Covariate Adjustment in Stratified Experiments*

Max Cytrynbaum†

August 23, 2024

Abstract

This paper studies covariate adjusted estimation of the average treatment effect in stratified experiments. We work in a general framework that includes matched tuples designs, coarse stratification, and complete randomization as special cases. Regression adjustment with treatment-covariate interactions is known to weakly improve efficiency for completely randomized designs. By contrast, we show that for stratified designs such regression estimators are generically inefficient, potentially even increasing estimator variance relative to the unadjusted benchmark. Motivated by this result, we derive the asymptotically optimal linear covariate adjustment for a given stratification. We construct several feasible estimators that implement this efficient adjustment in large samples. In the special case of matched pairs, for example, the regression including treatment, covariates, and pair fixed effects is asymptotically optimal. We also provide novel asymptotically exact inference methods that allow researchers to report smaller confidence intervals, fully reflecting the efficiency gains from both stratification and adjustment. Simulations and an empirical application demonstrate the value of our proposed methods.

Keywords: Matched Pairs, Analysis of Covariance, Blocking, Robust Standard Error, Treatment Effects.

JEL Codes: C10, C14, C90

*I thank the anonymous referees for helpful suggestions during the revision process.

†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 treatment assignment and ex-post covariate adjustment to improve the precision of experimental estimates. Indeed, out of a survey of over 50 experimental papers published in the AER and AEJ between 2018-2023, we found that 57% use stratified randomization, and 80% used some form of ex-post covariate adjustment. An influential paper by Lin (2013) showed in a design-based setting that the regression estimator with full treatment-covariate interactions is always asymptotically weakly more efficient than difference of means estimation for completely randomized designs. Negi and Wooldridge (2021) extended these results to estimation of the ATE using data sampled from a superpopulation. However, questions remain about the interaction between stratification and regression adjustment and the implications of combining these methods for both estimator efficiency and the power and validity of inference methods. To study these questions, we work in the stratified randomization framework of Cytrynbaum (2023), which includes matched tuples designs (e.g. matched pairs), coarse stratification, and complete randomization as special cases.

We show that the Lin (2013) interacted regression adjustment is generically inefficient in the family of linearly adjusted estimators, with asymptotic efficiency only in the limiting case of complete randomization. Motivated by this finding, we characterize the efficient linear covariate adjustment for a given stratified design, providing several new estimators that achieve the optimal variance.

Our first result derives the optimal linear adjustment coefficient for a given stratification. We show that asymptotically the interacted regression estimator uses the wrong objective function, minimizing a marginal variance objective that is totally insensitive to the stratification. By contrast, the optimal adjustment coefficient minimizes a *mean-conditional* variance objective, conditional on the covariates used to stratify. Intuitively, the efficient covariate adjustment is tailored to the stratification, ignoring fluctuations of the estimator that are predictable by the stratification covariates. Section 3.2 draws an interesting connection with partially linear regression (Robinson (1988)), showing that efficient linear adjustment of a stratified design is asymptotically equivalent to doubly-robust semiparametric adjustment of an iid design. Intuitively, stratification contributes the nonparametric component of the semiparametric adjustment function.

Our second set of results develops feasible versions of the optimal linear adjustment derived in Section 3.1. First, we show that if the conditional expectation of the adjustment covariates is linear in a known set of transformations of the stratification variables, then adding the latter to the interacted regression restores optimality. Next we relax this assumption, providing four different regression estimators that are asymptotically efficient under weak conditions. For matched pairs experiments or in settings with lim-

ited treatment effect heterogeneity, the non-interacted regression with a full set of pair fixed effects is asymptotically efficient. More generally, we show asymptotic optimality of within-stratum (inconsistently) partialled versions of the Lin and tyranny-of-the-minority estimators (Lin (2013)). We also define a "group OLS" estimator, extending a proposal of Imbens and Rubin (2015) for matched pairs experiments to a larger class of designs. We show that this group OLS estimator is also asymptotically optimal.

Our final contribution is to develop novel asymptotically exact inference methods for covariate adjusted estimation under stratified designs. Confidence intervals based on the usual heteroskedasticity robust variance estimator are known to be conservative in stratified experiments (Bai et al. (2021)). By contrast, the coverage probabilities of our proposed confidence intervals converge to the specified nominal level, with no overcoverage. Our approach applies to a generic family of linear covariate adjustments and randomization schemes, including as special cases non-interacted regression adjustment, the Lin (2013) interacted regression, and all of the other estimators considered in this paper. Simulations and an empirical application to the experiment in Baysan (2022a) suggest that the usual robust confidence intervals can substantially overcover in stratified experiments, while our confidence intervals have close to nominal coverage.

We present several extensions of our main results in the appendix. In the first, we consider estimation and inference in stratified experiments with noncompliance. As a simple corollary of our results on ATE estimation, we characterize the optimal linearly adjusted Wald estimator for the LATE (Imbens and Angrist (1994)), construct a feasible implementation of the efficient adjustment, and provide asymptotically exact inference methods. We also study efficient linear adjustment for finely stratified designs with non-constant treatment proportions, as in Cytrynbaum (2023), and briefly consider the problem of efficient nonlinear adjustment.

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 (2024), Ma et al. (2022), Reluga et al. (2024), Wang et al. (2021), Ye et al. (2022), Zhu et al. (2024), and Chang (2023). 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 superpopulation, (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. (2024) shows asymptotic efficiency of a projection-based estimator numerically equivalent to the "partialled Lin" approach considered in Section 3.4.2. In the same setting, Lu and Liu (2024) prove efficiency of a tyranny-of-the-minority style regression similar but not equivalent to one the considered in Section 3.4.4. Both papers give 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. Remarks 3.19 and 3.22 in Section 3.4 below provide a detailed comparison.

Relative to the above papers, the superpopulation framework considered here creates some 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 similar to Cytrynbaum (2022). This setting also has analytical advantages, which allow us to establish new conceptual results. For example, the population level characterization of the optimal adjustment coefficent in Section 3.1 allows us to give explicit necessary and sufficient conditions for the efficiency of several commonly used regression estimators. The efficiency of interacted regression under a "rich covariates" condition, as well as the equivalence between optimal linear adjustment of stratified designs and doubly-robust semiparametric adjustment appear to be new observations in this literature. To the best of our knowledge, we give the first asymptotically exact inference on the ATE for general covariate adjusted estimators under finely stratified randomization.

Independently, Bai et al. (2024b) study covariate adjustment under matched pairs randomization in a superpopulation framework. They also find that regression adjustment without pair fixed effects may be inefficient, while adding pair fixed effects restores efficiency. Relative to our work, they additionally study regularized regression adjustment under high-dimensional asymptotics, which we do not consider. By contrast, we study more general forms of stratification, allowing coarse and fine stratification with arbitrary treatment proportions $p \neq 1/2$. For such designs, the strata fixed effects estimator may still be inefficient. To fix this, we introduce novel forms of linear adjustment that are efficient under general stratified designs.

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 both simulations and an empirical application to the experiment in Baysan (2022a). 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 superpopulation 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 (2022).

Definition 2.1 (Local Randomization)**.** Let treatment proportions $p = a/k$ with $\gcd(a, k) = 1$.[1] Suppose that n is divisible by k for notational simplicity. Partition the experimental units into n/k groups g with $\{1, \ldots, n\} = \bigcup_g g$ disjointly and $|g| = k$. Let $\psi(X) \in \mathbb{R}^{d_\psi}$ denote a vector of stratification variables. Suppose that the groups that satisfy a homogeneity condition with respect to $\psi(X)$ such that

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

Require that the groups only depend on the stratification variables $\psi_{1:n}$ and data-independent randomness π_n, so that $g = g(\psi_{1:n}, \pi_n)$ for each g. 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 $D_{1:n} \sim \text{Loc}(\psi, p)$.

Example 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 (2023) provides an iterative pairing algorithm to match units into groups that provably satisfy this condition for any k. 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 $p = 1/2$.

Example 2.3 (Complete Randomization)**.** We say variables $D_{1:n}$ are completely randomized with treatment probability p if $D_{1:n}$ is drawn uniformly from all vectors $d_{1:n}$ with $d_i = 1$ for exactly proportion p of the units. Formally, $P(D_{1:n} = d_{1:n}) = 1/\binom{n}{pn}$ for all such vectors. We denote complete randomization by $D_{1:n} \sim \text{CR}(p)$. Complete randomization may be obtained in our framework by setting $\psi = 1$ and forming groups $|g| = k$ at random, which automatically satisfies Equation 2.1. For example, assigning 2 out of 3 units in each group to treatment gives a "random matched triples" representation of complete randomization with $p = 2/3$.

Remark 2.4 (Coarse Stratification)**.** Similarly, coarse stratification with large fixed strata $S(X) \in \{1, \ldots, m\}$ can also be obtained in our framework by setting $\psi(X) = S(X)$ and matching units with identical $S(X)$ values into groups at random. Because of this, our framework enables a unified asymptotic analysis for a wide range of stratifications.

Experiment Timing: Suppose that the experimenter does the following

[1] $\gcd(a, k)$ stands for greatest common divisor.

(1) Samples units and observes their baseline covariates.

(2) Partitions the units into data-dependent groups $g = g(\psi_{1:n}, \pi_n)$ 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 the ATE, potentially adjusting for covariates $h(X)$.

We are agnostic about the exact time at which the covariates are observed, subject to the constraints above. For example, it could be that only $\psi(X)$ is observed at the design stage, while the full vector X is collected later with the outcomes, and the experimenter chooses to adjust for $h(X) \subseteq X$. Alternatively, the full vector X could be observed at the design stage, but the experimenter chooses to only stratify on $\psi(X)$, and adjusts for $h(X) \subseteq X$ at step (3). We may or may not have $\psi(X) \subseteq h(X)$.[2]

Consider the unadjusted estimator given by the coefficient $\widehat{\theta}$ on D in the regression $Y \sim 1 + D$. Before discussing covariate adjustment, we first state a helpful variance decomposition for $\widehat{\theta}$ 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 (2022) shows that if $D_{1:n} \sim \text{Loc}(\psi, p)$ then $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with

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

The variance V is in fact the Hahn (1998) semiparametric variance bound[3] for the ATE (with covariates $\psi(X)$), providing 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, in this notation the difference in asymptotic 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 to covariates $h(X)$, which may strictly contain $\psi(X)$. One may try to further improve the efficiency of ATE estimation by regression adjustment

[2] Our asymptotic framework lets $h(X)$, $\psi(X)$ be fixed as $n \to \infty$.

[3] Armstrong (2022) shows that this variance bound also holds for stratified designs.

using these covariates, either using standard the regression $Y \sim 1+D+h$ or the regression $Y \sim 1 + D + h + Dh$ (with de-meaned covariates) studied in Lin (2013). We study 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 of the ATE under stratified randomization. Lin (2013) showed that in a completely randomized experiment, equivalent to $D_{1:n} \sim \text{Loc}(\psi, p)$ with $\psi = 1$, regression adjustment with full treatment-covariate interactions is asymptotically weakly more efficient than difference of means estimation. Negi and Wooldridge (2021) extended this result to ATE estimation in the superpopulation framework that we use in this paper. 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, minimizing a marginal variance objective when, due to the stratification, it should instead minimize a mean-conditional variance objective, conditional on the stratification variables ψ. In fact, the Lin estimator is totally insensitive to the stratification, estimating the same adjustment coefficient for any stratified design $D_{1:n} \sim \text{Loc}(\psi, p)$. Because of this, interacted regression is generically sub-optimal and in some cases can even be strictly less efficient than difference of means. Before proceeding, we state our main assumption.

Assumption 3.1 (Smoothness and Moment Conditions)**.** *Assume the following:*

(i) The conditional expectations $E[h(X)|\psi]$ and $E[Y(d)|\psi]$ for $d \in \{0, 1\}$ are Lipschitz continuous in the stratification variables ψ.

(ii) The moments $E[Y(d)^4] < \infty$ for $d \in \{0, 1\}$ and $E[|h_t(X)|^4] < \infty$ for all $1 \leq t \leq \dim(h)$, $|\psi(X)|_2 < K < \infty$ a.s. and $\text{Var}(h) \succ 0$.

Now we are ready to define the Lin estimator and state our first result. Denote $h_i = h(X_i)$ and de-meaned covariates $\tilde{h}_i = h_i - E_n[h_i]$, with $E_n[h_i] \equiv n^{-1} \sum_{i=1}^n h_i$. The Lin estimator $\widehat{\theta}_L$ is the coefficient on D_i in the interacted regression

$$Y_i \sim 1 + D_i + \tilde{h}_i + D_i \tilde{h}_i. \tag{3.1}$$

Define the within treatment arm covariate means $\bar{h}_1 = E_n[h_i D_i]/E_n[D_i]$ and $\bar{h}_0 = E_n[h_i(1 - D_i)]/E_n[1 - D_i]$. The Lin estimator $\widehat{\theta}_L$ can be related to the difference of

means estimator $\widehat{\theta}$ as

$$\widehat{\theta}_L = \widehat{\theta} - \widehat{\gamma}_L'(\bar{h}_1 - \bar{h}_0). \tag{3.2}$$

Here, the *adjustment coefficient* $\widehat{\gamma}_L$ is $\widehat{\gamma}_L = (1-p)(\widehat{a}_1 + \widehat{a}_0) + p\widehat{a}_0$, where $\widehat{a}_0$ and $\widehat{a}_1$ are the coefficients on $\tilde{h}_i$ and $D_i\tilde{h}_i$ in Equation 3.1. The following theorem characterizes the asymptotic properties of this estimator under stratified designs.

Theorem 3.2. *Let Assumption 3.1 hold. If* $D_{1:n} \sim \text{Loc}(\psi, p)$ *then the Lin estimator* $\sqrt{n}(\widehat{\theta}_L - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ *with*

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

The adjustment coefficient satisfies $\widehat{\gamma}_L \xrightarrow{p} \gamma_L$ *with* $\gamma_L = \text{argmin}_{\gamma \in \mathbb{R}^{d_h}} \text{Var}(b - \gamma' h)$.

The variance V differs from the variance of the unadjusted estimator only in the middle term, which changes from $E[\text{Var}(b|\psi)]$ in the unadjusted case to $E[\text{Var}(b - \gamma_L' h|\psi)]$ for the interacted regression. Crucially, the second statement of Theorem 3.2 shows that the adjustment coefficient γ_L attempts to minimize a marginal variance, instead of the mean-conditional variance that shows up in V above. Because of this, the estimator may be inefficient for general stratifications $\psi \neq 1$, since in general

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

Observe that the Lin estimator is completely insensitive to the experimental design, estimating the same adjustment coefficient $\gamma_L = \text{argmin}_\gamma \text{Var}(b - \gamma' h)$ for any stratification variables $\psi(X)$. The following example shows that this can lead to strict inefficiency relative to difference of means estimation.

