Electronic Journal of Statistics
Vol. 18 (2024) 1436–1454
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2233

A unified analysis of regression adjustment in randomized experiments

Katarzyna Reluga[1], **Ting Ye**[2] **and Qingyuan Zhao**[3]

[1] *School of Mathematics, Fry Building, University of Bristol*
Woodland Road, Bristol, BS8 1UG, U.K.
e-mail: katarzyna.reluga@bristol.ac.uk

[2] *Department of Biostatistics, University of Washington*
3980 15th Avenue NE, Box 351617, Seattle, WA 98195, U.S.A.
e-mail: tingye1@uw.edu

[3] *Department of Pure Mathematics and Mathematical Statistics, University of Cambridge*
Centre for Mathematical Sciences
Wilberforce Road, Cambridge CB3 0WB, U.K.
e-mail: qyzhao@statslab.cam.ac.uk

Abstract: Regression adjustment is broadly applied in randomized trials under the premise that it usually improves the precision of a treatment effect estimator. However, previous work has shown that this is not always true. To further understand this phenomenon, we develop a unified comparison of the asymptotic variance of a class of linear regression-adjusted estimators. Our analysis is based on the classical theory for linear regression with heteroscedastic errors and thus does not assume that the postulated linear model is correct. For a randomized Bernoulli trial, we provide sufficient conditions under which some regression-adjusted estimators are guaranteed to be more asymptotically efficient than others. We comment on the extension of our theory to other settings such as general treatment mechanisms and generalized linear models, and find that the variance dominance phenomenon no longer occurs

MSC2020 subject classifications: Primary 62F10; secondary 62J99.
Keywords and phrases: Average treatment effect, covariate adjustment, heteroscedasticity, randomized controlled trials.

Received October 2022.

1. Introduction

Randomized experiments are the gold standard to answer questions about causality. Many researchers use multiple linear regression with a treatment indicator and some baseline covariates to analyze randomized experiments, in which the treatment coefficient is often interpreted as a causal effect. In some fields, this is known as the “analysis of covariance” (ANCOVA), which was first proposed by Fisher [3] to unify “two very widely applicable procedures known as regression and analysis of variance”. This common practice is motivated by the belief that regression adjustments can increase precision if covariates in the regression are predictive of the outcome.

arXiv: 2210.04360

However, as pointed out by many authors, this is not always true especially when there is a lot of treatment effect heterogeneity. Regression adjustment in randomized experiments has been studied in two different frameworks, namely the finite-population model and the super-population model. In the former, the experimental units are treated as fixed and the randomness comes from the process of randomising the treatments [12, 14] whereas the latter assumes that the experimental units are drawn independently from an infinite population [see e.g. 8, Chapter 7]. Three estimators have been extensively studied in the literature: the simple difference-in-means or analysis of variance (ANOVA) estimator; the ANCOVA estimator that includes covariate main effects; and the analysis of heterogeneous covariance (ANHECOVA)[1] estimator that includes covariate main effects and all treatment-covariate interactions. Regardless of whether the finite-population model [4, 5, 16, 10, 7] or super-population model [9, 22, 19, 16, 15, 23] is used, the main conclusions about the asymptotic efficiency of these estimators are the same. Consider two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ that converge to the same limit. We say that $\hat{\beta}_1$ (asymptotically) *uniformly dominates* $\hat{\beta}_2$ if the (asymptotic) variance of $\hat{\beta}_1$ is always smaller than or equal to that of $\hat{\beta}_2$, no matter what the underlying distribution is. In both the finite-population model and the super-population model, it has been found that ANHECOVA uniformly dominates the other two, but, somewhat surprisingly, ANCOVA does not uniformly dominate ANOVA.

A major limitation of the existing analysis of regression adjustment is that the investigations are restricted to specific estimators and provide limited insights into the phenomenon of uniform dominance. In particular, authors rarely discuss whether their theory, established within the framework of Bernoulli randomized trials or other completely randomized experiments without stratification, extends to other cases such as stratified experiments or generalized linear models. The variance calculations are often quite technical, which further makes the theoretical results less accessible to practitioners.

