}{(1-p)^{1/2}}E[\mathrm{Cov}(h,m_0|\psi)] + o_p(1)\end{aligned}$$

The final line is $\frac{k-1}{k}E[\mathrm{Cov}(h,b|\psi)] + o_p(1)$. Then again by Lemma 8.9 and continuous mapping gives

$$\widehat{\gamma}_4 = E_n[\check{h}_i\check{h}_i']^{-1}E_n[\check{h}_i Y_i^w] = \frac{k-1}{k}\frac{k}{k-1}E[\mathrm{Var}(h|\psi)]^{-1}E[\mathrm{Cov}(h,b|\psi)] + o_p(1)$$

Applying Theorem 3.3 completes the proof. □

Proof of Theorem 3.23. First, consider the fixed effects estimator with

$$Y_i = \widehat{c} + \widehat{\tau}_1 D_i + \widehat{\gamma}_1'\check{h}_i + \widehat{\gamma}_z' z_i + e_{i,1}$$

Note that $\tilde{D}_i = D_i - p$ and $\check{h}_i - E_n[\check{h}_i] = \check{h}_i - (E_n[h_i] - E_n[E_n[h_i|g_i = g]) = \check{h}_i$. By Frisch-Waugh, we may instead study $Y_i = \widehat{\tau}_1(D_i - p) + \widehat{\gamma}_1'\check{h}_i + \widehat{\gamma}_z'\tilde{z}_i + e_{i,2}$. Let $\check{w}_i = (\check{h}_i, \tilde{z}_i)$ and $w_i = (h_i, z_i)$. Then by work in Theorem 3.16, $\widehat{\tau}_1 = E_n[(\bar{D}_i)^2]^{-1}E_n[\bar{D}_i Y_i]$ with

$$\bar{D}_i = (D_i - p) - (E_n[\check{w}_i\check{w}_i']^{-1}E_n[\check{w}_i(D_i - p)])'\check{w}_i$$

Previous work suffices to show that $E_n[\check{w}_i(D_i - p)] = O_p(n^{-1/2})$. Then as before, $E_n[(\bar{D}_i)^2]^{-1} = (p-p^2)^{-1} + O_p(n^{-1})$. Then we have

$$\begin{aligned}\widehat{\tau}_1 &= \widehat{\theta} - (p-p^2)^{-1}(E_n[\check{w}_i\check{w}_i']^{-1}E_n[\check{w}_i(D_i-p)])'E_n[\check{w}_i Y_i] \\ &= \widehat{\theta} - (\bar{w}_1 - \bar{w}_0)'E_n[\check{w}_i\check{w}_i']^{-1}E_n[\check{w}_i Y_i]\end{aligned}$$

The second equality uses $E_n[\check{h}_i(D_i - p)] = E_n[h_i(D_i - p)]$ and $E_n[\tilde{z}_i(D_i - p)] = E_n[z_i(D_i - p)]$ as noted before. This shows the claim about estimator representation.

Next, consider the coefficient $\widehat{\gamma}_1$. Define $g_i = (D_i - p, \tilde{z}_i)$. Let $\bar{h}_i = \check{h}_i - (E_n[g_i g_i']^{-1}E_n[g_i\check{h}_i])'g_i$. Then by Frisch-Waugh $\widehat{\gamma}_1 = E_n[\bar{h}_i\bar{h}_i']^{-1}E_n[\bar{h}_i Y_i]$. Consider $E_n[\tilde{z}_i\check{h}_i] = E_n[z_i\check{h}_i]$ since

$E_n[\check{h}_i] = 0$. We have $E_n[z_i\check{h}_i] = o_p(1)$ by Lemma 8.9. Then by previous work $E_n[g_i\check{h}_i] = o_p(1)$. Then $E_n[\bar{h}_i\bar{h}_i{}'] = E_n[\check{h}_i\check{h}_i'] + o_p(1)$. Similarly, $E_n[\bar{h}_iY_i] = E_n[\check{h}_iY_i] + o_p(1)$. Then by continuous mapping $\widehat{\gamma}_1 = E_n[\bar{h}_i\bar{h}_i{}']^{-1}E_n[\bar{h}_iY_i] = E_n[\check{h}_i\check{h}_i']^{-1}E_n[\check{h}_iY_i] + o_p(1)$, the coefficient from the regression without strata variables z_i included shown in Theorem 3.16.

Consider the coefficient $\widehat{\gamma}_z$. Let $q_i = (D_i - p, \check{h}_i)$ and $\bar{z}_i = \tilde{z}_i - (E_n[q_iq_i']^{-1}E_n[q_i\tilde{z}_i])'q_i$. The last paragraph shows that $E_n[q_i\tilde{z}_i] = o_p(1)$. Then by similar reasoning as above and Frisch-Waugh

$$\begin{aligned}\widehat{\gamma}_z &= E_n[\bar{z}_i\bar{z}_i{}']^{-1}E_n[\bar{z}_iY_i] = E_n[\tilde{z}_i\tilde{z}_i']^{-1}E_n[\tilde{z}_iY_i] + o_p(1) \\ &= \operatorname{Var}(z)^{-1}\operatorname{Cov}(z, pm_1 + (1-p)m_0) + o_p(1) = c_p\operatorname{Var}(z)^{-1}\operatorname{Cov}(z, f) + o_p(1)\end{aligned}$$

Our work so far also shows that $E_n[\check{w}_i\check{w}_i'] \xrightarrow{p} \operatorname{Diag}(E_n[\check{h}_i\check{h}_i'], E_n[\tilde{z}_i\tilde{z}_i'])$. Then it's easy to see from our expression for $\widehat{\tau}_1$ that we may identify $\widehat{\gamma}_z = \widehat{\alpha}_1 + o_p(1)$. This finishes the proof for $\widehat{\tau}_1$. The proof for the stratifaction variable adjusted Lin estimator $\widehat{\tau}_2$ is similar and is omitted for brevity.