Example 3.3 (Random Assignment to Class Size)**.** Suppose $Y(d)$ are student test scores under random assignment to one of two class sizes $d \in \{0, 1\}$. Let $h(X)$ be parent's wealth and $\psi(X)$ previous year (baseline) test scores. Suppose parent's wealth is predictive of future test scores marginally so that $\text{Cov}(h, Y(d)) > 0$. Then $\text{Cov}(h, b) > 0$ and the Lin coefficient is $\gamma_L = \text{Var}(h)^{-1}\text{Cov}(h, b) > 0$. However, if on average parent's wealth has no predictive power for test scores *conditional* on a student's baseline scores (a proxy for ability) then $E[\text{Cov}(h, Y(d)|\psi)] = 0$. In this case, regression adjustment for parent's wealth $h(X)$ in an experiment stratified on the earlier scores $\psi(X)$ will be strictly less efficient than unadjusted estimation since

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

An important special case occurs when the design is completely randomized ($\psi = 1$) or if the covariates and stratification variables are independent $h(X) \perp\!\!\!\perp \psi(X)$. In this case, the Lin estimator is weakly more efficient than difference of means since we have

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

An analogue of Theorem 3.2 also holds for the non-interacted regression estimator $Y_i \sim 1 + D_i + h_i$ under stratified designs $D_{1:n} \sim \mathrm{Loc}(\psi, p)$. The non-interacted estimator is known to be inefficient relative to difference of means even for completely randomized experiments unless $p = 1/2$ or treatment effects are homogeneous. For completeness, we give asymptotic theory and optimality conditions for this estimator under stratified randomization in Section A.3 in the appendix.

We noted above that the Lin estimator $\widehat{\theta}_L$ can be written in the canonical form $\widehat{\theta}_L = \widehat{\theta} - \widehat{\gamma}_L'(\bar{h}_1 - \bar{h}_0)$. In fact, most commonly used adjusted estimators can be written in the standard form $\widehat{\theta}_{adj} = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)$ for some $\widehat{\gamma}$, up to order $O_p(n^{-1})$ factors. The following theorem describes the asymptotic properties of general covariate-adjusted estimators $\widehat{\theta}_{adj}$ of this form. To avoid carrying around factors of p in our variance expressions, in what follows we scale adjusted estimators by the normalization constant $c_p = \sqrt{p(1-p)}$.

Theorem 3.4. *Let Assumption 3.1 hold. Suppose $\widehat{\gamma} \xrightarrow{p} \gamma$ and consider the adjusted estimator*

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

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

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

We define a linearly-adjusted estimator to be asymptotically efficient if it globally minimizes the asymptotic variance $V(\gamma)$ in the previous theorem.

Definition 3.5 (Optimal Linear Adjustment)**.** The estimator $\widehat{\theta}_{adj} = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p$ is *efficient* for the design $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ and covariates $h(X)$ if $\widehat{\gamma} \xrightarrow{p} \gamma^*$ for an optimal adjustment coefficient

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

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 $h(X)$. Setting $\gamma = 0$ recovers unadjusted estimation, so any optimal estimator is in particular weakly more efficient than difference of means.

Optimal Adjustment Coefficient. If $E[\mathrm{Var}(h|\psi)] \succ 0$, then the optimization in

Definition 3.5 is solved uniquely by a mean-conditional OLS coefficient

$$\gamma^* = E[\mathrm{Var}(h|\psi)]^{-1} E[\mathrm{Cov}(h, b|\psi)]. \tag{3.4}$$

Intuitively, fine stratification makes treatment-control imbalances in the covariates $h(X)$ and the potential outcomes $Y(d)$ that are predictable by ψ small enough that they do not contribute to first-order asymptotic variance. Because of this, the optimal covariate-adjusted estimator $\widehat{\theta} - \gamma^*(\bar{h}_1 - \bar{h}_0)c_p$ ignores such fluctuations, minimizing the *mean-conditional* variance objective $E[\mathrm{Var}(b - \gamma' h|\psi)]$, instead of the marginal variance $\mathrm{Var}(b - \gamma' h)$ targeted by the Lin estimator.

Optimal Covariates. Intuitively, the form of the variance in Equation 3.3 suggests adjusting for variables h that contain predictive information not already contained in ψ. The (unknown) optimal covariates are $h^* = b$. In this case, $\gamma^* = 1$ makes the middle variance term identically zero, and $\widehat{\theta}_{adj}$ achieves the Armstrong (2022) semiparametric variance bound.

Sample Average Treatment Effect. Theorem 3.4 may be extended to covariate-adjusted estimation of the sample average treatment effect $\mathrm{SATE} = E_n[Y_i(1) - Y_i(0)]$. Defining the conditional treatment effect variance $\sigma^2_\tau(X) = \mathrm{Var}(Y(1) - Y(0)|X)$, one can show that $\sqrt{n}(\widehat{\theta} - \mathrm{SATE}) \Rightarrow \mathcal{N}(0, V_S(\gamma))$ with

$$V_S(\gamma) = E[\mathrm{Var}(b - \gamma' h|\psi)] + E\left[\frac{\sigma^2_1(X)}{p} + \frac{\sigma^2_0(X)}{1-p} - \sigma^2_\tau(X)\right]. \tag{3.5}$$

In particular, the optimal adjustment for estimating the ATE and the SATE are the same, with $\gamma^*_{SATE} = \gamma^*$.

Remark 3.6 (Non-Uniqueness)**.** In general, the optimal adjustment coefficient γ^* may not be unique. For example, if $h(x) = (z(\psi), w(x))$ with $z(\psi)$ a Lipschitz function of the stratification variables, then the variance objective is constant in the coefficient on $z(\psi)$

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

In fact, our analysis shows that the adjustment term $\gamma_z'(\bar{z}_1 - \bar{z}_0) = o_p(n^{-1/2})$ for any coefficient γ_z in this case. Intuitively, since the covariate $z(\psi)$ is already finely balanced by stratifying on $\psi(X)$, ex-post adjustment by $z(\psi)$ cannot improve first-order efficiency. However, there may still be finite sample efficiency gains from such adjustments, if the covariates $z(\psi)$ are not completely balanced by the stratification. Section 3.5 below provides methods to further adjust for covariates that are functions of the stratification variables.

3.1.1 Extensions

Before continuing, we briefly mention some extensions to the framework above that are studied in detail in Appendices A.1-A.4.

Experiments with Noncompliance. In settings with noncompliance, we may instead consider estimation and inference on the local average treatment effect (LATE) of Imbens and Angrist (1994). As a simple application of our main results, Section A.1 characterizes the optimal linear adjustment for estimating the LATE, constructs feasible efficient estimators, and provides asymptotically exact inference on the LATE under stratified randomization with ex-post covariate adjustment.

Varying Treatment Proportions. Cytrynbaum (2023) extends Definition 2.1 to allow fine stratification with non-constant assignment propensity $p(\psi)$. Section A.2 in the appendix characterizes the optimal adjustment coefficient for such designs and derives a feasible efficient estimator.

Nonlinear Adjustment. In some settings, it may be more natural to use nonlinear or nonparametric covariate adjustment to improve efficiency, for example in experiments with binary outcomes. Section A.4 in the appendix characterizes the optimal adjustment over a general function space $\mathcal{H}$ for finely stratified designs with varying propensity $p(\psi)$. Feasible estimation of the optimal nonlinear adjustment is an interesting problem that we leave for future work.

3.2 Equivalence with Partially Linear Regression Adjustment

This section shows that optimal linear adjustment of a stratified design is as efficient as semiparametric *partially linear* regression adjustment in an experiment with iid treatments, with adjustment function that is linear in $h(X)$ and nonparametric in $\psi(X)$. This suggests that experimenters stratify on a small set of covariates expected to be most predictive of outcomes at design-time, and (efficiently) adjust for the remaining covariates ex-post. See below for a more detailed discussion of stratification vs. adjustment.

The main result of this section shows first-order asymptotic equivalence of the following (design, estimator) pairs

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

To define the latter, consider the within-arm partially linear regression models

$$(g_d^*, \gamma_d^*) = \underset{g \in L_2(\psi), \gamma \in \mathbb{R}^{d_h}}{\text{argmin}} \; E[(Y(d) - g(\psi) - \gamma' h)^2] \tag{3.6}$$

for $d \in \{0, 1\}$. Define the partially linear adjustment function $F_d(x) = g_d^*(\psi(x)) + h(x)'\gamma_d^*$

and consider a Robins and Rotnitzky (1995) style augmented inverse propensity weighting (AIPW) estimator

$$\widehat{\theta}_{AIPW} = E_n[F_1(X_i) - F_0(X_i)] + E_n\left[\frac{D_i(Y_i - F_1(X_i))}{p}\right] - E_n\left[\frac{(1-D_i)(Y_i - F_0(X_i))}{1-p}\right].$$

The next theorem shows that optimal linear adjustment of the design $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ is asymptotically equivalent to optimal semiparametric adjustment with nonparametric $\psi(X)$ and linear $h(X)$ components.

Theorem 3.7. *Require Assumption 3.1 and suppose* $D_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. *Then* $\sqrt{n}(\widehat{\theta}_{AIPW} - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$ *with*

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

The limiting variance V^* is the same as the optimal linearly adjusted variance $V(\gamma^*)$ from Definition 3.5. Intuitively, stratification contributes the nonlinear component of the optimal model $F_d(x)$ above, while optimal adjustment contributes the linear component. The optimal adjustment coefficient $\gamma^* = \sqrt{\frac{1-p}{p}}\gamma_1 + \sqrt{\frac{p}{1-p}}\gamma_0$, for partially linear coefficients γ_1, γ_0 defined in Equation 3.6 above.

Stratification vs. Regression Adjustment. Theorem 3.7 shows that stratification provides nonparametric control over the fluctuations of the outcomes predictable by $\psi(X)$, while (linear) adjustment only provides linear control. In first-order asympotics, this suggests that we stratify on all available covariates, since the variance V^* above is minimized by setting $\psi(X) = X$. However, this may perform poorly in finite samples due to a curse of dimensionality for stratification as $\dim(\psi)$ increases. For example, Cytrynbaum (2023) shows the variance convergence rate $n\,\mathrm{Var}(\widehat{\theta}) = V + O_p(n^{-2/(\dim(\psi)+1)})$ for the variance V in Equation 3.3, which may be slow even for moderate $\dim(\psi)$. Intuitively, this suggests stratifying on a small set[4] of covariates $\psi(X)$ expected to be most predictive of outcomes at design time, and planning to optimally adjust for less predictive covariates $h(X)$ ex-post.

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

This section provides a "rich covariates" style condition on the relationship between adjustment covariates and stratification variables under which a simple parametric correc-

[4]It is difficult to give concrete guidance for choosing $\dim(\psi)$, since the relevant quantities such as $E[\mathrm{Var}(b|\psi)]$ are not estimable at design-time, before we have outcome data. The rate above suggests $\dim(\psi) = o(\log n)$ to achieve the variance V in Equation 3.3.

tion of the Lin estimator is fully efficient. The basic idea is to include rich functions $z(\psi)$ of the stratification variables in the adjustment set alongside the additional covariates we would like to adjust for ex-post. The main result of this section shows that including $z(\psi)$ as covariates forces the Lin estimator to solve the mean-conditional variance minimization problem of Definition 3.5, restoring asymptotic optimality. An analogous result holds for the non-interacted regression estimator $Y \sim 1+D+h$ if $p = 1/2$. As a simple application of this section's results, Example 3.12 shows that for coarsely stratified designs the Lin estimator with leave-one-out strata indicators is efficient.

Consider adjusting for covariates $h(X) = (w(X), z(\psi))$. The main assumption of this section requires that the conditional mean $E[w|\psi]$ is well-approximated by known 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 3.1 and 3.8 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, $\sqrt{n}(\widehat{\theta}_L - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$ with*

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

Moreover, the asymptotic variance has

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

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 A.6 of the supplement. The following corollary follows shows that if $p = 1/2$ (matched pairs) or if treatment effect heterogeneity is limited then the non-interacted regression $Y \sim 1 + D + w + z$ with rich strata controls $z(\psi)$ is also asymptotically efficient.

Corollary 3.10. *Suppose additionally that $p = 1/2$ or $E[\mathrm{Cov}(Y(1) - Y(0), w|\psi)] = 0$. Then the coefficient $\widehat{\theta}_N$ on D_i in the regression $Y \sim 1 + D + w + z$ is asymptotically efficient.*

The condition $E[\mathrm{Cov}(Y(1) - Y(0), w|\psi)] = 0$ limits the explanatory power of covariates w for treatment effect heterogeneity, conditional on the stratification variables.

Remark 3.11 (Indirect Efficiency Gain)**.** The second statement of the theorem shows that optimal adjustment for $h(X)$ is as efficient as optimal adjustment for the subvector

$w(X) \subseteq h(X) = (w(X), z(\psi))$. In this sense, the efficiency improvement due to including $z(\psi)$ is indirect. Indeed, our analysis shows that $\widehat{\theta} - \gamma_z'(\bar{z}_1 - \bar{z}_0) = \widehat{\theta} + o_p(n^{-1/2})$ for any $\gamma_z \in \mathbb{R}^{d_z}$, so adjustment for $z(\psi)$ alone cannot affect the first-order asymptotic variance. Intuitively, we are just using the inclusion of $z(\psi)$ as a device to "tilt" the coefficient on $w(X)$, forcing the Lin estimator to solve the correct mean-conditional variance optimization problem.

The next example uses Theorem 3.9 to show that including leave-one-out strata indicators as covariates in the Lin estimator restores asymptotic efficiency for coarsely stratified designs.

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 the adjustment covariates be $h(x) = (w(x), z(s))$ with leave-one-out strata indicators $z(S_i) = (\mathbb{1}(S_i = k))_{k=1}^{m-1}$. In this case, Assumption 3.8 is automatically satisfied since we can write $E[w|S] = c + \Lambda z$ with $c = E[w|S = m]$ and $\Lambda_{jk} = (E[w_j|S = k] - E[w_j|S = m])_{jk}$. Then by Theorem 3.9, the Lin estimator $\widehat{\theta}_L$ with covariates $h_i = (w_i, z_i)$ is efficient. In particular, we have $\sqrt{n}(\widehat{\theta}_L - \text{ATE}) \Rightarrow \mathcal{N}(0, V^*)$ with optimal variance

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

Similarly, by Corollary 3.10 if $p = 1/2$ then including leave-one-out strata fixed effects in the non-interacted regression restores efficiency.

Remark 3.13 (Fine Stratification)**.** Note that the argument in 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 number of strata $m \asymp n$, so Theorem 3.23 does not apply. Indeed, for matched pairs the Lin regression in Example 3.12 would have $n + 2\dim(h) > n$ covariates, producing collinearity. This collinearity problem occurs more generally, see Remark 3.18 below for further discussion.