In this article, we provide a unified analysis for a large class of linear-regression adjusted estimators in the super-population framework. Besides the estimators mentioned above, our theory also applies to regression estimators with some coefficients fixed (such as the difference-in-differences estimator for which the coefficient corresponding to the baseline value of the response is included into the covariate vector, and it is fixed to be 1) or with treatment-covariate interactions only. By a simple application of the textbook theory for linear regression with heteroscedastic errors, this analysis not only recovers the known relationships between ANOVA, ANCOVA, and ANHECOVA, but also immediately provides a sufficient condition for uniform dominance when the expectation of the covariates is known (see Theorem 3.4 below). In the more practical situation when the covariate expectation is unknown, a slightly different sufficient condition is obtained (see Theorem 3.6 below). This condition shows that, for example, the so-called lagged-dependent-variable regression estimator is more efficient

[1] The term ‘ANHECOVA’ was coined by Ye et al. [23] but the estimator was used before by, among others, [10, 19, 20, 22].

than the difference-in-differences estimator in randomized experiments, despite them having a bracketing relationship in observational studies [2]. This unified analysis allows us to explore whether the uniform dominance extends to more complicated settings and provide numerical counterexamples. Some further remarks are provided at the end of this article, whereas proofs of the technical Lemmas can be found in Appendix A.

2. Linear regression adjustment in randomized trials

Consider a random sample $\{(A_i, X_i^T, Y_i)\}_{i=1}^n$ of n units, where $A_i \in \{0,1\}$ is a binary treatment indicator, $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T \in \mathbb{R}^p$ is a vector of unit covariates observed before treatment assignment, and $Y_i \in \mathbb{R}$ is a real-valued outcome of the unit. We assume that $(A_i, X_i^T, Y_i), i = 1, \dots, n$ are independent and identically distributed, which is often a good approximation when the units are randomly sampled from a large population. To simplify the notation, we drop the subscript i when referring to a generic unit from the population.

Unless mentioned otherwise, we assume that each unit receives the treatment independently with equal probability $\mathrm{pr}(A = 1 \mid X) = \pi$, where $0 < \pi < 1$ is a known constant. In other words, treatment is assigned by a simple Bernoulli trial, which approximates random sampling without replacement that is often studied in the finite-population model [4, 5, 10]. Under this assignment mechanism and standard assumptions in causal inference, the average treatment effect $\beta_{\mathrm{ATE}} = E[Y(1) - Y(0)]$, where $Y(a)$ is the potential outcome of unit i under treatment level a, can be identified as [see e.g. 8, Chapter 7]:

$$\beta_{\mathrm{ATE}} = E(Y \mid A = 1) - E(Y \mid A = 0). \tag{1}$$

In this article, we consider the following class of regression adjusted estimators of β_{ATE}. Let $\Gamma = \Gamma^{(1)} \times \cdots \times \Gamma^{(p)} \subseteq \mathbb{R}^p$ and $\Delta = \Delta^{(1)} \times \cdots \times \Delta^{(p)} \subseteq \mathbb{R}^p$ be two user-specified sets, where the individual components $\Gamma^{(j)}$ and $\Delta^{(j)}$ are either the real line $\mathbb{R}$ or a singleton. Define the constrained ordinary least squares estimator as

$$\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}) = \underset{\gamma \in \Gamma, \delta \in \Delta}{\arg\min} \frac{1}{n} \sum_{i=1}^n \{Y_i - \alpha - \beta A_i - \gamma^T X_i - A_i(\delta^T X_i)\}^2. \tag{2}$$

We sometimes use the notation $\hat{\theta}(\Gamma, \Delta)$ (and similarly for the components of $\hat{\theta}$) to emphasize the dependence of the estimator on the sets Γ and Δ. Lemma 3.1 in Section 3.1 shows that $\hat{\beta}$ is a reasonable estimator of β_{ATE} when the covariates are centered, i.e. $E(X) = 0$; otherwise β_{ATE} can be estimated by $\tilde{\beta} = \hat{\beta} + \hat{\delta}^T \bar{X}$, where $\bar{X} = \sum_{i=1}^n X_i/n$. Before examining the asymptotic properties of $\hat{\beta}$ and $\tilde{\beta}$, we give several examples in the class of estimators (2).