Next, consider the augmented group OLS estimator $\widehat{\tau}_3$.

$$\begin{aligned}\operatorname{Cov}_n(z_i, \eta_i) &= \operatorname{Cov}\left(z_i, \frac{(1-p)D_iY_i}{p} + \frac{p(1-D_i)Y_i}{1-p}\right) \\ &= \operatorname{Cov}(z_i, (1-p)m_1 + pm_0) + o_p(1)\end{aligned}$$

Then $c_p^{-1}\operatorname{Cov}_n(z_i, \eta_i - \widehat{\gamma}_3'h_i) = \operatorname{Cov}(z, b - c_p^{-1}h'\gamma_3) + o_p(1)$ by continuous mapping, so that

$$\widehat{\alpha}_3 = \operatorname{Var}(z)^{-1}\operatorname{Cov}(z, b - c_p^{-1}h'\gamma_3) + o_p(1) = \operatorname*{argmin}_{\gamma}\operatorname{Var}(b - c_p^{-1}h'\gamma_3 - \gamma'z) + o_p(1)$$

This completes the proof. □

8.5 Proofs for Section 4

Proof of Theorem 4.2. By Theorem 3.3 we have

$$\begin{aligned}V(\gamma) &= \operatorname{Var}(c(X)) + E\left[\operatorname{Var}(b - \gamma'h|\psi)\right] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right] \\ &= \operatorname{Var}(c(X)) + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right] + E[\operatorname{Var}(b|\psi)] \\ &\quad - 2\gamma'E[\operatorname{Cov}(h, b|\psi)] + \gamma'E[\operatorname{Var}(h|\psi)]\gamma\end{aligned}$$

As mentioned in the text, Theorem 5.3 of C21 shows that $\widehat{V} \xrightarrow{p} \operatorname{Var}(c(X)) + E[\operatorname{Var}(b|\psi)] + E\left[\frac{\sigma_1^2(X)}{p} + \frac{\sigma_0^2(X)}{1-p}\right]$. Then it suffices to show that $\widehat{V}_b \xrightarrow{p} E[\operatorname{Cov}(h, b|\psi)]$ and $\widehat{V}_h \xrightarrow{p} E[\operatorname{Var}(h|\psi)]$.

To see this, note that by Lemma 8.9

$$\begin{aligned}
\widehat{V}_b &= \frac{k}{k-1}E_n[\check{h}_i Y_i^w] = \frac{k}{k-1}\left[\frac{(1-p)^{1/2}}{p^{3/2}}E_n[D_i\check{h}_i Y_i] + \frac{p^{1/2}}{(1-p)^{3/2}}E_n[(1-D_i)\check{h}_i Y_i]\right] \\
&= \frac{k}{p(k-1)}\sqrt{\frac{1-p}{p}}E_n[D_i\check{h}_i Y_i] + \frac{k}{(1-p)(k-1)}\sqrt{\frac{p}{1-p}}E_n[(1-D_i)\check{h}_i Y_i] \\
&= \frac{k}{p(k-1)}\sqrt{\frac{1-p}{p}}E_n[D_i\check{h}_i Y_i] + \frac{k}{(1-p)(k-1)}\sqrt{\frac{p}{1-p}}E_n[(1-D_i)\check{h}_i Y_i] \\
&= E\left[\mathrm{Cov}\left(h, \sqrt{\frac{1-p}{p}}m_1 + \sqrt{\frac{p}{1-p}}m_0|\psi\right)\right] + o_p(1) = E[\mathrm{Cov}(h,b|\psi)] + o_p(1)
\end{aligned}$$

By the same lemma, we have $\frac{k}{k-1}E_n[\check{h}_i\check{h}_i'] = E[\mathrm{Var}(h|\psi)] + o_p(1)$. The conclusion then follows by continuous mapping, using $\widehat{\gamma} \xrightarrow{p} \gamma$.

Alternatively, consider defining $\widehat{V}_b = kc_p \cdot E_g[y_g h_g]$ and $\widehat{V}_h = kc_p^2 \cdot E_g[h_g h_g']$. In this case, by Lemma 8.7 and the proof of Theorem 3.22 we have

$$\sqrt{a(k-a)} \cdot E_g[y_g h_g] \xrightarrow{p} E[\mathrm{Cov}(h,b|\psi)] \qquad \frac{a(k-a)}{k} \cdot E_g[h_g h_g'] \xrightarrow{p} E[\mathrm{Var}(h|\psi)]$$

Since $\widehat{\gamma} \xrightarrow{p} \gamma$, the conclusion follows by continuous mapping. □

8.6 Lemmas

Lemma 8.4 (Conditional Convergence)**.** *Let $(\mathcal{G}_n)_{n\geq 1}$ and $(A_n)_{n\geq 1}$ a sequence of σ-algebras and RV's. Define conditional convergence*

$$\begin{aligned}
A_n = o_{p,\mathcal{G}_n}(1) &\iff P(|A_n| > \epsilon|\mathcal{G}_n) = o_p(1) \quad \forall \epsilon > 0 \\
A_n = O_{p,\mathcal{G}_n}(1) &\iff P(|A_n| > s_n|\mathcal{G}_n) = o_p(1) \quad \forall s_n \to \infty
\end{aligned}$$

Then the following results hold

(i) $A_n = o_p(1) \iff A_n = o_{p,\mathcal{G}_n}(1)$ *and* $A_n = O_p(1) \iff A_n = O_{p,\mathcal{G}_n}(1)$

(ii) $E[|A_n||\mathcal{G}_n] = o_p(1)/O_p(1) \implies A_n = o_p(1)/O_p(1)$

(iii) $\mathrm{Var}(A_n|\mathcal{G}_n) = o_p(c_n^2)/O_p(c_n^2) \implies A_n - E[A_n|\mathcal{G}_n] = o_p(c_n)/O_p(c_n)$ *for all positive* $(c_n)_n$

(iv) If $(A_n)_{n\geq 1}$ *has* $A_n \leq \bar{A} < \infty$ $\mathcal{G}_n$*-a.s.* $\forall n$ *and* $A_n = o_p(1) \implies E[|A_n||\mathcal{G}_n] = o_p(1)$

Proof. (i) Consider that for any $\epsilon > 0$

$$P(|A_n| > \epsilon) = E[\mathbb{1}(|A_n| > \epsilon)] = E[E[\mathbb{1}(|A_n| > \epsilon)|\mathcal{G}_n]] = E[P(|A_n| > \epsilon|\mathcal{G}_n)]$$

If $A_n = o_p(1)$, then $E[P(|A_n| > \epsilon|\mathcal{G}_n)] = o(1)$, so $P(|A_n| > \epsilon|\mathcal{G}_n) = o_p(1)$ by Markov inequality. Conversely, if $P(|A_n| > \epsilon|\mathcal{G}_n) = o_p(1)$, then $E[P(|A_n| > \epsilon|\mathcal{G}_n)] = o(1)$ since $(P(|A_n| > \epsilon|\mathcal{G}_n))_{n\geq 1}$ is uniformly bounded, hence UI. Then $P(|A_n| > \epsilon) = o(1)$. The second equivalence follows directly from the first. (ii) follows from (i) and conditional

Markov inequality. (iii) is an application of (ii). For (iv), note that for any $\epsilon > 0$

$$E[|A_n||\mathcal{G}_n] \leq \epsilon + E[|A_n|\mathbb{1}(|A_n| > \epsilon)|\mathcal{G}_n] \leq \epsilon + \bar{A}P(|A_n| > \epsilon|\mathcal{G}_n) = \epsilon + o_p(1)$$

The equality is by (i) and our assumption. Since $\epsilon > 0$ was arbitrary $E[|A_n||\mathcal{G}_n] = o_p(1)$. □