Leaving behind the rich covariates Assumption 3.8, the next section provides new adjusted estimators that are fully efficient for any design in the class $D_{1:n} \sim \text{Loc}(\psi, p)$ under weak conditions.

3.4 Generic Efficient Adjustment

In this section, we study several adjusted estimators that are asymptotically efficient under weak conditions for any stratified design $D_{1:n} \sim \text{Loc}(\psi, p)$. For matched pairs designs, or in settings with limited treatment effect heterogeneity, the non-interacted regression

including treatment, covariates, and pair fixed effects is efficient. More generally, we show that the following estimators are efficient under weak assumptions.

(1) **PL** - A partialled Lin estimator with within-stratum (inconsistently) partialled covariates.

(2) **GO** - A "Group OLS" estimator, generalizing a proposal of Imbens and Rubin (2015) for matched pairs designs.

(3) **TM** - A tyranny-of-the-minority (ToM) estimator for stratified designs.

The main new condition we impose in this section is that the adjustment covariates are not collinear, conditionally on the stratification variables. This guarantees that the optimal adjustment coefficient γ^* is unique with $\gamma^* = E[\mathrm{Var}(h|\psi)]^{-1}E[\mathrm{Cov}(h,b|\psi)]$, as discussed in Section 3.1.

Assumption 3.14. *The conditional variance satisfies* $E[\mathrm{Var}(h|\psi)] \succ 0$.

Note that this assumption rules out adjustment for functions $h(\psi)$ of the stratification variables. To see why it is necessary, consider that, for example, in a regression with full strata fixed effects $Y \sim D + h + z^n$, covariates $h_i = h(\psi_i)$ would be asymptotically collinear with the strata fixed effects $z^n = (\mathbb{1}(i \in g_j))_{j=1}^{n/k}$. More intuitively, the problem is that $h(\psi)$ has too little residual variation within local regions of $\psi(X)$ space defining the fine strata. We noted earlier that $\widehat{\theta} - \alpha'(\bar{h}_1 - \bar{h}_0) = \widehat{\theta} - o_p(n^{-1/2})$ for any $\alpha \in \mathbb{R}^{d_h}$, so such adjustment cannot improve first-order efficiency. Nevertheless, one may still wish to adjust for $h(\psi)$ to correct finite sample imbalances not controlled by the design. Adjustment for such variables needs to be handled slightly differently, and we construct modified efficient estimators for this purpose in Section 3.5 below.

3.4.1 Strata Fixed Effects Estimator

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

$$Y_i = \widehat{\theta}_{FE} D_i + \widehat{\gamma}'_{FE} h_i + \sum_{j=1}^{n/k} \widehat{a}_j \mathbb{1}(i \in g_j) + e_i. \tag{3.7}$$

The next theorem shows that $\widehat{\theta}_{FE}$ is fully efficient in the case of matched pairs or if treatment effect heterogeneity is limited, but may be inefficient in general.

Theorem 3.15. *Suppose Assumptions 3.1 and 3.14 hold. The estimator has representation* $\widehat{\theta}_{FE} = \widehat{\theta} - \widehat{\gamma}'_{FE}(\bar{h}_1 - \bar{h}_0) + O_p(n^{-1})$. *If* $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ *then* $\sqrt{n}(\widehat{\theta}_{FE} - \mathrm{ATE}) \Rightarrow$

$\mathcal{N}(0, V)$ *with variance*

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

and coefficient $\gamma_{FE} = \text{argmin}_{\gamma \in \mathbb{R}^{d_h}} E[\text{Var}(f - \gamma' h|\psi)]$ *for target function*

$$f(x) = m_1(x)\sqrt{\frac{p}{1-p}} + m_0(x)\sqrt{\frac{1-p}{p}}.$$

The function $f \neq b$ *in general. If* $p = 1/2$*, then* $f = b$ *and the fixed effects estimator is efficient. If* $p \neq 1/2$*, it is efficient if and only if* $E[\text{Cov}(h, Y(1) - Y(0)|\psi)] = 0$*.*

See Section A.7 for the proof. Asymptotically exact inference for the ATE using $\widehat{\theta}_{FE}$ is available using the tools in Section 4.

Remark 3.16 (Conditions for Efficiency)**.** If $p = 1/2$ then $f = b$ and $\widehat{\theta}_{FE}$ is efficient. More generally, $f(x) \neq b(x)$ and $\widehat{\theta}_{FE}$ solves the wrong variance minimization problem, effectively targeting the wrong linear combination of outcomes. The necessary and sufficient condition $E[\text{Cov}(h, Y(1) - Y(0)|\psi)] = 0$ requires that treatment effect heterogeneity is not explained by the covariates $h(X)$, conditional on the stratification variables.

In the rest of this section, we develop estimators that are fully efficient for any finely stratified design, without imposing any assumptions on treatment effect heterogeneity or treatment proportions.

3.4.2 Partialled Lin Estimator

First, we define a partialled version of the Lin estimator. Let $g(i)$ denote the group that unit i belongs to and define the within-group partialled covariates

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

For example, if $k = 2$ this is just the within-pair covariate difference $\check{h}_i = (1/2)(h_i - h_{m(i)})$, where i is matched to $m(i)$. We can think of $\check{h}_i$ as an inconsistent but approximately unbiased signal for the non-parametrically residualized covariate $h_i - E[h_i|\psi_i]$. Next, we use these partialled covariates in the Lin regression

$$Y \sim 1 + D_i + \check{h}_i + D_i\check{h}_i. \tag{3.8}$$

Define the *partialled* Lin estimator $\widehat{\theta}_{PL}$ to be the coefficient on D_i in this regression. For reference, similarly to the Lin regression we may write this in the standard form $\widehat{\theta}_{PL} = \widehat{\theta} - \widehat{\gamma}_{PL}'(\bar{h}_1 - \bar{h}_0)c_p$ with adjustment coefficient $\widehat{\gamma}_{PL} = (\widehat{a}_1 + \widehat{a}_0)\sqrt{\frac{1-p}{p}} + \widehat{a}_0\sqrt{\frac{p}{1-p}}$,

where $\widehat{a}_0$ and $\widehat{a}_1$ are coefficients on $\check{h}_i$ and $D_i\check{h}_i$.

Our main result in Theorem 3.23 below shows that the partialled Lin estimator $\widehat{\theta}_{PL}$ is asymptotically efficient in the sense of Definition 3.5, with $\widehat{\gamma}_{PL} \xrightarrow{p} \gamma^*$ for the optimal adjustment coefficient γ^*.

Remark 3.17 (Intuition for Optimality)**.** Theorem 3.4 showed that an estimator $\widehat{\theta} - \widehat{\gamma}(\bar{h}_1 - \bar{h}_0)c_p$ is efficient if $\widehat{\gamma} \xrightarrow{p} \gamma^*$ and γ^* solves the conditional-mean variance problem $\gamma^* \in \operatorname{argmin}_\gamma E[\operatorname{Var}(b - \gamma' h|\psi)]$. By using within-stratum partialled regressors $\check{h}_i$, we force the Lin estimator to only use covariate signal $h_i - E[h_i|\psi_i]$ that is mean-independent of the stratification variables.

Remark 3.18 (Treatment-Strata Interactions)**.** As an alternative to $\widehat{\theta}_{PL}$, one may attempt to use the Lin regression $Y_i \sim (1, h_i, g^n(i)) + D_i(1, h_i, g^n(i))$ with leave-one-out strata fixed effects $g^n(i) = (\mathbb{1}(i \in g_j))_{j=1}^{n/k-1}$. Unfortunately, this produces collinear regressors for $p = a/k$ if either $a = 1$ or $a = k - 1$, which includes the case of matched pairs. To see the issue, one can show by Frisch-Waugh that in contrast to Equation 3.8 above, this estimator partials covariates h_i separately in each treatment arm, using $\check{h}_{i1} = h_i - a^{-1}\sum_{j\in g(i)} h_j D_j$ if $D_i = 1$ and $\check{h}_{i0} = h_i - (k-a)^{-1}\sum_{j\in g(i)} h_j(1 - D_j)$ if $D_i = 0$. For instance, if $a = 1$ then this is $\check{h}_i = h_i - h_i = 0$ for all i, showing collinearity. In the case $1 < a < k - 1$ where this estimator is feasible, it is asymptotically equivalent to the partialled Lin estimator. However, finite sample properties will be worse due to noisier within-arm partialling.

Remark 3.19. A calculation shows that our estimator $\widehat{\theta}_{PL}$ is numerically equivalent to a regression estimator proposed in Zhu et al. (2024), which the authors derive alternately through an optimal projection argument. They study estimation of the SATE under stratified randomization in a finite population framework, providing conservative inference. They do not derive the exact form of the asymptotic variance, instead leaving it as an infinite sum, which they assume converges to some limit. By contrast, we derive the exact form of the asymptotic variance under the data-adaptive stratifications in Definition 2.1, enabling asymptotically exact inference on the ATE using $\widehat{\theta}_{PL}$.

3.4.3 Group OLS Estimator

Next, we generalize an estimator proposed by Imbens and Rubin (2015) for covariate adjustment in matched pairs experiments to more general stratified designs. For each group of units $g = 1, \ldots, n/k$ in the design $D_{1:n} \sim \operatorname{Loc}(\psi, p)$, define the within-group difference of means of outcomes and covariates

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

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

$$y_g = \widehat{\theta}_G + \widehat{\gamma}_G' h_g + e_g \tag{3.9}$$

with $E_g[(1, h_g)e_g] = 0$. For motivation, note that if $h = 0$ then this becomes $y_g = \widehat{\theta}_G + e_g$ and $\widehat{\theta}_G$ is just the unadjusted estimator $\widehat{\theta}_G = \bar{Y}_1 - \bar{Y}_0$. More generally, the adjusted version can be written $\widehat{\theta}_G = E_g[y_g] - \widehat{\gamma}_G' E_g[h_g] = \widehat{\theta} - \widehat{\gamma}_G'(\bar{h}_1 - \bar{h}_0)$ with adjustment coefficient $\widehat{\gamma}_G = \mathrm{Var}_g(h_g)^{-1}\,\mathrm{Cov}_g(h_g, y_g)$. The estimators $\widehat{\theta}_G$ and $\widehat{\theta}_{PL}$ are numerically identical for the case of matched pairs, but not for $p \neq 1/2$. The main result of this section shows that $\widehat{\theta}_G$ is asymptotically equivalent to the partialled Lin estimator $\widehat{\theta}_{PL}$, and both are asymptotically optimal.

Remark 3.20 (Intuition for Efficiency)**.** The estimator $\widehat{\theta}_G$ uses within-group differences of covariates $\bar{h}_{g1} - \bar{h}_{g0}$ to predict within-group outcome differences $\bar{Y}_{1g} - \bar{Y}_{0g}$. Similar to the partialled Lin strategy, by doing this we only measure the variation in covariates and potential outcomes orthogonal to the stratification variables. This forces least squares to compute a conditional variance-covariance tradeoff, solving the optimal adjustment problem in Definition 3.5. In particular, the proof of Theorem 3.23 shows that if $D_{1:n} \sim \mathrm{Loc}(\psi, p)$ then the adjustment coefficient

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

Remark 3.21. Imbens and Rubin (2015) propose $\widehat{\theta}_G$ in the case of matched pairs $p = 1/2$. Their analysis uses a toy sampling model where the pairs themselves are drawn "pre-matched" from a superpopulation. By contrast, we model the experimental units as being sampled from a superpopulation, with units matched into data-dependent strata post-sampling. This more realistic model complicates the analysis, producing different limiting variances and requiring different inference procedures. In a design-based setting, Fogarty (2018) shows that the Imbens and Rubin (2015) estimator is weakly more efficient than difference of means for matched pairs designs. By contrast, we extend this estimator to a larger family of fine stratifications strictly containing matched pairs, and show that it is asymptotically optimal among linearly adjusted estimators.

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

Finally, we define tyranny-of-the-minority (ToM) adjustment, extending Lin (2013). To do so, define the adjustment coefficient

$$\widehat{\gamma}_{TM} = \mathrm{Var}_n(\check{h}_i)^{-1} \left(\mathrm{Cov}_n(\check{h}_i, Y_i | D_i = 1) \sqrt{\frac{1-p}{p}} + \mathrm{Cov}_n(\check{h}_i, Y_i | D_i = 0) \sqrt{\frac{p}{1-p}} \right). \tag{3.10}$$

Define the ToM estimator $\widehat{\theta}_{TM} = \widehat{\theta} - \widehat{\gamma}'_{TM}(\bar{h}_1 - \bar{h}_0)c_p$. The main difference between the ToM and Partialled Lin adjustment coefficients is that $\widehat{\gamma}_{TM}$ estimates the conditional variance $E[\mathrm{Var}(h|\psi)]$ only once, using the sample variance $\mathrm{Var}_n(\check{h}_i)$ for the full experimental sample. By contrast, partialled Lin estimates this term separately in each treatment arm, using $\mathrm{Var}_n(\check{h}_i | D_i = 1)$ and $\mathrm{Var}_n(\check{h}_i | D_i = 0)$. Because of this, we expect $\widehat{\theta}_{TM}$ to be more stable than $\widehat{\theta}_{PL}$ in small experiments.

Remark 3.22. Lu and Liu (2024) propose an alternate ToM regression adjustment for stratified experiments. To compare the approaches, for propensity $p = a/k$ define the within-arm partialling $\check{h}_{i1} = h_i - a^{-1} \sum_{i \in g} D_i h_i$ and $\check{h}_{i0} = h_i - (k-a)^{-1} \sum_{i \in g} (1 - D_i) h_i$. Their estimator takes the form $\widehat{\theta}_{LL} = \widehat{\theta} - \widehat{\gamma}'_{LL}(\bar{h}_1 - \bar{h}_0)$. In our notation, their adjustment coefficient $\widehat{\gamma}_{LL} = \widehat{S}_{hh}^{-1} \widehat{S}_{hY}$ has

$$\widehat{S}_{hh} = E_n \left[\frac{D_i \check{h}_{i1} \check{h}'_{i1}}{p^2} \frac{a}{a-1} + \frac{(1-D_i) \check{h}_{i0} \check{h}'_{i0}}{(1-p)^2} \frac{k-a}{k-a-1} \right]$$

and similarly for $\widehat{S}_{hY}$. Their approach is infeasible if $a = 1$ or $a = k - 1$. For example, this prohibits its use in matched pairs and matched triples experiments.

3.4.5 Main Result

The main result of this section shows that all three estimators above are asymptotically equivalent and efficient in the sense of Definition 3.5.