Example 1. The ANOVA, ANCOVA, ANHECOVA estimators correspond to setting $\Gamma = \Delta = \{0\}$; $\Gamma = \mathbb{R}^p$ and $\Delta = \{0\}$; $\Gamma = \mathbb{R}^p$ and $\Delta = \mathbb{R}^p$.

Example 2. In some applications, the covariate vector X includes the baseline value of the response before the treatment is assigned (let us call it Y_0). For simplicity, suppose the first entry of X is Y_0, so $X = (X_1 = Y_0, X_2, \dots, X_p)^T$. The difference-in-differences estimator corresponds to setting $\Gamma = \{1\} \times \mathbb{R}^{p-1}$ and $\Delta = \{0\} \times \mathbb{R}^{p-1}$, while the lagged-dependent-variable regression estimator corresponds to setting $\Gamma \subseteq \mathbb{R}^p$ and $\Delta = \{0\} \times \mathbb{R}^{p-1}$. In observational studies, these two estimators rely on different identification assumptions [2] and may converge to different limits. In the randomized experiment described above, both estimators should converge to the average treatment effect, but we are unaware of any comparison of their statistical efficiency in presence of covariates besides Y_0.

Remark 1. The assumption that the individual components $\Gamma^{(j)}$ and $\Delta^{(j)}$ are either the real line $\mathbb{R}$ or a singleton is essential to prove the results in Theorems 3.4 and 3.6. Otherwise, we would need to consider the uniform dominance in regions defined by the intervals of Γ and Δ which would lead to the loss of generality of our theory.

3. A unified analysis of linear regression-adjusted estimators

3.1. Covariates with known expectation

We first consider estimation of β_{ATE} when the covariates X have known expectation. As will be seen in a moment, the proof of uniform dominance is fairly straightforward in this case.

Consider the population counterpart to (2):

$$\theta = (\alpha, \beta, \gamma, \delta) = \underset{\gamma \in \Gamma, \delta \in \Delta}{\arg\min}\, E\{Y - \alpha - \beta A - \gamma^T X - A(\delta^T X)\}^2. \tag{3}$$

Clearly, $\theta = \theta(\Gamma, \Delta)$, and we often suppress the dependence of θ on (Γ, Δ) if it is clear from the context.

Lemma 3.1. *For any Γ and Δ of the form described in Section 2, we have $\beta = \beta_{ATE} - \delta^T E(X)$.*

Without loss of generality, we shall assume that $E(X) = 0$ for the rest of Section 3.1; otherwise, we can simply replace X with $X - E(X)$ since $E(X)$ is known. When $E(X) = 0$, Lemma 3.1 shows that $\hat{\beta}$ is a reasonable estimator of β_{ATE}. To study the asymptotic properties of $\hat{\beta}$, we first state a classical result for linear regression with heteroskedastic error. For a proof of this result, see e.g. White [21].

Lemma 3.2. *Consider a linear regression of an independent and identically distributed sample of response $Y \in \mathbb{R}$ on regressors $Z \in \mathbb{R}^p$. Let $\hat{\theta}$ and θ be sample and population least squares estimators and $\epsilon(\theta) = Y - \theta^T Z$. Suppose that $E(ZZ^T)$ and $E\{\epsilon(\theta)^2 ZZ^T\}$ are positive definite and Y, Z have bounded* *fourth moments. Then, as $n \to \infty$, $\hat{\theta} \to \theta$ in probability and*

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to N\left(0, \{E(ZZ^T)\}^{-1}E(ZZ^T\epsilon(\theta)^2)\{E(ZZ^T)\}^{-1}\right) \text{ in distribution.} \tag{4}$$

Note that these results do not require that the linear model is correctly specified. By applying Lemma 3.2 to our problem with $Z = (1, A, X^T, AX^T)^T$ and regression error

$$\epsilon = \epsilon(\theta) = Y - \alpha - \beta A - \gamma^T X - A(\delta^T X), \tag{5}$$

we obtain the expression for the asymptotic variance of $\hat{\beta}$. The proof of this result is straightforward due to the block diagonal structure of $E(ZZ^T)$. This is made possible by the assumption that $E(X) = 0$ (for details, see the proof of Lemma 3.2 in Appendix A.3).