Lemma 8.5. *Let $(a_i), (b_i), (c_i)$ be positive scalar arrays for $i \in [n]$. Then the following*

$$\sum_{\substack{i,j,s \in g \\ i \neq j, j \neq s}} a_i b_j c_s \leq 3 \sum_i (a_i^3 + b_i^3 + c_i^3)$$

Proof. Note that by AM-GM inequality and Jensen, for non-negative x, y, z we have $xyz \leq ((1/3)(x+y+z))^3 \leq (1/3)(x^3+y^3+z^3)$. Applying this gives

$$\begin{aligned}\sum_{\substack{i,j,s \in g \\ i \neq j, j \neq s}} a_i b_j c_s &\leq \left(\sum_i a_i\right)\left(\sum_j b_j\right)\left(\sum_s c_s\right) \\ &\leq (1/3)\left[\left(\sum_i a_i\right)^3 + \left(\sum_j b_j\right)^3 + \left(\sum_s c_s\right)^3\right] \leq 3\sum_i (a_i^3 + b_i^3 + c_i^3)\end{aligned}$$

□

Assumption 8.6 (Group OLS)**.** *Consider the following assumptions*

(i) $E[Y(d)^4] < \infty$ *and* $E[\sigma_d^2(X)^2] < \infty$ *for* $d \in \{0,1\}$. $E[h_{it}^4] < \infty$ *for all* $t \in [d_h]$.

(ii) The functions $E[h_i|\psi_i = \psi]$, $E[z_i|\psi_i = \psi]$ *and* $E[Y_i(d)|\psi_i = \psi]$ *for* $d \in \{0,1\}$ *are Lipschitz continuous.*

(iii) $\mathrm{Var}(h) \succ 0$.

(iv) $|\psi(X)|_2 < K$ *a.s.*

Lemma 8.7 (Group OLS)**.** *Let $h, w : \mathcal{X} \to \mathbb{R}$. Denote $h_i = h(X_i)$ and $w_i = w(X_i)$ and suppose $E[h_i|\psi_i = \psi]$ and $E[w_i|\psi_i = \psi]$ are Lipschitz continuous. Suppose $E[h_i^4] < \infty$ and $E[w_i^4] < \infty$. Let $\epsilon_i^d = Y_i(d) - m_d(X_i)$ for $d \in \{0,1\}$. Then we have*

$$A_n = n^{-1}\sum_g \left(k^{-1}\sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2}\right)\left(k^{-1}\sum_{i \in g} \frac{w_i(D_i - p)}{p - p^2}\right) = \frac{E[\mathrm{Cov}(h, w|\psi)]}{a(k-a)} + o_p(1)$$

$$B_n = n^{-1}\sum_g \left(k^{-1}\sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2}\right)\left(k^{-1}\sum_{i \in g} w_i\right) = O_p(n^{-1/2})$$

$$C_n = \sum_g \left(k^{-1}\sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2}\right)\left(k^{-1}\sum_{i \in g} \frac{D_i \epsilon_i^1}{p} - \frac{(1 - D_i)\epsilon_i^0}{1-p}\right) = O_p(n^{-1/2})$$

Proof. Define the following

$$\bar{h}_{g1} = a^{-1}\sum_{i \in g} h_i \mathbb{1}(D_i = 1) \quad \bar{h}_{g0} = (k-a)^{-1}\sum_{i \in g} h_i \mathbb{1}(D_i = 0) \quad \bar{w}_g = k^{-1}\sum_{i \in g} w_i$$

Recall that $g \in \sigma(\psi_{1:n}, \pi_n)$ for each g and $D_{1:n} \in \sigma(\psi_{1:n}, \pi_n, \tau)$ with exogenous randomization variable $\tau \perp\!\!\!\perp (X_{1:n}, Y(d)_{1:n} : d = 0, 1)$. Notice that $k^{-1} \sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2} = \bar{h}_{g1} - \bar{h}_{g0}$. First consider B_n. By Lemma 9.19 of Cytrynbaum (2021), we have $E[B_n | X_{1:n}, \pi_n] = 0$. Next, we have

$$E[B_n^2 | X_{1:n}, \pi_n] = E\left[n^{-2} \sum_{g,g'} \left(k^{-1} \sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2} \right) \left(k^{-1} \sum_{i \in g'} \frac{h_i(D_i - p)}{p - p^2} \right) \bar{w}_g \bar{w}_{g'} \middle| X_{1:n}, \pi_n \right]$$
$$= E\left[n^{-2} \sum_{g} \left(k^{-1} \sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2} \right)^2 \bar{w}_g^2 \middle| X_{1:n}, \pi_n \right]$$

The second equality follows by Lemma 9.19 of Cytrynbaum (2021), since $\mathrm{Cov}(D_i, D_j | X_{1:n}, \pi_n) = 0$ if i, j are in different groups. We may calculate

$$E\left[\left(k^{-1} \sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2} \right)^2 \middle| X_{1:n}, \pi_n \right] = \frac{1}{k^2 (p - p^2)^2} \sum_{i \in g} h_i^2 \, \mathrm{Var}(D_i | X_{1:n}, \pi_n)$$
$$+ \frac{1}{k^2 (p - p^2)^2} \sum_{i \neq j \in g} h_i h_j \, \mathrm{Cov}(D_i, D_j | X_{1:n}, \pi_n) = \frac{1}{k^2 (p - p^2)} \left[\sum_{i \in g} h_i^2 - (k-1)^{-1} \sum_{i \neq j \in g} h_i h_j \right]$$

Note that $\sum_{i \neq j \in g} |h_i h_j| \leq \left(\sum_{i \in g} |h_i| \right)^2 = k^2 \left(k^{-1} \sum_{i \in g} |h_i| \right)^2 \leq k \sum_{i \in g} |h_i|^2$. The final inequality by Jensen. Then by triangle inequality, a simple calculation gives

$$\frac{1}{k^2} \left| \sum_{i \in g} h_i^2 - (k-1)^{-1} \sum_{i \neq j \in g} h_i h_j \right| \leq \frac{1}{k^2} \frac{2k - 1}{k - 1} \sum_{i \in g} h_i^2 \leq 3k^{-2} \sum_{i \in g} h_i^2$$

Then continuing from above

$$E[B_n^2 | X_{1:n}, \pi_n] \lesssim k^{-2} n^{-2} \sum_g \left(\sum_{i \in g} h_i^2 \right) \left(\sum_{i \in g} w_i \right)^2 \leq \frac{1}{kn^2} \sum_g \left(\sum_{i \in g} h_i^2 \right) \left(\sum_{i \in g} w_i^2 \right)$$
$$\leq \frac{1}{2kn^2} \sum_g \left[\left(\sum_{i \in g} h_i^2 \right)^2 + \left(\sum_{i \in g} w_i^2 \right)^2 \right] = (2n)^{-1} E_n[h_i^4 + w_i^4] = O_p(n^{-1})$$

The second inequality follows from Jensen, and the third by Young's inequality. The first equality by Jensen and final equality by our moment assumption. So by Lemma 9.16 in Cytrynbaum (2021), $B_n = O_p(n^{-1/2})$.