Theorem 3.23. *Suppose Assumptions 3.1 and 3.14 hold. If $D_{1:n} \sim \mathrm{Loc}(\psi, p)$, then $\widehat{\theta}_{PL} - \widehat{\theta}_G = o_p(n^{-1/2})$ and $\widehat{\theta}_{PL} - \widehat{\theta}_{TM} = o_p(n^{-1/2})$. We have $\sqrt{n}(\widehat{\theta}_{PL} - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V^*)$ with the optimal variance*

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

Methods for asymptotically exact inference on the ATE using these estimators are discussed in Section 4 below. Our simulations and empirical results show that the partialled Lin, Group OLS, and ToM estimators behave very similarly in finite samples.

3.5 Further Adjustment for Stratification Variables

In this section, we provide modified versions of the previous estimators that allow further adjustment for covariates $z(\psi)$ that are functions of the stratification variables. As discussed above, this cannot improve first-order efficiency but may improve finite sample performance by correcting for any remaining imbalances in ψ not controlled by the stratification.

Denote $z_i = z(\psi_i)$. For each estimator $\widehat{\theta}_k$ above with $k \in \{FE, PL, G, TM\}$, we define a modified estimator of the form $\widehat{\tau}_k = \widehat{\theta}_k - \widehat{\alpha}_k'(\bar{z}_1 - \bar{z}_0)c_p$. For the fixed effects estimator, define $\widehat{\tau}_{FE}$ to be the coefficient on D_i in the regression $Y_i \sim (1, D_i, \check{h}_i, z_i)$. For the partialled Lin estimator, define $\widehat{\tau}_{PL}$ to be the coefficient on D_i in the regression $Y_i \sim (1, \check{h}_i, z_i) + D_i(1, \check{h}_i, z_i)$. Define the modified ToM estimator to be as in Equation 3.10, with $(\check{h}_i, z_i)$ in place of $\check{h}_i$. Finally, define the modified group OLS estimator $\widehat{\tau}_G = \widehat{\theta}_G - \widehat{\alpha}_G'(\bar{z}_1 - \bar{z}_0)c_p$, with $\widehat{\alpha}_G = \widehat{\alpha}_{PL}$. Our next theorem shows that these estimators are asymptotically equivalent to the original versions of each estimator that do not adjust for $z(\psi)$. However, the simulations in Sections 5 and 6 show that they may perform better in small experiments.

Theorem 3.24. *Suppose Assumptions 3.1 and 3.14 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* $\widehat{\tau}_k = \widehat{\theta}_k + o_p(n^{-1/2})$ *for* $k \in \{FE, PL, G, TM\}$. *Each estimator has the form* $\widehat{\tau}_k = \widehat{\theta}_k - \widehat{\alpha}_k'(\bar{z}_1 - \bar{z}_0)c_p$ *with* $\widehat{\alpha}_{FE} \xrightarrow{p} \mathrm{argmin}_\alpha \mathrm{Var}(f - \alpha' z)$ *for* f *as in Theorem 3.15 and* $\widehat{\alpha}_{PL}, \widehat{\alpha}_G, \widehat{\alpha}_{TM} \xrightarrow{p} \mathrm{argmin}_\alpha \mathrm{Var}(b - \alpha' z)$.

From the second statement of the theorem, we can interpret the modified estimators as taking a conservative approach that ignores stratification on ψ and adjusts for imbalances in $z(\psi)$ as if the experiment were completely randomized.

4 Inference

In this section, we provide asymptotically exact confidence intervals for the ATE in stratified experiments using generic linearly adjusted estimators. Overcoverage is known to be a problem for inference based on the usual Eicker-Huber-White (EHW) variance estimator in stratified experiments. For example, Bai et al. (2021) shows that the EHW variance estimators for $Y \sim 1 + D + h$ and the fixed effects regression $Y \sim D + h + z^n$ are asymptotically conservative for matched pairs designs if $h = 0$. To the best of our knowledge, we give the first asymptotically exact inference methods for covariate-adjusted ($h \neq 0$) ATE estimation under general stratified designs. Our main inference result applies to any estimator of the form $\widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p + o_p(n^{-1/2})$. In particular, this enables asymptotically exact inference on the ATE using any of the estimators in this paper. Our confidence intervals are shorter than those produced by EHW in the

simulations and empirical application below, taking full advantage of the efficiency gains from both stratification and covariate adjustment.

To define our inference methods, consider such an estimator $\widehat{\theta}(\widehat{\gamma}) = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)c_p$ with $\widehat{\gamma} \xrightarrow{p} \gamma$. Define the augmented potential outcomes $Y_i^a(d) = Y_i(d) - c_p\widehat{\gamma}'h_i$ for $d \in \{0,1\}$ and the augmented outcome $Y_i^a = Y_i - c_p\widehat{\gamma}'h_i$. Then apparently

$$\widehat{\theta}(\widehat{\gamma}) = \bar{Y}_1 - \bar{Y}_0 - c_p\widehat{\gamma}'(\bar{h}_1 - \bar{h}_0) = \bar{Y}_1^a - \bar{Y}_0^a. \tag{4.1}$$

Our strategy is to apply the inference results of Cytrynbaum (2023) for difference of means estimation $\widehat{\theta} = \bar{Y}_1 - \bar{Y}_0$ to the difference of augmented potential outcomes $\bar{Y}_1^a - \bar{Y}_0^a$. To do so, let $\mathcal{G}_n$ denote the set of groups in Definition 2.1. For each $g \in \mathcal{G}_n$ define the group centroid $\bar{\psi}_g = |g|^{-1}\sum_{i\in g}\psi_i$. Let $\nu : \mathcal{G}_n \to \mathcal{G}_n$ be a bijective matching between groups satisfying $\nu(g) \neq g$, $\nu^2 = \mathrm{Id}$, and the homogeneity condition

$$\frac{1}{n}\sum_{g\in\mathcal{G}_n} |\bar{\psi}_g - \bar{\psi}_{\nu(g)}|_2^2 = o_p(1). \tag{4.2}$$

In practice, ν is obtained by simply matching the group centroids $\bar{\psi}_g$ into pairs using the Derigs (1988) matching algorithm. Let $\mathcal{G}_n^\nu = \{g \cup \nu(g) : g \in \mathcal{G}_n\}$ be the unions of paired groups formed by this matching. Define $a(g) = \sum_{i\in g} D_i$ and $k(g) = |g|$. Finally, define the variance estimator components

$$\begin{aligned}
\widehat{v}_1 &= n^{-1}\sum_{g\in\mathcal{G}_n^\nu} \frac{1}{a(g)-1}\sum_{i\neq j\in g} \frac{Y_i^a Y_j^a D_i D_j(1-p)}{p^2} \\
\widehat{v}_0 &= n^{-1}\sum_{g\in\mathcal{G}_n^\nu} \frac{1}{(k-a)(g)-1}\sum_{i\neq j\in g} \frac{Y_i^a Y_j^a (1-D_i)(1-D_j)p}{(1-p)^2} \\
\widehat{v}_{10} &= n^{-1}\sum_{g\in\mathcal{G}_n} \frac{k}{a(k-a)}(g)\sum_{i,j\in g} Y_i^a Y_j^a D_i(1-D_j).
\end{aligned}$$

Next, define the variance estimator

$$\widehat{V} = \mathrm{Var}_n\left(\frac{(D_i-p)Y_i^a}{p-p^2}\right) - \widehat{v}_1 - \widehat{v}_0 - 2\widehat{v}_{10}. \tag{4.3}$$

Our inference strategy begins with the sample variance of adjusted estimator, which is consistent for the asymptotic variance of $\widehat{\theta}_{adj}$ under an iid design, but too large under stratified designs. We correct this sample variance using the estimators above, which measure how well the stratification variables predict augmented outcomes in local regions of the covariate space. This section's main result shows that $\widehat{V}$ is consistent for the limiting variance of Theorem 3.4, enabling asymptotically exact inference on the ATE using adjusted estimators.

Theorem 4.1 (Inference)**.** *Under the conditions of Theorem 3.4, if* $D_{1:n} \sim \text{Loc}(\psi, p)$*, then* $\widehat{V} = V + o_p(1)$*.*

By Theorem 4.1 and our previous asymptotic results in Theorem 3.4, the confidence interval $\widehat{C} = [\widehat{\theta}(\widehat{\gamma}) \pm \widehat{V}^{1/2} c_{1-\alpha/2}/\sqrt{n}]$ with $c_\alpha = \Phi^{-1}(\alpha)$ is asymptotically exact in the sense that $P(\text{ATE} \in \widehat{C}) = 1 - \alpha + o(1)$.

5 Simulations

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

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

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 $h_i = \psi_i' Q_h \psi_i + \psi_i' L_h + u_i$ with $E[u_i | \psi_i] = 0$. 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 $Q_h = (1/100)A$.

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 3.2. **Naive** refers to the non-interacted regression $Y \sim (1, D, h)$, (Theorem A.4). **FE** refers to the fixed effects estimator (Theorem 3.15) and **Plin** the partialled Lin estimator (Theorem 3.23). **GO** refers to Group OLS and **ToM** refers to Tyranny-of-the-Minority estimation (Theorem 3.23). **Strata Controls** refer to modified versions of each of the previous estimators that further adjust for parametric strata controls $z(\psi)$, as discussed in Section 3.5. In our simulations, we set $z(\psi) = \psi$. **Ad** refers to

an adaptive[5] estimator that sets $\widehat{\theta}_{adj} = \widehat{\theta}_L$ if $\widehat{V}(\widehat{\gamma}_L) \leq \widehat{V}(\widehat{\gamma}_{PL})$ and $\widehat{\theta}_{adj} = \widehat{\theta}_{PL}$ otherwise,[6] including parametric controls $z(\psi) = \psi$ in both cases.

		No Strata Controls							Strata Controls $z(\psi)$						
$(n, \dim(\psi))$	Model	Unadj	Naive	Lin	FE	Plin	GO	ToM	Naive	Lin	FE	Plin	GO	ToM	Ad
(600, 2)	1	100	113	102	49	48	49	48	36	35	35	37	37	36	34
	2	100	126	102	64	57	58	57	60	46	52	47	47	47	45
	3	100	116	116	38	38	38	38	48	48	36	36	37	36	37
	4	100	27	31	31	27	27	27	26	26	38	32	33	32	26
	5	100	28	28	18	18	18	18	21	21	19	19	19	19	19
	6	100	100	100	11	11	11	11	7	7	9	9	9	9	7
(1200, 2)	1	100	114	103	44	44	44	44	35	34	31	33	33	33	32
	2	100	126	102	60	56	56	56	61	47	50	47	46	47	45
	3	100	116	116	38	38	38	38	48	48	37	37	37	37	37
	4	100	26	30	29	25	25	25	23	24	36	30	30	30	24
	5	100	28	28	17	17	17	17	20	20	17	18	17	18	18
	6	100	101	101	9	9	9	9	7	7	8	8	8	8	7
(1200, 5)	1	100	142	127	85	84	84	84	25	24	41	46	55	46	24
	2	100	145	123	94	86	87	86	45	34	57	54	62	54	34
	3	100	137	137	81	81	81	81	40	40	54	54	57	54	40
	4	100	27	31	31	27	27	27	25	20	54	45	49	45	20
	5	100	32	32	24	24	24	24	18	18	38	38	39	38	18
	6	100	138	138	67	67	67	67	15	15	36	36	39	37	15
	R_k	73	65	60	17	15	15	15	5	2	10	8	10	8	0.2

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

Table 1 studies finite sample efficiency. We present the mean squared error (MSE) ratio, relative to unadjusted estimation, for each of the adjusted estimators above. The bottom line of the table reports the *excess risk* R_k of each estimator k relative to the optimal estimator. To define this, let $\text{MSE}_{k,s}$ be the relative MSE of estimator k in simulation s. Then we set $R_k = (1/S)\sum_s(\text{MSE}_{k,s} - \min_j \text{MSE}_{j,s})$, averaging over all simulations in the table. All results are calculated using 2000 Monte Carlo repetitions.

In models 1, 2, and 3, both Naive and Lin style linear adjustment are strictly inefficient relative to unadjusted estimation. These models have marginal covariance $\text{Cov}(Y(d), h) > 0$ but conditional covariance $E[\text{Cov}(Y(d), h|\psi)] < 0$, conditional on the stratification variables. Because of this, the optimal adjustment coefficient $\gamma^* < 0$, while the Naive and Lin regressions estimate positive adjustment coefficients $\gamma_N, \gamma_L > 0$, leading to even worse performance than unadjusted estimation in some cases. For Models 4 and 5, the **Naive** and **Lin** methods are competitive with the generic efficient methods from Section 3.4. This is because in these cases we made it so that $\text{Cov}(Y(d), h) \approx E[\text{Cov}(Y(d), h|\psi)]$, so that "by

[5]This estimator is pointwise asymptotically equivalent to $\widehat{\theta}_{PL}$. Issues with post model-selection inference (e.g. Leeb and Potscher (2005)) appear to be less worrying here, since even under the fixed alternative $\gamma^* \neq \gamma_L$, the Lin estimator is still $\sqrt{n}$-consistent and asymptotically unbiased.

[6]We could also use a cross-fit version of $\widehat{V}(\gamma)$ to reduce bias. However, the in-sample criterion performed quite well in our simulations.

chance" γ^* is close to γ_N and γ_L. However, the parametric coefficients γ_N and γ_L are estimated more precisely than the semiparametric object $\gamma^* = E[\text{Var}(h|\psi)]^{-1}E[\text{Cov}(h, b|\psi)]$. For Model 6, **Lin** with $z(\psi) = \psi$ controls is (approximately) optimal by Theorem 3.9, since $E[w|\psi]$ is (approximately) linear in ψ.

Summarizing our findings, the **Lin**, **Plin**, and **Naive** estimators with parametric **Strata Controls** $z(\psi) = \psi$ had low excess risk across specifications, while the **Ad** estimator was the most efficient overall. The **Naive** and **Lin** estimators without strata controls or within-stratum partialling had large MSE. The **Plin**, **GO**, and **ToM** estimators had similar MSE across model specifications. These generic methods performed the best in regimes with large n, small $\dim(\psi)$, and nonlinear $E[h|\psi]$. In these cases, the gap $\gamma_L - \gamma^*$ between the sub-optimal Lin coefficient and optimal coefficient γ^* dominates the additional variability $\text{Var}(\widehat{\gamma}^*) > \text{Var}(\widehat{\gamma}_L)$ required to estimate γ^* (this variability increases with $\dim(\psi)$). For example, **Plin** with $z(\psi)$ controls performs the best when $(n, \dim(\psi)) = (1200, 2)$, but **Lin** with $z(\psi)$ controls is much better when $\dim(\psi) = 5$. The **Ad** estimator used a variance pre-test to choose between **Plin** and **Lin** (including $z(\psi)$ controls), allowing it to perform well in both regimes.