Next, consider A_n. Using the within-group covariances above, we compute

$$\begin{aligned}
E[A_n|X_{1:n},\pi_n] &= \frac{1}{nk^2(p-p^2)^2}\sum_g\sum_{i,j\in g}\mathrm{Cov}(D_i,D_j|X_{1:n},\pi_n)h_iw_j \\
&= \frac{1}{nk^2(p-p^2)^2}\sum_g\left(\sum_{i\in g}(p-p^2)h_iw_i - \sum_{i\neq j\in g}\frac{a(k-a)}{k^2(k-1)}h_iw_j\right) \\
&= \frac{1}{k^2(p-p^2)}\left(E_n[h_iw_i] - \frac{1}{n(k-1)}\sum_g\sum_{i\neq j\in g}h_iw_j\right)
\end{aligned}$$

Define $u_i = w_i - E[w_i|\psi_i]$ and $v_i = h_i - E[h_i|\psi_i]$. Consider the second term. We have

$$n^{-1}\sum_g\sum_{i\neq j\in g}h_iw_j = n^{-1}\sum_g\sum_{i\neq j\in g}(E[h_i|\psi_i]+v_i)(E[w_j|\psi_j]+u_j) \equiv \sum_{l=1}^4 A_{n,l}$$

First, note that for any scalars $a_ib_j + a_jb_i = a_ib_i + a_jb_j + (a_i-a_j)(b_j-b_i)$. Then we have

$$\begin{aligned}
A_{n,1} &\equiv n^{-1}\sum_g\sum_{i\neq j\in g}E[h_i|\psi_i]E[w_j|\psi_j] = n^{-1}\sum_g\sum_{i<j\in g}E[h_i|\psi_i]E[w_j|\psi_j]+E[h_j|\psi_j]E[w_i|\psi_i] \\
&= n^{-1}\sum_g\sum_{i<j\in g}E[h_i|\psi_i]E[w_i|\psi_i]+E[h_j|\psi_j]E[w_j|\psi_j] \\
&+ n^{-1}\sum_g\sum_{i<j\in g}(E[h_i|\psi_i]-E[h_j|\psi_j])(E[w_j|\psi_j]-E[w_i|\psi_i]) \equiv B_{n,1}+C_{n,1}
\end{aligned}$$

By counting ordered tuples (i,j), it's easy to see that

$$\begin{aligned}
B_{n,1} &= n^{-1}\sum_g\sum_{i\in g}(k-1)E[h_i|\psi_i]E[w_i|\psi_i] = (k-1)E_n[E[h_i|\psi_i]E[w_i|\psi_i]] \\
&= (k-1)E[E[h_i|\psi_i]E[w_i|\psi_i]] + o_p(1) = (k-1)(E[h_iw_i]-E[v_iu_i]) + o_p(1)
\end{aligned}$$

For the second term, by our Lipschitz assumptions we have

$$|C_{n,1}| \lesssim n^{-1}\sum_g\sum_{i<j\in g}|\psi_i-\psi_j|_2^2 = o_p(1)$$

Next, claim that $A_{n,l} = o_p(1)$ for $l = 2,3,4$. For instance, we have

$$E[A_{n,2}|\psi_{1:n},\pi_n] = n^{-1}\sum_g\sum_{i\neq j\in g}E\left[E[h_i|\psi_i]u_j|\psi_{1:n},\pi_n\right] = 0$$

Since $E[u_j|\psi_{1:n},\pi_n] = E[u_j|\psi_j] = 0$ by Lemma 9.21 of Cytrynbaum (2021). Moreover, we have

$$E[A_{n,2}^2|\psi_{1:n},\pi_n] = n^{-2}\sum_{g,g'}\sum_{i\neq j\in g}\sum_{s\neq t\in g'}E[h_i|\psi_i]E[h_s|\psi_s]E[u_ju_t|\psi_{1:n},\pi_n]$$

For $j\neq t$, we have $E[u_ju_t|\psi_{1:n},\pi_n] = E[u_j|\psi_j]E[u_t|\psi_t] = 0$ by Lemma 9.21 of the paper

above. Since the groups g are disjoint, and using $E[u_j^2|\psi_{1:n}, \pi_n] = E[u_j^2|\psi_j]$

$$\begin{aligned} E[A_{n,2}^2|\psi_{1:n}, \pi_n] &= n^{-2} \sum_g \sum_{\substack{i,j,s \in g \\ i \neq j, j \neq s}} E[h_i|\psi_i] E[h_s|\psi_s] E[u_j^2|\psi_j] \\ &\leq 3n^{-2} \sum_g \sum_{i \in g} 2E[h_i|\psi_i]^3 + E[u_i^2|\psi_i]^3 \\ &= 3n^{-1} E_n[2E[h_i|\psi_i]^3 + E[u_i^2|\psi_i]^3] = O_p(n^{-1}) \end{aligned}$$

Then we have shown $A_{n,2} = O_p(n^{-1/2})$ by Lemma 9.16 of Cytrynbaum (2021). The proof for $l = 3, 4$ is almost identical. Summarizing, the work above has shown that

$$\begin{aligned} E[A_n|X_{1:n}, \pi_n] &= \frac{1}{k^2(p-p^2)} \left(E_n[h_i w_i] - \frac{1}{k-1}(k-1)(E[h_i w_i] - E[v_i u_i]) \right) + o_p(1) \\ &= \frac{1}{k^2(p-p^2)} E[v_i u_i] + o_p(1) = \frac{E[\mathrm{Cov}(h, w|\psi)]}{a(k-a)} + o_p(1) \end{aligned}$$

Next, we claim that $\mathrm{Var}(A_n|X_{1:n}, \pi_n) = o_p(1)$. Define $\Delta_{h,g} = k^{-1} \sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2}$, then

$$\mathrm{Var}(A_n|X_{1:n}, \pi_n) = n^{-2} \sum_{g,g'} \mathrm{Cov}\left(\Delta_{h,g}\Delta_{w,g}, \Delta_{h,g'}\Delta_{w,g'}|X_{1:n}, \pi_n\right)$$

Note that $\Delta_{h,g}\Delta_{w,g} \perp\!\!\!\perp \Delta_{h,g'}\Delta_{w,g'}|X_{1:n}, \pi_n$ for $g \neq g'$, since treatment assignments are (conditionally) independent between groups. Then the on-diagonal terms are

$$\begin{aligned} \mathrm{Var}(A_n|X_{1:n}, \pi_n) &= n^{-2} \sum_g \mathrm{Var}\left(\left(k^{-1} \sum_{i \in g} \frac{h_i(D_i - p)}{p - p^2} \right) \left(k^{-1} \sum_{i \in g} \frac{w_i(D_i - p)}{p - p^2} \right) \middle| X_{1:n}, \pi_n \right) \\ &= n^{-2} k^{-4} (p-p)^{-4} \sum_g \mathrm{Var}\left(\sum_{i,j \in g} h_i w_j (D_i - p)(D_j - p) \middle| X_{1:n}, \pi_n \right) \end{aligned}$$

The inner variance term can be expanded as

$$\sum_{i,j \in g} \sum_{s,t \in g} h_i w_j h_s w_t \,\mathrm{Cov}\left((D_i - p)(D_j - p), (D_s - p)(D_t - p) \middle| X_{1:n}, \pi_n \right)$$

We have $|\,\mathrm{Cov}((D_i - p)(D_j - p), (D_s - p)(D_t - p)|X_{1:n}, \pi_n)| \leq 2$ since $|(D_i - p)| \leq 1$ for all $i \in [n]$. Using Lemma 9.17 in Cytrynbaum (2021), the previous display is bounded above by

$$\sum_{i,j \in g} \sum_{s,t \in g} |h_i w_j h_s w_t| \cdot 2 \leq 2k^3 \sum_{i \in g} (h_i^4 + w_i^4)$$

Putting this all together, we have

$$\begin{aligned} \mathrm{Var}(A_n|X_{1:n}, \pi_n) &\leq 2n^{-2} k^{-4} (p-p)^{-4} k^3 \sum_g \sum_{i \in g} (h_i^4 + w_i^4) \\ &= 2n^{-1} k^{-1} (p-p)^{-4} E_n[h_i^4 + w_i^4] = O_p(n^{-1}) \end{aligned}$$

By conditional Markov, this shows that $A_n - E[A_n|X_{1:n}, \pi_n] = O_p(n^{-1/2})$. Then we have shown that $A_n = \frac{E[\text{Cov}(h,w|\psi)]}{a(k-a)} + o_p(1)$.

Finally, we consider C_n. Note that $g, D_{1:n} \in \sigma(X_{1:n}, \pi_n, \tau)$ and $E[\epsilon_i^d|X_{1:n}, \pi_n, \tau] = E[\epsilon_i^d|X_i] = 0$ for $d = 0, 1$ by Lemma 9.21 of Cytrynbaum (2021), so we have $E[C_n|X_{1:n}, \pi_n, \tau] = 0$. Next, we claim that $E[C_n^2|X_{1:n}, \pi_n, \tau] = O_p(n^{-1})$. Note that C_n^2 can be written

$$\frac{1}{n^2k^4}\sum_{g,g'}\left(\sum_{i,j\in g}\sum_{s,t\in g'}\frac{h_i(D_i-p)}{p-p^2}\left(\frac{D_j\epsilon_j^1}{p}-\frac{(1-D_j)\epsilon_j^0}{1-p}\right)\frac{h_s(D_s-p)}{p-p^2}\left(\frac{D_t\epsilon_t^1}{p}-\frac{(1-D_t)\epsilon_t^0}{1-p}\right)\right)$$

We have $E[\epsilon_j^d\epsilon_t^{d'}|X_{1:n}, \pi_n, \tau] = E[\epsilon_j^d|X_j]E[\epsilon_t^{d'}|X_t] = 0$ for any $j \neq t$ by Lemma 9.21 of Cytrynbaum (2021). By group disjointness, the term $E[C_n^2|X_{1:n}, \pi_n, \tau]$ simplifies to

$$\frac{1}{n^2k^4}\sum_{g}\left(\sum_{i,j,s\in g}\frac{h_i(D_i-p)}{p-p^2}\frac{h_s(D_s-p)}{p-p^2}E\left[\left(\frac{D_j\epsilon_j^1}{p}-\frac{(1-D_j)\epsilon_j^0}{1-p}\right)^2\middle|X_{1:n}, \pi_n, \tau\right]\right)$$

We have $E[(\epsilon_i^d)^2|X_{1:n}, \pi_n, \tau] = E[(\epsilon_i^d)^2|X_i] = \sigma_d^2(X_i)$. Then by Young's inequality and Lemma 9.21 of the paper above

$$E\left[\left(\frac{D_j\epsilon_j^1}{p}-\frac{(1-D_j)\epsilon_j^0}{1-p}\right)^2\middle|X_{1:n}, \pi_n, \tau\right] \leq 2(p\wedge(1-p))^{-1}(\sigma_1^2(X_j)+\sigma_0^2(X_j))$$

Taking the absolute value of the second to last display and using triangle inequality gives the upper bound

$$\begin{aligned}
&2[n^2k^4(p-p^2)^2(p\wedge(1-p))]^{-1}\sum_g\left(\sum_{i,j,s\in g}|h_ih_s|(\sigma_1^2(X_j)+\sigma_0^2(X_j))\right)\\
\lesssim& n^{-2}\sum_g\left(\sum_{i,j,s\in g}|h_ih_s|^2+(\sigma_1^2(X_j)+\sigma_0^2(X_j))^2\right)\\
\leq& n^{-1}k^2E_n[(\sigma_1^2(X_i)+\sigma_0^2(X_i))^2]+n^{-2}k\sum_g\sum_{i,s\in g}|h_ih_s|^2
\end{aligned}$$

By Young's inequality and assumption $E[E_n[(\sigma_1^2(X_i)+\sigma_0^2(X_i))^2]] \leq 2E[\sigma_1^2(X_i)^2+\sigma_0^2(X_i)^2] < \infty$. For the second term, using Jensen we have

$$n^{-1}\sum_g\sum_{i,s\in g}|h_ih_s|^2 = n^{-1}\sum_g\left(\sum_{i\in g}|h_i|^2\right)^2 \leq kn^{-1}E_n[h_i^4] = O_p(1)$$

Then we have shown that $E[C_n^2|X_{1:n}, \pi_n, \tau] = O_p(n^{-1})$, so by conditional Markov inequality in Lemma 8.4, $C_n = O_p(n^{-1/2})$. This finishes the proof. □

Assumption 8.8 (Partialled Lin)**.** *Consider the following assumptions*

(i) $E[Y(d)^4] < \infty$ *for* $d \in \{0,1\}$ *and* $E[h_{it}^4] < \infty$ *for all* $t \in [d_h]$.