		No Strata Controls							Strata Controls $z(\psi)$						
	Model	Unadj	Naive	Lin	FE	Plin	GO	ToM	Naive	Lin	FE	Plin	GO	ToM	Ad
%ΔCI Length vs. Unadj	1	0	17	11	-5	-5	-5	-5	-49	-50	-34	-29	-26	-29	-50
	2	0	18	10	-3	-4	-4	-4	-33	-41	-25	-25	-22	-25	-41
	3	0	16	16	-6	-6	-6	-6	-36	-36	-24	-24	-24	-24	-36
	4	0	-46	-43	-42	-46	-46	-46	-50	-55	-22	-31	-26	-30	-55
	5	0	-44	-44	-49	-49	-49	-49	-56	-56	-34	-34	-31	-34	-56
	6	0	16	16	-12	-12	-12	-12	-59	-59	-35	-35	-35	-35	-59
Coverage (Exact)	1	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.95	0.95	0.96	0.96	0.95	0.96	0.95
	2	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.96
	3	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
	4	0.95	0.95	0.94	0.96	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.95
	5	0.94	0.95	0.95	0.96	0.96	0.96	0.96	0.95	0.95	0.96	0.96	0.97	0.96	0.95
	6	0.95	0.95	0.95	0.97	0.97	0.97	0.97	0.96	0.96	0.97	0.97	0.96	0.97	0.96
Coverage (EHW)	1	0.99	0.95	0.96	0.97	0.99			0.99	0.98	0.99	0.97			
	2	0.99	0.95	0.95	0.94	0.99			0.98	0.93	0.98	0.96			
	3	1.00	0.96	0.95	0.96	1.00			0.99	0.93	0.98	0.97			
	4	0.99	0.99	0.90	0.98	1.00			0.97	0.68	0.98	0.97			
	5	0.99	0.97	0.90	0.98	1.00			0.96	0.65	0.99	0.99			
	6	1.00	0.97	0.96	0.96	1.00			0.99	0.97	1.00	0.99			

Table 2: Properties of Inference

Table 2 reports finite sample efficiency and coverage properties of the asymptotically exact inference methods developed in Section 4. We let $n = 1200$ and $\dim(\psi) = 5$. The first panel shows % change in confidence interval length relative to unadjusted estimation. All confidence intervals are computed using the method in Theorem 4.1. 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 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 Application

In this section we apply our methods to the experiment in Baysan (2022a),[7] who estimates the effect of a political information campaign on support for a 2017 Turkish referendum removing checks and balances on executive power. The campaign was administered by the opposition Republican People's Party (CHP), who opposed the referendum. Randomization was performed at the neighborhood level, stratified on quartiles of CHP vote share in the previous 2015 elections. The main outcome is the "No" vote share in the 2017 referendum.[8] Due to the cost of administering the campaign, $p = 2/11$ out of $n = 550$ total neighborhoods were treated. In the original analysis, Baysan (2022a) performed non-interacted covariate adjustment (Theorem A.4) for $h(X)$ = number of registered voters, number of valid votes, number of votes for the CHP in 2015, CHP vote share in 2015, voter turnout, and CHP vote share quartile fixed effects.

	No Strata Controls							Strata Controls $z(\psi)$					
Model	Unadj	Naive	Lin	FE	Plin	GO	ToM	Naive	Lin	FE	Plin	GO	ToM
$\widehat{\theta}_{adj}$	-0.0054	0.0040	0.0047	0.0041	0.0021	0.0034	0.0021	0.0041	0.0040	0.0038	0.0037	0.0031	0.0019
SE	0.0088	0.0074	0.0074	0.0074	0.0078	0.0077	0.0081	0.0074	0.0073	0.0077	0.0076	0.0078	0.0083
HC2	0.0155	0.0075	0.0073	0.0075	0.0149			0.0075	0.0070	0.0736	0.0071		

Table 3: Empirical Results

In the first block of Table 3, we replicate the neighborhood-level analysis of Baysan (2022a). $\widehat{\theta}_{adj}$ is the point estimate from each adjustment strategy, SE is the asymptotically exact standard error from Section 4, and EHW is the usual robust standard error (HC2). Estimates in the "strata controls $z(\psi)$" section include quartile fixed effects, while the leftmost section does not. The results in Section 3.3 show that Lin adjustment with quartile fixed effects is efficient in this case, and indeed this has the smallest estimated standard error. The generic efficient estimators have slightly larger SE. The asymptotically exact standard errors from Section 4 are generally similar to or smaller than EHW, except for the Lin, FE, and Plin estimators with $z(\psi)$ controls. However, our simulation also showed that EHW standard errors may severely undercover in these cases.[9]

[7] The data is available from Baysan (2022b).

[8] Baysan (2022a) estimates effects of the campaign on vote share at both the ballot box and neighborhood level. We focus on the neighborhood level effects.

[9] We also note that Bai et al. (2024c) have found the EHW standard error from a linear regression with block fixed effects to be potentially invalid in a related problem.

		No Strata Controls							Strata Controls $z(\psi)$					
	Model	Unadj	Naive	Lin	FE	Plin	GO	ToM	Naive	Lin	FE	Plin	GO	ToM
Coarse	Est	0.0000	0.0001	0.0003	0.0001	0.0002	0.0003	0.0000	0.0001	0.0004	0.0001	0.0003	0.0002	0.0000
	SE	0.0085	0.0076	0.0075	0.0075	0.0077	0.0077	0.0078	0.0076	0.0074	0.0078	0.0076	0.0079	0.0079
	HC2	0.0144	0.0078	0.0078	0.0077	0.0141			0.0078	0.0077	0.0736	0.0080		
Fine	Est	-0.0001	0.0000	0.0000	0.0000	0.0000	0.0004	-0.0001	0.0000	0.0002	0.0000	0.0002	0.0004	-0.0001
	SE	0.0077	0.0081	0.0080	0.0076	0.0077	0.0077	0.0077	0.0075	0.0075	0.0075	0.0075	0.0077	0.0077
	HC2	0.0144	0.0141	0.0142	0.0078	0.0145			0.0078	0.0078	0.0733	0.0078		
Fine $p = 1/2$	Est	-0.0001	-0.0001	-0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	-0.0001	0.0001	0.0001	0.0001	0.0001
	SE	0.0072	0.0073	0.0073	0.0070	0.0070	0.0070	0.0070	0.0066	0.0066	0.0066	0.0066	0.0066	0.0066
	HC2	0.0113	0.0111	0.0111	0.0059	0.0114			0.0060	0.0060	0.0565	0.0061		
Fine $\dim(\psi) = 3$	Est	0.0000	0.0001	0.0002	0.0001	0.0002	0.0000	0.0003	0.0001	0.0002	0.0000	0.0002	0.0001	0.0001
	SE	0.0155	0.0155	0.0155	0.0155	0.0155	0.0155	0.0155	0.0146	0.0146	0.0146	0.0146	0.0147	0.0147
	HC2	0.0145	0.0145	0.0145	0.0089	0.0145			0.0079	0.0078	0.0740	0.0078		

Table 4: Simulated Designs

Overall, changing the adjustment method did not have an economically meaningful effect on the conclusions of the study, and we recover the null effect of Baysan (2022a) in all cases. The covariate $h_k =$ "CHP vote share in 2015" is highly predictive of $Y =$ "CHP vote share in 2017," so adjusting for this variable ex-post provides a modest variance reduction even after stratifying on 2015 vote share quartiles. However, the estimated optimal coefficient $\gamma_k^* \approx 0.27$ and Lin coefficient $\gamma_{L,k} \approx 0.31$ are quite similar, so (inefficient) Lin adjustment still performs quite well. The other covariates such as $h_j =$ "voter turnout" are very weak predictors of outcomes, so changing the adjustment coefficient on these variables doesn't matter much.

Next, we ask how each estimator would have performed in the experiment in Baysan (2022a) under counterfactual randomization procedures, such as fine stratification.[10] To do so, we follow the nonparametric imputation strategy in Bai (2022), defining potential outcomes $\widehat{Y}_i(d) = Y_i$ if $D_i = d$ and matching imputation $\widehat{Y}_i(d) = Y_{j(i)}(d)$ with $j(i) = \text{argmin}_{j:D_j=d} |X_i - X_j|_2$ if $D_i \neq d$. We let the matching variables X_i include all controls used in the analysis of Baysan (2022a). Given the imputed data $(X_i, \widehat{Y}_i(0), \widehat{Y}_i(1))_{i=1}^n$, we do the following simulation exercise: (1) draw treatment assignments $D_{1:n} \sim \text{Loc}(\psi, p)$, (2) reveal outcomes $\widehat{Y}_i = \widehat{Y}_i(D_i)$ and (3) form each estimator $\widehat{\theta}_{adj}$. We report average point estimates and standard errors over $N = 2000$ Monte Carlo repetitions of this procedure.

The first block of Table 4 uses this imputation procedure to reproduce the empirical results in Table 3, stratifying by quartiles of CHP vote share and adjusting for exactly the same covariates. The standard errors are very similar to those in the empirical analysis, which provides some validation for this imputation exercise. In the second block of Table 4, we simulate a design with fine stratification on 2015 CHP vote share, rather than just stratifying by quartiles of the vote share as in Baysan (2022a). We used a matched 11-tuples design, letting $D_{1:n} \sim \text{Loc}(\psi, p)$ for $p = 2/11$ and $\psi =$ (2015 CHP vote share). Covariates $h(X)$ are as above, with $z(\psi) = \psi$. In the third block, we simulate a matched

[10] Algorithms and inference methods for fine stratification with $p \neq 1/2$ have only been developed recently, e.g. Bai (2022) and Cytrynbaum (2023).

pairs design $D_{1:n} \sim \text{Loc}(\psi, 1/2)$. Note that $p = 1/2$ was infeasible in the original experiment due to the high cost of treatment. The last block uses the design $D_{1:n} \sim \text{Loc}(\psi_{alt}, p)$ for $\psi_{alt} = (\text{CHP vote share}, \text{Num. of registered voters}, \text{Num. of valid votes})$, $p = 2/11$, and covariates $h = \text{Turnout}$.

We make some brief observations about this simulation exercise. First, note that the Naive and Lin adjustment are strictly less efficient than unadjusted estimation under simulated fine stratification, consistent with Theorems 3.2 and Section A.4. Lin and partialled Lin with $z(\psi)$ controls are the most efficient. Adjustment for extra covariates h doesn't significantly improve efficiency relative to the baseline efficiency gain from finely stratifying on ψ and adjusting for $z(\psi)$ ex-post. Using a matched pairs design $p = 1/2$ improves efficiency, though the improvement is small considering that this design would require providing the information campaign to 175 extra neighborhoods. Finally, fine stratification on ψ_{alt} significantly reduces efficiency. This is because the extra covariates are not very predictive of outcomes, but stratifying on these covariates force us to use worse matches on the important covariate $\psi = (\text{2015 CHP vote share})$.

7 Discussion and Recommendations for Practice

Stratified randomization and covariate adjustment are both commonly used in the design and analysis of experiments. In general, experimenters should stratify on a few variables $\psi(X)$ expected to be most predictive of outcomes at design-time, and plan to adjust for imbalances in the remaining covariates $h(X)$ ex-post, as discussed in Section 3.2. Our analysis showed that under stratified randomization, the usual regression adjusted estimators can be inefficient. Motivated by this, we provide feasible alternatives that are asymptotically optimal in the class of linearly adjusted estimators. We conclude by giving some recommendations for empirical practice based on the theory, simulations, and empirical results above.

We recommend that applied researchers use either (1) the Lin estimator with parametric strata controls $z(\psi)$ (e.g. $z(\psi) = \psi$) or (2) the partialled Lin estimator with parametric controls $z(\psi)$, since these estimators performed the best across our simulations and empirical application. Lin with parametric controls $z(\psi)$ is efficient under a rich covariates condition (Section 3.3), while partialled Lin is generically efficient (Section 3.5). Both estimators are robust to treatment effect heterogeneity, while the strata fixed effects estimator (Theorem 3.15) is not unless $p = 1/2$.

In our simulations, partialled Lin had good finite sample performance in regimes where n was large relative to $\dim(\psi)$, especially when $E[h|\psi]$ was very nonlinear. Lin with $z(\psi) = \psi$ controls performed better when $\dim(\psi)$ was large relative to n, or if $E[h|\psi]$ was approximately linear. To decide which regime we are in, we suggest model

selection using a variance pre-test, choosing Lin if $\widehat{V}(\widehat{\gamma}_L) \leq \widehat{V}(\widehat{\gamma}_{PL})$ and partialled Lin otherwise. This adaptive estimator (**Ad** in Section 6) was efficient in both regimes and had good coverage properties. We leave a more general study of such post model-selection estimators in this context to future work.

Regardless of the adjustment strategy, we recommend using the asymptotically exact confidence intervals provided 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 often had significant overcoverage.

References

Alberto Abadie and Guido W. Imbens. Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, pages 175–187, 2008.

Jason Ansel, Han Hong, and Jessie Li. Ols and 2sls in randomized and conditionally randomized experiments. *Jahrbücher für National ökonomie und Statistik*, 2018.

Tim Armstrong. Asymptotic efficiency bounds for a class of experimental designs. *arXiv preprint arXiv:2205.02726*, 2022.

Yuehao Bai. Optimality of matched-pair designs in randomized controlled trials. *American Economic Review*, 2022.

Yuehao Bai, Joseph P. Romano, and Azeem M. Shaikh. Inference in experiments with matched pairs. *Journal of the American Statistical Association*, 2021.

Yuehao Bai, Hongchang Guo, Azeem M. Shaikh, and Max Tabord-Meehan. Inference in experiments with matched pairs and imperfect compliance. *arXiv preprint arXiv:2307.13094*, 2024a.

Yuehao Bai, Liang Jiang, Joseph P. Romano, Azeem M. Shaikh, and Yichong Zhang. Covariate adjustment in experiments with matched pairs. *Journal of Econometrics*, 2024b.

Yuehao Bai, Max Tabord-Meehan, and Jizhou Liu. Inference for matched tuples and fully blocked factorial designs. *Quantitative Economics*, 2024c.

Ceren Baysan. Persistent polarizing effects of persuasion: Experimental evidence from turkey. *American Economic Review*, November 2022a.

Ceren Baysan. Data and code for: Persistent polarizing effects of persuasion: Experimental evidence from turkey. Nashville, TN: American Economic Association, 2022. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2022-10-19. https://doi.org/10.3886/E172061V1, December 2022b.

Federico A. Bugni, Ivan A. Canay, and Azeem M. Shaikh. Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 2018.

Haoge Chang. Design-based estimation theory for complex experiments. *arXiv preprint arXiv:2311.06891*, 2023.

Max Cytrynbaum. Designing representative and balanced experiments by local randomization. *arXiv preprint arXiv:2111.08157v1*, 2022.

Max Cytrynbaum. Optimal stratification of survey experiments. *arXiv preprint arXiv:2111.08157*, 2023.

Ulrich Derigs. Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, 13:225–261, 1988.