(ii) The functions $E[h_i|\psi_i = \psi]$, $E[z_i|\psi_i = \psi]$ *and* $E[Y_i(d)|\psi_i = \psi]$ *for* $d \in \{0,1\}$ *are Lipschitz continuous.*

(iii) $|\psi(X)|_2 < K$ *a.s.*

Lemma 8.9 (Partialled Lin)**.** *Under assumptions,* $E_n[\check{h}_i z_i] = o_p(1)$*. Also, we have*

$$E_n[D_i\check{h}_i\check{h}_i'] = \frac{p(k-1)}{k}E[\mathrm{Var}(h|\psi)] + o_p(1) \quad E_n[\check{h}_i\check{h}_i'] = \frac{k-1}{k}E[\mathrm{Var}(h|\psi)] + o_p(1)$$
$$E_n[D_i\check{h}_iY_i] = \frac{p(k-1)}{k}E[\mathrm{Cov}(h, m_1|\psi)] + o_p(1)$$
$$E_n[(1-D_i)\check{h}_iY_i] = \frac{(1-p)(k-1)}{k}E[\mathrm{Cov}(h, m_0|\psi)] + o_p(1)$$

Proof. First, observe that

$$\check{h}_i = h_i - k^{-1}\sum_{j\in g(i)} h_j = \frac{k-1}{k}\cdot h_i - k^{-1}\sum_{j\in g(i)\setminus\{i\}} h_j = k^{-1}\sum_{j\in g(i)\setminus\{i\}}(h_i - h_j)$$

Note that $E_n[D_i\check{h}_i\check{h}_i] = E_n[(D_i-p)\check{h}_i\check{h}_i] + pE_n[\check{h}_i\check{h}_i]$. We claim that $E_n[(D_i-p)\check{h}_i\check{h}_i] = O_p(n^{-1/2})$. For $1 \le t, t' \le d_h$, by Lemma 9.20 of Cytrynbaum (2021) and Cauchy-Schwarz we have

$$\mathrm{Var}(\sqrt{n}E_n[(D_i-p)\check{h}_{it}\check{h}_{it'}]|X_{1:n}, \pi_n) \le 2E_n[\check{h}_{it}^2\check{h}_{it'}^2] \le 2E_n[\check{h}_{it}^4]^{1/2}E_n[\check{h}_{it'}^4]^{1/2}$$

Next, note that by Jensen's followed by Young's inequality

$$\check{h}_{it}^4 = \frac{(k-1)^4}{k^4}\left(\frac{1}{k-1}\sum_{j\in g(i)\setminus\{i\}}(h_{it}-h_{jt})\right)^4 \le \frac{(k-1)^3}{k^4}\sum_{j\in g(i)\setminus\{i\}}(h_{it}-h_{jt})^4$$
$$\le 8\frac{(k-1)^3}{k^4}\sum_{j\in g(i)\setminus\{i\}}(h_{it}^4 + h_{jt}^4) \le 8\frac{(k-1)^3}{k^4}\left((k-1)h_{it}^4 + \sum_{j\in g(i)\setminus\{i\}} h_{jt}^4\right)$$

By counting, we have $E_n\left[\sum_{j\in g(i)\setminus\{i\}} h_{jt}^4\right] = (k-1)E_n[h_{it}^4]$. Putting this all together, $E_n[\check{h}_{it}^4] \lesssim E_n[h_{it}^4] = O_p(1)$. Then $\mathrm{Var}(\sqrt{n}E_n[(D_i-p)\check{h}_{it}\check{h}_{it'}]|X_{1:n}, \pi_n) = O_p(1)$ so that $E_n[(D_i-p)\check{h}_{it}\check{h}_{it'}] = O_p(n^{-1/2})$ by Lemma 8.4. Then it suffices to show the claim for $E_n[\check{h}_i\check{h}_i]$. Let $f_{it} = E[h_t(X_i)|\psi_i]$ and write $h_{it} = f_{it} + u_{it}$. Then we have

$$E_n[\check{h}_{it}\check{h}_{it'}] = \frac{1}{nk^2}\sum_i\left(\sum_{j\in g(i)\setminus\{i\}} h_{it} - h_{jt}\right)\left(\sum_{l\in g(i)\setminus\{i\}} h_{it'} - h_{lt'}\right)$$
$$= \frac{1}{nk^2}\sum_i D_i \sum_{j,l\in g(i)\setminus\{i\}}(h_{it}-h_{jt})(h_{it'}-h_{lt'})$$

We can expand the expression above as

$$\frac{1}{nk^2}\sum_i\sum_{j,l\in g(i)\setminus\{i\}}\Big[(f_{it}-f_{jt})(f_{it'}-f_{lt'}) + (f_{it}-f_{jt})(u_{it'}-u_{lt'})$$
$$+ (u_{it}-u_{jt})(f_{it'}-f_{lt'}) + (u_{it}-u_{jt})(u_{it'}-u_{lt'})\Big] \equiv A_n + B_n + C_n + D_n$$

First consider A_n. By the Lipschitz assumption in 8.8 and Young's inequality

$$|A_n| \leq \frac{1}{nk^2} \sum_i \sum_{j,l \in g \setminus \{i\}} |f_{it} - f_{jt}||f_{it'} - f_{lt'}| \lesssim \frac{1}{nk^2} \sum_i \sum_{j,l \in g \setminus \{i\}} |\psi_i - \psi_j|_2 |\psi_i - \psi_l|_2$$
$$\leq \frac{2}{nk^2} \sum_i \sum_{j,l \in g \setminus \{i\}} (|\psi_i - \psi_j|_2^2 + |\psi_i - \psi_l|_2^2) = \frac{4(k-1)}{nk^2} \sum_g \sum_{i,j \in g} |\psi_i - \psi_j|_2^2 = o_p(1)$$

The second to last equality by counting and the final equality by Assumption 2.1. Next consider B_n. Note that each $g \in \sigma(\psi_{1:n}, \pi_n)$ and $E[u_{it}|\psi_{1:n}, \pi_n] = E[u_{it}|\psi_i] = 0$, so $E[B_n|\psi_{1:n}, \pi_n] = 0$. We can rewrite the sum

$$\sum_i \sum_{j,l \in g \setminus \{i\}} (f_{it} - f_{jt})(u_{it'} - u_{lt'}) = \sum_g \sum_{\substack{i,j,l \in g \\ j,l \neq i}} (f_{it} - f_{jt})(u_{it'} - u_{lt'})$$

Then we may compute $\text{Var}(\sqrt{n} B_n|\psi_{1:n}, \pi_n) = E[nB_n^2|\psi_{1:n}, \pi_n]$ as follows. By Lemma 9.21 of Cytrynbaum (2021), $E[u_{it'}u_{jt'}|\psi_{1:n}, \pi_n] = 0$ for any $g(i) \neq g(j)$, so we only consider the diagonal

$$0 \leq \frac{1}{nk^4} \sum_g \sum_{\substack{i,j,l \in g \\ j,l \neq i}} \sum_{\substack{a,b,c \in g \\ b,c \neq a}} E[(f_{it} - f_{jt})(f_{at} - f_{bt})(u_{it'} - u_{lt'})(u_{at'} - u_{ct'})|\psi_{1:n}, \pi_n]$$
$$\leq n^{-1} \sum_g \sum_{\substack{i,j,l \in g \\ j,l \neq i}} \sum_{\substack{a,b,c \in g \\ b,c \neq a}} |f_{it} - f_{jt}||f_{at} - f_{bt}||E[(u_{it'} - u_{lt'})(u_{at'} - u_{ct'})|\psi_{1:n}, \pi_n]|$$
$$\lesssim n^{-1} \sum_g \max_{i,j \in g} |\psi_i - \psi_j|_2^2 \sum_{\substack{i,j,l \in g \\ j,l \neq i}} \sum_{\substack{a,b,c \in g \\ b,c \neq a}} |E[(u_{it'} - u_{lt'})(u_{at'} - u_{ct'})|\psi_{1:n}, \pi_n]|$$

Next, by Lemma 9.21 of Cytrynbaum (2021), $E[(u_{it'} - u_{lt'})(u_{at'} - u_{ct'})|\psi_{1:n}, \pi_n]$ is equal to

$$\delta_{ai} E[u_{it'}^2|\psi_i] - \delta_{la} E[u_{at'}^2|\psi_a] - \delta_{ci} E[u_{it'}^2|\psi_i] + \delta_{lc} E[u_{lt'}^2|\psi_l]$$

Applying the triangle inequality and summing out using this relation, the above is

$$\leq \frac{4k(k-1)^3}{n} \sum_g \max_{i,j \in g} |\psi_i - \psi_j|_2^2 \sum_{i \in g} E[u_{it'}^2|\psi_i]$$
$$\lesssim n^{-1} \sum_g \left(\max_{i,j \in g} |\psi_i - \psi_j|_2^4 + \sum_{i \in g} E[u_{it'}^2|\psi_i]^2 \right)$$
$$\leq n^{-1} \sum_g \text{Diam}(\text{Supp}(\psi))^2 \sum_{i,j \in g} |\psi_i - \psi_j|_2^2 + E_n[E[u_{it'}^2|\psi_i]^2]$$

We claim that $E[u_{it'}^4] < \infty$. Note that $E[u_{it'}^4] = E[(h_{it'} - f_{it'})^4] \leq 8E[h_{it'}^4] + 8E[f_{it'}^4]$ by Young's inequality. We have $E[h_{it'}^4] < \infty$ by assumption. Note that $E[f_{it'}^4] \leq C_f |\psi_i|^4 \leq C_f \,\text{Diam}(\text{Supp}(\psi))^4 < \infty$ by Assumption 8.8, with Lipschitz constant C_f. Then $E[u_{it'}^4] < \infty$, so $E[E_n[E[u_{it'}^2|\psi_i]^2]] = E[E[u_{it'}^2|\psi_i]^2] \leq E[u_{it'}^4] < \infty$. The inequality follows by conditional Jensen and tower law. Then $E_n[E[u_{it'}^2|\psi_i]^2 = O_p(1)$ by

Markov inequality. Then using Assumption 2.1 in the display above, we have shown $E[nB_n^2|\psi_{1:n},\pi_n] = O_p(1)$ and by Lemma 8.4 we have shown $B_n = O_p(n^{-1/2})$. We have $C_n = O_p(n^{-1/2})$ by symmetry. Finally, consider D_n. By Lemma 9.21 of Cytrynbaum (2021) compute $E[(u_{it}-u_{jt})(u_{it'}-u_{lt'})|\psi_{1:n},\pi_n] = E[u_{it}u_{it'}|\psi_i]+E[u_{jt}u_{jt'}|\psi_j]\delta_{jl}$ for $j,l \neq i$. Then

$$E[D_n|\psi_{1:n},\pi_n] = \frac{1}{nk^2}\sum_i \sum_{j,l\in g(i)\setminus\{i\}} E[u_{it}u_{it'}|\psi_i] + E[u_{jt}u_{jt'}|\psi_j]\mathbb{1}(j=l)$$
$$= \frac{1}{nk^2}\sum_i (k-1)^2 E[u_{it}u_{it'}|\psi_i] + \frac{1}{nk^2}\sum_i \sum_{j\in g(i)\setminus\{i\}} E[u_{jt}u_{jt'}|\psi_j]$$
$$= \frac{(k-1)^2}{nk^2}\sum_i E[u_{it}u_{it'}|\psi_i] + \frac{k-1}{nk^2}\sum_i E[u_{it}u_{it'}|\psi_i] = \frac{k(k-1)}{nk^2}\sum_i E[u_{it}u_{it'}|\psi_i]$$

Now $E[E[u_{it}u_{it'}|\psi_i]^2] \leq E[u_{it}^2u_{it'}^2] \leq 2E[u_{it}^4] + 2E[u_{it'}^4] < \infty$ by Jensen, tower law, Young's, and work above. Then by Chebyshev $\frac{(k-1)}{nk}\sum_i E[u_{it}u_{it'}|\psi_i] = \frac{k-1}{k}E[u_{it}u_{it'}] + O_p(n^{-1/2}) = \frac{k-1}{k}E[\mathrm{Cov}(h_{it},h_{it'}|\psi_i)] + O_p(n^{-1/2})$. Then we have shown $E[D_n|\psi_{1:n},\pi_n] = \frac{k-1}{k}E[\mathrm{Cov}(h_{it},h_{it'}|\psi_i)] + O_p(n^{-1/2})$. Next, we claim that $\mathrm{Var}(\sqrt{n}D_n|\psi_{1:n},\pi_n) = O_p(1)$. Following the steps above for B_n replacing terms shows that $\mathrm{Var}(\sqrt{n}D_n|\psi_{1:n},\pi_n)$ is

$$0 \leq \frac{1}{nk^4}\sum_g \sum_{\substack{i,j,l\in g\\ j,l\neq i}} \sum_{\substack{a,b,c\in g\\ b,c\neq a}} \mathrm{Cov}((u_{it}-u_{jt})(u_{it'}-u_{lt'}),(u_{at}-u_{bt})(u_{at'}-u_{ct'})|\psi_{1:n},\pi_n)$$