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

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

Guido Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62, 1994.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

Liang Jiang, Oliver B. Linton, Haihan Tang, and Yichong Zhang. Improving estimation efficiency via regression-adjustment in covariate-adaptive randomizations with imperfect compliance. *Review of Economics and Statistics*, 2024.

Hannes Leeb and Benedikt M. Potscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 2005.

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

Hanzhong Liu and Yuehan Yang. Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 2020.

Xin Lu and Hanzhong Liu. Tyranny-of-the-minority regression adjustment in randomized experiments. *Journal of the American Statistical Association*, June 2024.

Wei Ma, Fuyi Tu, and Hanzhong Liu. Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective. *Statistics in Medicine*, 2022.

Akanksha Negi and Jeffrey M. Wooldridge. Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 2021.

Katarzyna Reluga, Ting Ye, and Qingyuan Zhao. A unified analysis of regression adjustment in randomized experiments. *Electronic Journal of Statistics*, 2024.

Jiyang Ren. Model-assisted complier average treatment effect estimates in randomized experiments with non-compliance and a binary outcome. *Journal of Business and Economic Statistics*, 2023.

James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129, 1995.

Peter M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4), 1988.

Xinhe Wang, Tingyu Wang, and Hanzhong Liu. Rerandomization in stratified randomized experiments. *Journal of the American Statistical Association*, 2021. Working Paper.

Ting Ye, Jun Shao, Yanyao Yi, and Qingyuan Zhao. Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*, 00(0), 2022.

Ke Zhu, Hanzhong Liu, and Yuehan Yang. Design-based theory for lasso adjustment in randomized block experiments with a general blocking scheme. *arXiv preprint arXiv:2109.11271*, 2024.

Supplement to "Covariate Adjustment in Stratified Experiments"

Max Cytrynbaum

A Appendix

A.1 Experiments with Noncompliance

In this section, we extend our main results to the case of experiments with imperfect compliance. The theorems in this section are simple corollaries of our main results. For completeness, full proofs are provided in Section A.9.

Previously, Ansel et al. (2018) studied covariate adjustment in experiments with noncompliance and iid or coarsely stratified treatment assignment. Bai et al. (2024a) study matched pairs experiments with noncompliance. See also Jiang et al. (2024) and Ren (2023) for nonlinear adjustment in coarsely stratified experiments and completely randomized experiments with noncompliance, respectively.

Let $z \in \{0,1\}$ denote a binary instrument. Let $D(z)$ be the potential treatments and $Y(d,z) = Y(d)$ the potential outcomes, satisfying exclusion. Define the intention-to-treat (ITT) potential outcomes $W_i(z) = Y_i(D_i(z))$, so that $Y_i = Z_i W_i(1) + (1-Z_i)W_i(0)$ and $D_i = Z_i D_i(1) + (1-Z_i)D_i(0)$. Impose monotonicity $D(1) \geq D(0)$ and positive compliance $\tau_D = P(D(1) > D(0)) > 0$. Define the ITT effect $\tau_W = E[W(1) - W(0)]$. Under these assumptions, the parameter $\tau_L \equiv \tau_W/\tau_D = E[Y(1) - Y(0)|D(1) > D(0)]$ is the local average treatment effect (LATE) (Imbens and Angrist (1994)). To estimate τ_L, we consider adjusted Wald estimators of the form

$$\widehat{\tau}_{adj} = \frac{\bar{W}_1 - \bar{W}_0 - \widehat{\gamma}_W'(\bar{h}_1 - \bar{h}_0)c_p}{\bar{D}_1 - \bar{D}_0 - \widehat{\gamma}_D'(\bar{h}_1 - \bar{h}_0)c_p} \tag{A.1}$$

To analyze $\widehat{\tau}_{adj}$, we require that Assumption 3.1 holds for both potential outcomes $W(z)$ and $D(z)$ and covariates $h(X)$, and also impose Assumption 3.14. Suppose the adjustment coefficients $(\widehat{\gamma}_W, \widehat{\gamma}_D) = (\gamma_W, \gamma_D) + o_p(1)$. Our first result is a consequence of Theorem 3.4. To state the result, we define the modified potential outcomes $Q(z) = W(z) - \tau_L D(z)$ for $z \in \{0,1\}$ and modified adjustment coefficient $\gamma_Q = \gamma_W - \tau_L \gamma_D$.

Theorem A.1. *If* $Z_{1:n} \sim \mathrm{Loc}(\psi, p)$ *then* $\sqrt{n}(\widehat{\tau}_{adj} - \tau_L) \Rightarrow \mathcal{N}(0, V(\gamma_Q)/\tau_D^2)$ *with*

$$V(\gamma_Q) = \mathrm{Var}(c_Q) + E\big[\mathrm{Var}(b_Q - \gamma_Q' h|\psi)\big] + E\left[\frac{\sigma_{1Q}^2(X)}{p} + \frac{\sigma_{0Q}^2(X)}{1-p}\right].$$

The terms $c_Q(X) = E[Q(1) - Q(0)|X]$, similarly for b_Q and σ_{zQ}^2, substituting the

potential outcomes $Q(z)$ for $Y(d)$ in each formula.

Optimal Adjustment. Let $\widehat{\gamma}_Q = \widehat{\gamma}_W - \tau_L \widehat{\gamma}_D$ and define the adjustment scheme $\widehat{\tau}_{adj}$ to be efficient if $\widehat{\gamma}_Q \xrightarrow{p} \gamma_Q^* \in \operatorname{argmin}_\gamma V(\gamma)$. We construct efficient adjusted Wald estimators using the generic efficient estimators of Section 3.4. Let $\widehat{\theta}_k^W$ and $\widehat{\theta}_k^D$ for $k \in \{PL, GO, TM\}$ be any of the generic efficient estimators of Section, plugging in outcomes W or D in place of Y. For example, $\widehat{\theta}_{PL}^W$ is the coefficient on Z_i in the regression $W_i \sim (1, \check{h}_i) + Z_i(1, \check{h}_i)$ and $\widehat{\theta}_{PL}^D$ the coefficient on Z_i in $D_i \sim (1, \check{h}_i) + Z_i(1, \check{h}_i)$. Define the LATE estimators $\widehat{\tau}_L^k = \widehat{\theta}_k^W / \widehat{\theta}_k^D$ for $k \in \{PL, GO, TM\}$. Our next theorem is a consequence of the efficiency results in Section 3.4.

Theorem A.2. *Suppose $Z_{1:n} \sim \text{Loc}(\psi, p)$. For each $k \in \{PL, GO, TM\}$, the estimator $\widehat{\tau}_L^k$ is efficient with $\sqrt{n}(\widehat{\tau}_L^k - \tau_L) \Rightarrow \mathcal{N}(0, V^*)$ for $V^* = \min_\gamma V(\gamma)$.*

Finally, we provide asymptotically exact inference on τ_L using the adjusted estimators $\widehat{\tau}_L^k$ above. Define the augmented outcomes $Q_i^a = W_i - \widehat{\tau}_L^k D_i - h_i'(\widehat{\gamma}_W - \widehat{\tau}_L^k \widehat{\gamma}_D)$. Let $\widehat{v}_1^q$, $\widehat{v}_0^q$, and $\widehat{v}_{10}^q$ be the variance estimators in Equation 4.3, plugging in Q_i^a in place of Y_i^a. Define the variance estimator

$$\widehat{V} = \frac{1}{(\widehat{\theta}_k^D)^2} \left[\text{Var}_n \left(\frac{(D_i - p) Q_i^a}{p - p^2} \right) - \widehat{v}_1^q - \widehat{v}_0^q - 2\widehat{v}_{10}^q \right] \tag{A.2}$$

Theorem A.3. *Suppose $Z_{1:n} \sim \text{Loc}(\psi, p)$. Then $\widehat{V} = V^* + o_p(1)$.*

Theorems A.1 and A.3 show that the confidence interval $\widehat{C} = [\widehat{\tau}_L^k \pm \widehat{V}^{1/2} c_{1-\alpha/2} / \sqrt{n}]$ with $c_\alpha = \Phi^{-1}(\alpha)$ is asymptotically exact in the sense that $P(\tau_L \in \widehat{C}) = 1 - \alpha + o(1)$.

A.2 Varying Propensities

In this section, we extend our results to fine stratification with varying propensities $p(\psi)$. To that end, let $p(\psi) \in \{a_l / k_l : l \in L\}$ with $|L| < \infty$ a finite index set. Cytrynbaum (2023) extends Definition 2.1 to non-constant $p(\psi)$ by the following double stratification procedure:

(1) Partition the units $\{1, \ldots, n\}$ into propensity strata $S_l \equiv \{i : p(X_i) = a_l / k_l\}$.

(2) In each propensity stratum S_l, draw samples $(D_i)_{i \in S_l} \sim \text{Loc}(\psi, a_l / k_l)$.

To implement this, we run the algorithm of Cytrynbaum (2023) to match units into groups of k_l separately in each propensity stratum S_l, drawing treatment assignments $(D_i)_{i \in g} \sim \text{CR}(a_l / k_l)$ independently for each $g \in \mathcal{G}_l$. Define $\widehat{\theta}_{adj}(\gamma)$ to be the AIPW estimator of Section 3.2, with linear models $f_d(X_i) = \gamma_d' h(X_i)$ for $d \in \{0, 1\}$, so that

$$\widehat{\theta}_{adj}(\gamma) = (\gamma_1 - \gamma_0)' E_n[h_i] + E_n \left[\frac{D_i (Y_i - \gamma_1' h_i)}{p(\psi_i)} \right] - E_n \left[\frac{(1 - D_i)(Y_i - \gamma_0' h_i)}{1 - p(\psi_i)} \right].$$

Define $\gamma = (\gamma_0, \gamma_1)$ and weighted covariates $h_i^p = \left(h_i\sqrt{\frac{p_i}{1-p_i}}, h_i\sqrt{\frac{1-p_i}{p_i}}\right)$. Under assumption 3.1, Theorem 3.4 may be extended to show that if $\widehat{\gamma} \xrightarrow{p} \gamma$ and $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ then $\sqrt{n}(\widehat{\theta}_{adj}(\widehat{\gamma}) - \text{ATE}) \Rightarrow \mathcal{N}(0, V(\gamma))$ with variance

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

The optimal adjustment coefficient is $\gamma^* = E[\text{Var}(h_i^p|\psi_i)]^{-1}E[\text{Cov}(h_i^p, b_i|\psi_i)]$ if the condition $E[\text{Var}(h_i^p|\psi_i)] \succ 0$ is satisfied. Let k_i denote the size of the group that unit i belongs to. Extending the work in Section 3.4, the estimator

$$\widehat{\gamma} = E_n\left[\check{h}_i^p(\check{h}_i^p)'\frac{k_i}{k_i - 1}\right]^{-1} E_n\left[\check{h}_i^p Y_i^{TM}\frac{k_i}{k_i - 1}\right]$$

with weighted outcomes $Y_i^{TM} = D_i Y_i (1-p_i)^{1/2} p_i^{-3/2} + (1 - D_i) Y_i p_i^{1/2}(1-p_i)^{-3/2}$ has $\widehat{\gamma} = \gamma^* + o_p(1)$. Then the estimator $\widehat{\theta}_{adj}(\widehat{\gamma})$ is efficient in the sense of achieving the minimal variance $\min_\gamma V(\gamma)$.

A.3 Non-Interacted Regression Adjustment

For completeness, before continuing we describe the asymptotic behavior of the commonly used non-interacted regression estimator under stratified designs. Let $\widehat{\theta}_N$ be the coefficient on D_i in $Y \sim 1 + D + h$.

Theorem A.4. *Suppose Assumptions 3.1 and 3.14 hold. The estimator has representation* $\widehat{\theta}_N = \widehat{\theta} - \widehat{\gamma}_N'(\bar{h}_1 - \bar{h}_0) + O_p(n^{-1})$. *If* $D_{1:n} \sim \text{Loc}(\psi, p)$ *then* $\sqrt{n}(\widehat{\theta}_N - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ *with variance*

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

The coefficient $\gamma_N = \text{argmin}_{\gamma \in \mathbb{R}^{d_h}} \text{Var}(f - \gamma' h)$ *for target function*

$$f(x) = m_1(x)\sqrt{\frac{p}{1-p}} + m_0(x)\sqrt{\frac{1-p}{p}}$$

with $f(x) \neq b(x)$ *in general. The non-interacted estimator is efficient if* $\psi = 1$ *and either* $p = 1/2$ *or* $\text{Cov}(h, Y(1) - Y(0)) = 0$.

Theorem A.4 shows that $\widehat{\theta}_N$ is generally inefficient since it uses the wrong objective function. In particular, the target function $f(x) \neq b(x)$ unless $p = 1/2$. Also, the limiting coefficient γ_N minimizes marginal instead of conditional variance. The results in Section 4 show how to construct asymptotically exact confidence intervals for the ATE using $\widehat{\theta}_N$.

A.4 Nonlinear Adjustment

Alternately, we may consider general nonlinear covariate adjustment strategies. Let $\widehat{h}(x)$ be a function estimated in some class $\mathcal{H}$ and consider the adjusted estimator

$$\widehat{\theta}_{adj}(\widehat{h}) = E_n\left[\frac{(Y_i - \widehat{h}(X_i))(D_i - p_i)}{p_i - p_i^2}\right].$$

For example, the usual AIPW estimator in Section 3.2 can be shown to take this form. Linear adjustment corresponds to the parametric family $\mathcal{H} = \{h(x)'\gamma : \gamma \in \mathbb{R}^{d_h}\}$. Similar to Bai et al. (2024b), suppose that for some function $h(X) \in L_2$ the equicontinuity condition holds

$$\sqrt{n}E_n\left[\frac{(\widehat{h} - h)(X_i)(D_i - p_i)}{p_i - p_i^2}\right] = o_p(1).$$

Theorem 3.4 can be extended to show that if $D_{1:n} \sim \text{Loc}(\psi, p(\psi))$ then $\sqrt{n}(\widehat{\theta}_{adj}(\widehat{h}) - \text{ATE}) \Rightarrow \mathcal{N}(0, V(h))$ with asymptotic variance

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

for $c_p(\psi) = \sqrt{p(\psi) - p(\psi)^2}$. One natural extension of the current work would be to solve a general version of the optimal adjustment problem over a nonlinear or general nonparametric function class $\mathcal{H}$.

$$\min_{h \in \mathcal{H}} E\left[\text{Var}\left(b - h/c_p(\psi)\middle|\psi\right)\right] \tag{A.3}$$

This requires new technical tools, the development of which we leave to future work.