For any variables A,B, $|\mathrm{Cov}(A,B)| \leq |E[AB]| + |E[A]E[B]| \leq 2|A|_2|B|_2$ by Cauchy-Schwarz and increasing $L_p(\mathbb{P})$ norms. By Young's inequality, $(a-b)^4 \leq 8(a^4+b^4)$ for any $a,b\in\mathbb{R}$. Then using these facts

$$\begin{aligned}
&|\mathrm{Cov}((u_{it}-u_{jt})(u_{it'}-u_{lt'}),(u_{at}-u_{bt})(u_{at'}-u_{ct'})|\psi_{1:n},\pi_n)|\\
\leq& 2E[(u_{it}-u_{jt})^2(u_{it'}-u_{lt'})^2|\psi_{1:n},\pi_n]^{1/2}E[(u_{at}-u_{bt})^2(u_{at'}-u_{ct'})^2|\psi_{1:n},\pi_n]^{1/2}\\
\leq& 4E[(u_{it}-u_{jt})^2(u_{it'}-u_{lt'})^2|\psi_{1:n},\pi_n] + 4E[(u_{at}-u_{bt})^2(u_{at'}-u_{ct'})^2|\psi_{1:n},\pi_n]\\
\leq& 2E[(u_{it}-u_{jt})^4 + (u_{it'}-u_{lt'})^4|\psi_{1:n},\pi_n] + 2E[(u_{at}-u_{bt})^4 + (u_{at'}-u_{ct'})^4|\psi_{1:n},\pi_n]\\
\leq& 16(E[u_{it}^4+u_{jt}^4+u_{it'}^4+u_{lt'}^4|\psi_{1:n},\pi_n] + E[u_{at}^4+u_{bt}^4+u_{at'}^4+u_{ct'}^4|\psi_{1:n},\pi_n])\\
=& 16(2E[u_{it}^4|\psi_i] + E[u_{jt}^4|\psi_j] + E[u_{lt'}^4|\psi_l] + 2E[u_{at}^4|\psi_a] + E[u_{bt}^4|\psi_b] + E[u_{ct'}^4|\psi_c])
\end{aligned}$$

Plugging this bound in above and summing out gives

$$\mathrm{Var}(\sqrt{n}D_n|\psi_{1:n},\pi_n) \leq \frac{32k^5}{nk^4}\sum_g\sum_{i\in g} E[u_{it}^4|\psi_i] \asymp E_n[E[u_{it}^4|\psi_i]] = O_p(1)$$

The final equality by Markov since $E[u_{it}^4] < \infty$. Then by conditional Markov 8.4 we have $D_n = \frac{k-1}{k}E[\mathrm{Cov}(h_{it},h_{it'}|\psi_i)] + O_p(n^{-1/2})$. Since t,t' were arbirary, this shows $E_n[\check{h}_i\check{h}_i'] = E[\mathrm{Var}(h|\psi)] + o_p(1)$.

Next, consider $E_n[D_i\check{h}_iY_i] = E_n[(D_i-p)\check{h}_iY_i(1)] + pE_n[\check{h}_iY_i(1)]$. We claim that $E_n[(D_i-p)\check{h}_iY_i(1)] = O_p(n^{-1/2})$. For $1 \leq t \leq d_h$, by Lemma 9.20 of Cytrynbaum (2021), and

Cauchy-Schwarz

$$\text{Var}(\sqrt{n}E_n[(D_i - p)\check{h}_{it}Y_i(1)]|X_{1:n}, Y(1)_{1:n}, \pi_n) \leq 2E_n[\check{h}_{it}^2 Y_i(1)^2] \leq 2E_n[\check{h}_{it}^4]^{1/2} E_n[Y_i(1)^4]^{1/2}$$

Note that $E_n[Y_i(1)^4] = O_p(1)$ by Markov inequality and Assumption 8.8 and $E_n[\check{h}_{it}^4] = O_p(1)$ was shown above. Then by Lemma 8.4 (conditional Markov), this shows the claim. Then it suffices to analyze $E_n[\check{h}_i Y_i(1)]$. Let $g_i = E[Y_i(1)|\psi_i]$ and $v_i = Y_i(1) - g_i$ with $E[v_i|\psi_i] = 0$. Then as above we may expand

$$\begin{aligned}
E_n[\check{h}_i Y_i(1)] &= \frac{1}{nk}\sum_i \left(\sum_{j \in g(i)\setminus\{i\}} f_{it} - f_{jt} + u_{it} - u_{jt}\right)(g_i + v_i) \\
&= \frac{1}{nk}\sum_i \sum_{j \in g(i)\setminus\{i\}} (f_{it} - f_{jt})g_i + (f_{it} - f_{jt})v_i + (u_{it} - u_{jt})g_i + (u_{it} - u_{jt})v_i \\
&\equiv H_n + J_n + K_n + L_n
\end{aligned}$$

First consider H_n. By Assumption 8.8, $\psi \to g(\psi)$ is continuous and $\text{Supp}(\psi) \subseteq \bar{B}(0, K)$ compact, so $\sup_{\psi \in \bar{B}(0,K)} |g(\psi)| \equiv K' < \infty$ and $|g_i| \leq K'$ a.s. Then we have

$$|H_n| \lesssim n^{-1} \sum_i \sum_{j \in g(i)\setminus\{i\}} |\psi_i - \psi_j|_2 |g_i| \lesssim n^{-1} \sum_g \sum_{i,j \in g} |\psi_i - \psi_j|_2 = o_p(1)$$

For the final equality, note that here we have the unsquared norm, different from Assumption 2.1. The construction in Proposition 8.6 of Cytrynbaum (2021) gave showed that this quantity is also $o_p(1)$ (with rates). By substituting z_i for g_i, which satisfies the same conditions, this also shows that $E_n[z_i \check{h}_i'] = o_p(1)$. The proof that $J_n, K_n = O_p(n^{-1/2})$ are similar to the terms B_n, C_n above. Next, consider L_n. We have

$$\begin{aligned}
E[L_n|\psi_{1:n}, \pi_n] &= \frac{1}{nk}\sum_i \sum_{j \in g(i)\setminus\{i\}} E[(u_{it} - u_{jt})v_i|\psi_{1:n}, \pi_n] \\
&= \frac{1}{nk}\sum_i \sum_{j \in g(i)\setminus\{i\}} E[u_{it}v_i|\psi_i] = \frac{k-1}{k} E_n[E[u_{it}v_i|\psi_i]] \\
&= \frac{k-1}{k} E[\text{Cov}(h_{it}, Y_i(1)|\psi_i)] + O_p(n^{-1/2})
\end{aligned}$$

The second equality follows since $j \neq i$ and by Lemma 9.21 of Cytrynbaum (2021). The third equality by counting. For the last equality, note that by Jensen, tower law, Young's inequality $E[E[u_{it}v_i|\psi_i]^2] \leq E[u_{it}^2 v_i^2] \leq (1/2)(E[u_{it}^4] + E[v_i^4])$. We showed $E[u_{it}^4] < \infty$ above and a similar proof applies to v_i. Then the final equality above follows by Chebyshev. The proof that $\text{Var}(L_n|\psi_{1:n}, \pi_n) = O_p(n^{-1/2})$ is similar to our analysis of D_n above. Then we have shown $E_n[D_i \check{h}_i Y_i] = p\frac{k-1}{k} E[\text{Cov}(h, Y(1)|\psi)] + o_p(1)$. The conclusion for $E_n[(1 - D_i)\check{h}_i Y_i]$ follows by symmetry. This finishes the proof. □