A.5 Proofs for Section 3.1

Proof of Theorem 3.4. First, note that since $E[|h|_2^2] < \infty$ we may apply Lemma A.2 of Cytrynbaum (2023) to show that

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

Define auxiliary potential outcomes $Z(d) = Y(d) - c_p\gamma'h(X)$ for $d \in \{0, 1\}$ with $Z_i = Z(D_i)$. Summarizing, we have shown that $\widehat{\theta}_{adj} = \bar{Z}_1 - \bar{Z}_0 + o_p(n^{-1/2})$. Observe that $E[Z(d)^2] \lesssim E[Y(d)^2] + c_p^2|\gamma|_2^2E[|h(X)|_2^2] < \infty$. Then we may apply the general version of Theorem 3.11 in Cytrynbaum (2023) (Equation 3.6). Setting $q = 1$ and

$\psi_1 = \psi_2$ and applying the theorem to the auxiliary potential outcomes $Z(d)$, we have $\sqrt{n}(\widehat{\theta}_{adj} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$

$$V = \text{Var}(c_Z(X)) + E[\text{Var}(b_Z(X;p)|\psi)] + E\left[\frac{\sigma^2_{1,Z}(X)}{p} + \frac{\sigma^2_{0,Z}(X)}{1-p}\right].$$

Calculating, we have $c_Z(X) = E[Z(1) - Z(0)|X] = c(X)$ and

$$b_Z(X) = E[Z(1)|X]\left(\frac{1-p}{p}\right)^{1/2} + E[Z(0)|X]\left(\frac{p}{1-p}\right)^{1/2} = b(X;p) - \gamma' h(X).$$

Finally, $\sigma^2_{d,Z}(X) = \text{Var}(Z(d)|X) = \text{Var}(Y(d)|X) = \sigma^2_d(X)$. Then the variance V above is

$$V = \text{Var}(c(X)) + E[\text{Var}(b - \gamma' h|\psi)] + E\left[\frac{\sigma^2_1(X)}{p} + \frac{\sigma^2_0(X)}{1-p}\right]$$

as claimed. □

Proof of Theorem 3.2. 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 $\widehat{c}_1 = E_n[Y_i|D_i = 1] - \widehat{\alpha}_1' E_n[\tilde{h}_i|D_i = 1]$ and $\widehat{\alpha}_1 = \text{Var}_n(\tilde{h}_i|D_i = 1)^{-1} \text{Cov}_n(\tilde{h}_i, Y_i|D_i = 1)$. 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 second equality since $E_n[D_i] = p$ identically. The third equality by expanding $D_i = D_i - p + p$ and using $E_n[\tilde{h}_i] = 0$ and $E_n[(D_i - p)E_n[h_i]] = 0$. The fourth equality is algebra and collecting terms. The fifth equality since $\bar{h}_1 - \bar{h}_0 = E_n[h_i(D_i - p)/p(1-p)]$ again using $E_n[D_i] = p$ and $c_p = \sqrt{p(1-p)}$ by definition.

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 $E_n[D_i \tilde{h}_{it} \tilde{h}_{it'}] = E_n[(D_i - p)\tilde{h}_{it}\tilde{h}_{it'}] + p E_n[\tilde{h}_{it}\tilde{h}_{it'}]$. Expanding the first term

$$\begin{aligned}
E_n[(D_i - p)\tilde{h}_{it}\tilde{h}_{it'}] &= E_n[(D_i - p)h_{it}h_{it'}] - E_n[h_{it}]E_n[(D_i - p)h_{it'}] - E_n[h_{it'}]E_n[(D_i - p)h_{it}] \\
&\quad + E_n[h_{it'}]E_n[h_{it}]E_n[D_i - p] = o_p(1).
\end{aligned}$$

The final equality follows since $E_n[(D_i - p)h_{it}h_{it'}] = o_p(1)$ by applying Lemma A.2 of Cytrynbaum (2023), using that $E[|h_{it}\tilde{h}_{it'}|] \le E[|h_i|_2^2] < \infty$, and similarly for the other terms. By WLLN, we also have $E_n[\tilde{h}_{it}\tilde{h}_{it'}] \xrightarrow{p} \mathrm{Var}(h)$. Then by continuous mapping $\mathrm{Var}_n(\tilde{h}_i | D_i = 1)^{-1} = \mathrm{Var}(h)^{-1} + o_p(1)$. Similar reasoning shows $\mathrm{Cov}_n(\tilde{h}_i, Y_i | D_i = 1) = \mathrm{Cov}(h_i, Y_i(1)) + o_p(1)$.

Then we have shown $\widehat{\alpha}_1 = \mathrm{Var}(h)^{-1}\mathrm{Cov}(h, Y(1)) + o_p(1) = \mathrm{Var}(h)^{-1}\mathrm{Cov}(h, m_1) + o_p(1)$. By symmetry, we also have $\widehat{\alpha}_0 = \mathrm{Var}(h)^{-1}\mathrm{Cov}(h, m_0) + o_p(1)$. 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 3.4, $\sqrt{n}(\widehat{\theta}_L - \mathrm{ATE}) \Rightarrow \mathcal{N}(0, V)$ with

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

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 A.4. 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 $\tilde{Y}_i = \widehat{\theta}_N(D_i - p) + \widehat{\gamma}_N'\tilde{h}_i + e_i$ and $\widehat{\theta}_N =$

$E_n[(\check{D}_i)^2]^{-1}E_n[\check{D}_iY_i]$ with partialled treatment $\check{D}_i = (D_i - p) - (E_n[\tilde{h}_i\tilde{h}_i']^{-1}E_n[\tilde{h}_i(D_i - p)])'\tilde{h}_i$. Squaring this expression gives

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

Using $E_n[\tilde{h}_i(D_i - p)] = O_p(n^{-1/2})$ by Lemma A.2 of Cytrynbaum (2023) and $E_n[\tilde{h}_i\tilde{h}_i'] \xrightarrow{p} \mathrm{Var}(h) \succ 0$, 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 shown that

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

Then the first claim follows from Theorem 3.4. For the efficiency claims, (a) if $p = 1/2$ and $\psi = 1$, then $f = b$ and $\gamma_N = \mathrm{argmin}_\gamma\,\mathrm{Var}(f - \gamma'h) = \mathrm{argmin}_\gamma\,E[\mathrm{Var}(b - \gamma'h|\psi)]$. For

(c), if $\psi = 1$ and $\text{Cov}(h, m_1 - m_0) = 0$, then we have

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

By expanding the variance, we have $\text{argmin}_\gamma \text{Var}(f - \gamma' h) = \text{argmin}_\gamma \text{Var}(b - \gamma' h)$. 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 $\text{argmin}_{g \in L_2(\psi)} E[(Y(d) - g(\psi) - \gamma' h)^2] = E[Y(d) - \gamma' h | \psi]$ 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

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

Let $F(x) = (1+p)f_1(x) + pf_0(x)$. Then by vanilla CLT we have $\sqrt{n}(\widehat{\theta}_{AIPW} - \text{ATE}) \Rightarrow \mathcal{N}(0, V)$ with $V = \text{Var}\left(\frac{D_i - p}{p - p^2}(Y_i - F(X_i))\right) \equiv \text{Var}(W_i)$ with $W_i = \frac{D_i - p}{p - p^2}(Y_i - F(X_i)) - \text{ATE}$. By fundamental expansion of the IPW estimator from Cytrynbaum (2023)

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

By the law of total variance and tower law

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

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. □

A.6 Proofs for Section 3.3

Proof of Theorem 3.9. By Theorem 3.2, 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 population 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. □

A.7 Proofs for Section 3.4

Proof of Theorem 3.15. By Frisch-Waugh $\check{Y}_i = \widehat{\theta}_{FE}\check{D}_i + \widehat{\gamma}'_{FE}\check{h}_i + e_i$ 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 $\widehat{\theta}_{FE} = E_n[(\bar{D}_i)^2]^{-1}E_n[\bar{D}_i Y_i]$ 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 A.8 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 A.2 of Cytrynbaum (2023). 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

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

By Lemma A.8 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 $c_p^{-1} E_n[\check{h}_i \check{h}_i']^{-1} E_n[\check{h}_i Y_i] \xrightarrow{p} E[\mathrm{Var}(h|\psi)]^{-1} E[\mathrm{Cov}(h, f|\psi)] = \mathrm{argmin}_\gamma E[\mathrm{Var}(f - \gamma' h|\psi)]$. Similar reasoning shows that $\widehat{\gamma}_{FE} = E_n[\check{h}_i \check{h}_i']^{-1} E_n[\check{h}_i Y_i] + O_p(n^{-1/2})$. Then we have representation $\widehat{\theta}_{FE} = \widehat{\theta} - (c_p^{-1}\widehat{\gamma}_{FE})'(\bar{h}_1 - \bar{h}_0)c_p + o_p(n^{-1/2})$. The efficiency claims follow identically to the reasoning in Theorem A.4. This finishes the proof. □

Proof of Theorem 3.23 (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 $Y_i = \widehat{c}_1 + \widehat{\alpha}_1' \check{h}_i + e_i$ with $E_n[e_i(1, \check{h}_i)|D_i = 1] = 0$. 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 A.8. 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 3.2, 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 3.4 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 3.23 (Part II). Next, we analyze the group OLS estimator. By Theorem 3.4, it suffices to show that $\widehat{\gamma}_G = \mathrm{Var}_g(h_g)^{-1}\,\mathrm{Cov}_g(h_g, y_g) = c_p \cdot E[\mathrm{Var}(h|\psi)]^{-1}E[\mathrm{Cov}(h, b|\psi)] + o_p(1)$. 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 A.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 A.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

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

Applying Theorem 3.4 completes the proof. □

Proof of Theorem 3.23 (Part III). Finally, we analyze the ToM estimator. From the work in part I of this proof we have

$$\begin{aligned}\widehat{\gamma}_{PL} = \operatorname{Var}_n(\check{h}_i|D_i=1)^{-1}\operatorname{Cov}_n(\check{h}_i,Y_i|D_i=1)\sqrt{\frac{1-p}{p}} \\ + \operatorname{Var}_n(\check{h}_i|D_i=0)^{-1}\operatorname{Cov}_n(\check{h}_i,Y_i|D_i=0)\sqrt{\frac{p}{1-p}}\end{aligned}$$

Comparing with Equation 3.10, it suffices to show that $\operatorname{Var}_n(\check{h}_i|D_i=1)^{-1}\operatorname{Var}_n(\check{h}_i) = o_p(1)$ and $\operatorname{Var}_n(\check{h}_i|D_i=0)^{-1}\operatorname{Var}_n(\check{h}_i) = o_p(1)$. This follows immediately from Lemma A.8. Applying Theorem 3.4 completes the proof. □

Proof of Theorem 3.24. First, consider the fixed effects estimator with

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

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 $Y_i = \widehat{\tau}_{FE}(D_i-p) + \widehat{\gamma}_{FE}'\check{h}_i + \widehat{\gamma}_z'\tilde{z}_i + e_{i,2}$. Let $\check{w}_i = (\check{h}_i, \tilde{z}_i)$ and $w_i = (h_i, z_i)$. Then by work in Theorem 3.15, $\widehat{\tau}_{FE} = E_n[(\bar{D}_i)^2]^{-1}E_n[\bar{D}_iY_i]$ 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

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

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 $\widehat{\gamma}_{FE}$. Define $g_i = (D_i - p, \tilde{z}_i)$. Let $\bar{h}_i = \check{h}_i - (E_n[g_ig_i']^{-1}E_n[g_i\check{h}_i])'g_i$. Then by Frisch-Waugh $\widehat{\gamma}_{FE} = E_n[\bar{h}_i\bar{h}_i']^{-1}E_n[\bar{h}_iY_i]$. 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 A.8. 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 $\widehat{\gamma}_{FE} = E_n[\bar{h}_i\bar{h}_i']^{-1}E_n[\bar{h}_iY_i] = E_n[\check{h}_i\check{h}_i']^{-1}E_n[\check{h}_iY_i] + o_p(1)$, the coefficient from the regression without strata variables z_i included shown in Theorem 3.15. Consider the coefficient $\widehat{\gamma}_z$ on $z(\psi)$. Let $q_i = (D_i - p, \check{h}_i)$ and $\bar{z}_i = \tilde{z}_i - (E_n[q_iq_i']^{-1}E_n[q_i\tilde{z}_i])'q_i$. We just showed 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 $E_n[\check{w}_i\check{w}_i'] \xrightarrow{p} \text{Diag}(E_n[\check{h}_i\check{h}_i'], E_n[\tilde{z}_i\tilde{z}_i'])$. Then it's easy to see from our expression for $\widehat{\tau}_{FE}$ that we may identify $\widehat{\gamma}_z = \widehat{\alpha}_1 + o_p(1)$. This finishes the proof for $\widehat{\tau}_{FE}$. The proofs for the modified partialled Lin estimator $\widehat{\tau}_{PL}$ and modified ToM estimators are similar and omitted for brevity.

□

A.8 Proofs for Section 4

Proof of Theorem 4.1. Define population augmented potential outcomes $Y^b(d) = Y(d) - c_p\gamma' h(X)$ for $d \in \{0,1\}$ with outcomes $Y_i^b = Y_i^b(D_i) = Y_i - c_p\gamma' h_i$. The proof of Theorem 3.4 showed that $\widehat{\theta}_{adj} = \bar{Y}_1^b - \bar{Y}_0^b + o_p(n^{-1/2})$. Define $\widehat{v}_1^b$, $\widehat{v}_0^b$, and $\widehat{v}_{10}^b$ to be the analogues of $\widehat{v}_1$, $\widehat{v}_0$, and $\widehat{v}_{10}$ substituting Y_i^b for Y_i^a. By applying Theorem 6.1 of Cytrynbaum (2023) to $\widehat{\theta}_b \equiv \bar{Y}_1^b - \bar{Y}_0^b$, we have $\widehat{V}_b = V + o_p(1)$ for variance estimator

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

Then it suffices to show the following claim: $\widehat{V} - \widehat{V}_b = o_p(1)$. We prove a slight generalization, letting $h_i(d)$ possibly have a potential outcomes structure and setting $h_i = h_i(D_i)$. The case with $h_i(1) = h_i(0) = h_i$ is a special case.

We work term by term. Define the weights $L_i = (D_i - p)/(p - p^2)$. Then we have $\text{Var}_n(L_iY_i^b) - \text{Var}_n(L_iY_i^a) = E_n[L_i^2(Y_i^b)^2] - E_n[L_iY_i^b]^2 - E_n[L_i^2(Y_i^a)^2] + E_n[L_iY_i^a]^2$. We have $E_n[L_iY_i^a]^2 - E_n[L_iY_i^b]^2 = \text{ATE}^2 - \text{ATE}^2 + o_p(1) = o_p(1)$ by previous work. Next, we have $|E_n[L_i^2(Y_i^b)^2] - E_n[L_i^2(Y_i^a)^2]| = |E_n[L_i^2(Y_i^b - Y_i^a)(Y_i^b + Y_i^a)]| \lesssim E_n[(Y_i^b - Y_i^a)^2]^{1/2} E_n[(Y_i^b + Y_i^a)^2]^{1/2}$. It's easy to see that $E_n[(Y_i^b + Y_i^a)^2]^{1/2} = O_p(1)$. We have $E_n[(Y_i^b - Y_i^a)^2] = c_p^2 E_n[(\gamma' h_i - \widehat{\gamma}' h_i)^2] = c_p^2(\widehat{\gamma} - \gamma)' E_n[h_ih_i'](\widehat{\gamma} - \gamma) = o_p(1)$. This shows that $\text{Var}_n(L_iY_i^b) - \text{Var}_n(L_iY_i^a) = o_p(1)$, completing the proof for the first term.

Next consider $\widehat{v}_1^b - \widehat{v}_1$. We may expand

$$\widehat{v}_1^b - \widehat{v}_1 = n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \frac{1}{a(g) - 1} \frac{1 - p}{p^2} \sum_{i \neq j \in g} D_iD_j(Y_i^aY_j^a - Y_i^bY_j^b).$$

Note that $Y_i^aY_j^a - Y_i^bY_j^b = (Y_i^a - Y_i^b)Y_j^a + Y_i^b(Y_j^a - Y_j^b) = c_p(\widehat{\gamma} - \gamma)'(h_iY_j^a + Y_i^bh_j)$. Then by triangle inequality and Cauchy-Schwarz

$$\begin{aligned} |\widehat{v}_1^b - \widehat{v}_1| &= \left| c_p(\widehat{\gamma} - \gamma)' n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \frac{1}{a(g) - 1} \frac{1 - p}{p^2} \sum_{i \neq j \in g} D_iD_j(h_iY_j^a + Y_i^bh_j) \right| \\ &\lesssim |\widehat{\gamma} - \gamma|_2 \left(n^{-1} \sum_{g \in \mathcal{G}_n^\nu} \sum_{i \neq j \in g} |h_i|_2 |Y_j^a| + |Y_i^b||h_j|_2 \right) \end{aligned}$$

Observe that

$$\sum_{i\neq j\in g} |h_i|_2|Y_j^a| \leq (1/2)\sum_{i\neq j\in g} |h_i|_2^2 + |Y_j^a|^2 = \frac{k-1}{2}\sum_{i\in g} |h_i|_2^2 + |Y_i^a|^2$$

Then since $\mathcal{G}_n^\nu$ is a partition of $[n]$ we have $|\widehat{v}_1^b - \widehat{v}_1| \lesssim |\widehat{\gamma} - \gamma|_2 E_n[|h_i|_2^2 + |Y_i^a|^2] = o_p(1)O_p(1) = o_p(1)$. Then by symmetry $\widehat{v}_0^b - \widehat{v}_0 = o_p(1)$ as well. A similar calculation shows that $\widehat{v}_{10}^b - \widehat{v}_{10} = o_p(1)$. Then we have shown that $\widehat{V}_b - \widehat{V} = o_p(1)$, which completes the proof. □

A.9 Proofs of Noncompliance Theorems

Proof of Theorems A.1, A.2, A.3. First we show Theorem A.1. Define $\widehat{\theta}^W(\alpha) = \bar{W}_1 - \bar{W}_0 - \alpha'(\bar{h}_1 - \bar{h}_0)c_p$ and similary for $\widehat{\theta}^D(\alpha)$. We claim that $\widehat{\tau}_{adj} = \widehat{\theta}^W(\gamma_W)/\widehat{\theta}^D(\gamma_D) + o_p(n^{-1/2})$. By algebra, we have

$$\widehat{\tau}_{adj} - \frac{\widehat{\theta}^W(\gamma_W)}{\widehat{\theta}^D(\gamma_D)} = \frac{\widehat{\theta}^D(\gamma_D)(\widehat{\gamma}_W - \gamma_W)'(\bar{h}_1 - \bar{h}_0)c_p + \widehat{\theta}^W(\gamma_W)(\gamma_D - \widehat{\gamma}_D)'(\bar{h}_1 - \bar{h}_0)c_p}{\widehat{\theta}^D(\gamma_D)\widehat{\theta}^D(\widehat{\gamma}_D)}$$

By Theorem 3.4, $\widehat{\theta}^D(\gamma_D), \widehat{\theta}^D(\widehat{\gamma}_D) = \tau_D + o_p(1)$ with $\tau_D > 0$, so the denominator is $O_p(1)$. The numerator is $o_p(n^{-1/2})$ since $\widehat{\theta}^D(\gamma_D), \widehat{\theta}^W(\gamma_W) = O_p(1)$ and $(\widehat{\gamma}_A - \gamma_A)'(\bar{h}_1 - \bar{h}_0)c_p = o_p(n^{-1/2})$ for $A = D, W$ by the first line of the proof of Theorem 3.4. Next, recall the potential outcomes $Q(z) = W(z) - \tau_L D(z)$ and define $\gamma_Q = \gamma_W - \tau_L\gamma_D$. Then we have

$$\frac{\widehat{\theta}^W(\gamma_W)}{\widehat{\theta}^D(\gamma_D)} - \tau_L = \frac{\widehat{\theta}^W(\gamma_W) - \tau_L\widehat{\theta}^D(\gamma_D)}{\widehat{\theta}^D(\gamma_D)} = \frac{\widehat{\theta}^Q(\gamma_Q)}{\widehat{\theta}^D(\gamma_D)}.$$

The ATE-like quantity $E[Q(1) - Q(0)] = 0$ by definition of τ_L. Then by Theorem 3.4, we have $\sqrt{n}\widehat{\theta}^Q(\gamma_Q) \Rightarrow \mathcal{N}(0, V_Q)$ with variance

$$V_Q = \mathrm{Var}(c_Q) + E\Big[\mathrm{Var}(b_Q - h'\gamma_Q|\psi)\Big] + E\left[\frac{\sigma_{1Q}^2(X)}{p} + \frac{\sigma_{0Q}^2(X)}{1-p}\right]. \tag{A.4}$$

The claim now follows by Slutsky since $\widehat{\theta}^D(\gamma_D) = E[D(1) - D(0)] + o_p(1)$ so that $\sqrt{n}(\widehat{\tau}_{adj} - \tau_L) = \sqrt{n}\widehat{\theta}^Q(\gamma_Q)/\widehat{\theta}^D(\gamma_D) + o_p(1) = \sqrt{n}\widehat{\theta}^Q(\gamma_Q)/E[D(1) - D(0)] + o_p(1)$.

Next, we prove Theorem A.2. By linearity of the balance function (Equation 2.2), we have $b_Q = b_W - \tau_L b_D$. The optimal coefficient is $\gamma_Q^* = E[\mathrm{Var}(h|\psi)]^{-1}E[\mathrm{Cov}(h, b_Q|\psi)] = E[\mathrm{Var}(h|\psi)]^{-1}(E[\mathrm{Cov}(h, b_W|\psi)] - \tau_L E[\mathrm{Cov}(h, b_D|\psi)]) = \gamma_W^* - \tau_L\gamma_D^*$. This shows that $\widehat{\tau}_{adj}$ is efficient if and only if $\gamma_W - \tau_L\gamma_D = \gamma_W^* - \tau_L\gamma_D^*$. In particular, this holds if $\gamma_W = \gamma_W^*$ and $\gamma_D = \gamma_D^*$. By the estimator representations in Section 3.4, the estimator $\widehat{\theta}_k^W = \bar{W}_1 - \bar{W}_0 - \widehat{\gamma}_{W,k}'(\bar{h}_1 - \bar{h}_0)c_p$ for $\widehat{\gamma}_{W,k} = \gamma_W^* + o_p(1)$ for $k \in \{PL, GO, TM\}$, and similarly for $\widehat{\theta}_k^D$. Then $\widehat{\tau}_L^k$ is efficient for each $k \in \{PL, GO, TM\}$.

Finally, we show Theorem A.3. With $\gamma_Q = \gamma_W - \tau_L \gamma_D$, define the "population" augmented potential outcomes $Q^b(z) = Q(z) - h'\gamma_Q$ and outcomes $Q_i^b = Q_i - h_i'\gamma_Q$. Let $\widehat{V}_Q^a$ denote the bracketed term in Equation 4.1, and let $\widehat{V}_Q^b$ denote the bracketed term with Q_i^a replaced by the population version Q_i^b. Note that we showed above that $\sqrt{n}(\bar{Q}_1^b - \bar{Q}_0^b) \Rightarrow N(0, V_Q)$. Then $\widehat{V}_Q^b = V_Q + o_p(1)$ by Theorem 4.1. Then it suffices to show that $\widehat{V}_Q^b - \widehat{V}_Q^a = o_p(1)$. To see this, note that we may write $Q_i^b = W_i - \beta' S_i$ and $Q_i^a = W_i - \widehat{\beta}' S_i$ with $\widehat{\beta} = \beta + o_p(1)$ for $\widehat{\beta} = (\widehat{\tau}_L^k, \widehat{\gamma}_Q)$, $\beta = (\tau_L, \gamma_Q)$ and $S_i = (D_i, h_i)$. Then the fact that $\widehat{V}_Q^b - \widehat{V}_Q^a = o_p(1)$ for outcomes of this form and $\widehat{\beta} = \beta + o_p(1)$ is exactly what we showed in the main claim in the proof of Theorem 4.1. This finishes the proof. □

A.10 Technical Lemmas

Lemma A.5 (Conditional Convergence)**.** *Let $(\mathcal{G}_n)_{n\geq 1}$ and $(A_n)_{n\geq 1}$ a sequence of σ-algebras and RV's. Then the following results hold*

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

(ii) $\text{Var}(A_n|\mathcal{G}_n) = o_p(c_n^2)/O_p(c_n^2) \implies A_n - E[A_n|\mathcal{G}_n] = o_p(c_n)/O_p(c_n)$ *for any positive sequence* $(c_n)_n$.

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

See Appendix C of Cytrynbaum (2023) for the proof.

Lemma A.6. *Let $(a_i), (b_i), (c_i)$ be positive scalar arrays for $i \in I$ for some index set I. Then we have $\sum_{\substack{i,j,s\in I \\ i\neq j, j\neq s}} a_i b_j c_s \leq 3\sum_{i\in I}(a_i^3 + b_i^3 + c_i^3)$.*

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

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

□

Lemma A.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 $\bar{h}_{g1} = a^{-1} \sum_{i \in g} h_i \mathbb{1}(D_i = 1)$, $\bar{h}_{g0} = (k-a)^{-1} \sum_{i \in g} h_i \mathbb{1}(D_i = 0)$, and $\bar{w}_g = k^{-1} \sum_{i \in g} w_i$. Recall that $g \in \sigma(\psi_{1:n}, \pi_n)$ for each g and $D_{1:n} \in \sigma(\psi_{1:n}, \pi_n, \tau)$ for an exogenous variable $\tau \perp\!\!\!\perp (X_{1:n}, Y(0)_{1:n}, Y(1)_{1:n})$ used to randomize treatments. 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 C.10 of Cytrynbaum (2023), we have $E[B_n | X_{1:n}, \pi_n] = 0$. Next, we have

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

The second equality follows by Lemma C.10 of Cytrynbaum (2023), 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. Then by Lemma A.5, $B_n = O_p(n^{-1/2})$.

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 (2022). Moreover, we have

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

For $j \neq t$, we have $E[u_j u_t|\psi_{1:n}, \pi_n] = E[u_j|\psi_j]E[u_t|\psi_t] = 0$ 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 A.5. 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

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

We have $|\operatorname{Cov}((D_i-p)(D_j-p),(D_s-p)(D_t-p)|X_{1:n},\pi_n)| \le 2$ since $|(D_i-p)|\le 1$ for all $i\in[n]$. Using Lemma 9.17 in Cytrynbaum (2022), the previous display is bounded above by $\sum_{i,j\in g}\sum_{s,t\in g}|h_i w_j h_s w_t|\cdot 2 \le 2k^3\sum_{i\in g}(h_i^4+w_i^4)$. Putting this all together,

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

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 $A_n = \frac{E[\operatorname{Cov}(h,w|\psi)]}{a(k-a)} + o_p(1)$.

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 (2022), 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 (2022). 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

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

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 A.5, $C_n=O_p(n^{-1/2})$. This finishes the proof. □

Lemma A.8 (Partialled Lin)**.** *Under assumptions, $E_n[\check{h}_iz_i]=o_p(1)$. Also, we have*

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

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 $1\leq t,t'\leq d_h$, by Lemma A.2 of Cytrynbaum (2022) and Cauchy-Schwarz we have $\mathrm{Var}(\sqrt{n}E_n[(D_i-p)\check{h}_{it}\check{h}_{it'}]|X_{1:n},\pi_n)\leq 2E_n[\check{h}_{it}^2\check{h}_{it'}^2]\leq 2E_n[\check{h}_{it}^4]^{1/2}E_n[\check{h}_{it'}^4]^{1/2}$. Next, note that by Jensen's followed by Young's inequality

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

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 A.5. 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

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

We can expand the expression above as

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

First consider A_n. By the Lipschitz assumption in 3.1 and Young's inequality

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

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 $\mathrm{Var}(\sqrt{n}B_n|\psi_{1:n},\pi_n) = E[nB_n^2|\psi_{1:n},\pi_n]$ as follows. By Lemma 9.21 of Cytrynbaum (2022), $E[u_{it'}u_{jt'}|\psi_{1:n},\pi_n] = 0$ 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 (2022), $E[(u_{it'} - u_{lt'})(u_{at'} - u_{ct'})|\psi_{1:n}, \pi_n]$ is

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

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 3.1, 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 A.5 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 (2022) 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 we calculate

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

Now $E[E[u_{it}u_{it'}|\psi_i]^2] \leq E[u_{it}^2 u_{it'}^2] \leq 2E[u_{it}^4] + 2E[u_{it'}^4] < \infty$ by Jensen, tower law,

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

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

For any variables A, B, $|\text{Cov}(A, B)| \leq |E[AB]| + |E[A]E[B]| \leq 2|A|_2|B|_2$ 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

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

Plugging this bound in above and summing out gives

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

The final equality by Markov since $E[u_{it}^4] < \infty$. Then by conditional Markov A.5 we have $D_n = \frac{k-1}{k}E[\text{Cov}(h_{it}, h_{it'}|\psi_i)] + O_p(n^{-1/2})$. Since t, t' were arbirary, this shows $E_n[\check{h}_i\check{h}_i'] = E[\text{Var}(h|\psi)] + o_p(1)$.

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 A.2 of Cytrynbaum (2023), and Cauchy-Schwarz

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

Note that $E_n[Y_i(1)^4] = O_p(1)$ by Markov inequality and Assumption 3.1 and $E_n[\check{h}_{it}^4] = O_p(1)$ was shown above. Then by Lemma A.5 (conditional Markov), this shows the claim. Then it suffices to analyze $E_n[\check{h}_iY_i(1)]$. 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 3.1, $\psi \to g(\psi)$ 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. Proposition 8.6 of Cytrynbaum (2022) showed that this quantity is also $o_p(1)$. 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 (2022). 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. □