Lemma 3.3. *Suppose that $E(X) = 0$ and the regularity conditions in Lemma 3.2 are satisfied. Then, as $n \to \infty$, we have*

$$\sqrt{n}(\hat{\beta} - \beta) \to N\left(0, \frac{E\{(A-\pi)^2\epsilon^2\}}{\pi^2(1-\pi)^2}\right) \text{ in distribution.}$$

To state our first main result about uniform dominance, we introduce an additional notation. Let $\mathcal{U}(\Gamma) \subseteq \{1, \ldots, p\}$ denote the unrestricted dimensions of Γ, i.e. $\mathcal{U}(\Gamma) = \{1 \leqslant j \leqslant p : \Gamma^{(j)} = \mathbb{R}\}$. In words, $\mathcal{U}(\Gamma)$ represents a subset of $\{1, \ldots, p\}$, where each element j of $\mathcal{U}(\Gamma)$ corresponds to an individual component $\Gamma^{(j)} \subseteq \Gamma$, which is a real line denoted as $\mathbb{R}$. Similarly, $\mathcal{U}(\Delta)$ denotes the unrestricted dimensions of Δ.

Theorem 3.4. *Suppose $E(X) = 0$. Consider two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained from the least squares problem (2) with $(\Gamma, \Delta) = (\Gamma_1, \Delta_1)$ and (Γ_2, Δ_2), respectively, and suppose $(\Gamma_1, \Delta_1) \neq (\Gamma_2, \Delta_2)$. Then $\hat{\beta}_1$ uniformly dominates $\hat{\beta}_2$ if*

$$\Gamma_1 \supseteq \Gamma_2,\ \Delta_1 \supseteq \Delta_2,\ \text{and either } \pi = 1/2 \text{ or } \mathcal{U}(\Delta_1) \supseteq \mathcal{U}(\Gamma_1). \tag{6}$$

Proof. The first-order condition for the unrestricted version of the least squares problem (3) can be written as

$$E(\epsilon) = 0,\ E(\epsilon A) = 0,\ E\{\epsilon X_{\mathcal{U}(\Gamma)}\} = 0,\ E\{\epsilon A X_{\mathcal{U}(\Delta)}\} = 0. \tag{7}$$

Let $\theta_k = (\alpha_k, \beta_k, \gamma_k, \delta_k)$ be the solution for when $(\Gamma, \Delta) = (\Gamma_k, \Delta_k)$ and $\epsilon_k = Y - \alpha_k - \beta_k A - \gamma_k^T X - A(\delta_k^T X)$ be the corresponding regression error, $k = 1, 2$. By Lemma 3.1, $\epsilon_2 = \epsilon_1 + (\gamma_1 - \gamma_2)^T X + A(\delta_1 - \delta_2)^T X$. Let $V_k = [E\{(A - \pi)^2\epsilon_k^2\}]/\{\pi^2(1-\pi)^2\}$. Then by Lemma 3.3,

$$\begin{aligned}\pi^2(1-\pi)^2(V_2 - V_1) =& E\{(A-\pi)^2(\epsilon_2^2 - \epsilon_1^2)\} \\ =& E\left[(A-\pi)^2 2\epsilon_1\{(\gamma_1 - \gamma_2)^T X + A(\delta_1 - \delta_2)^T X\}\right] \\ &+ E\left[(A-\pi)^2\{(\gamma_1 - \gamma_2)^T X + A(\delta_1 - \delta_2)^T X\}^2\right]\end{aligned}$$

$$\geq 2E\left[(A-\pi)^2\epsilon_1\{(\gamma_1-\gamma_2)^T X + A(\delta_1-\delta_2)^T X\}\right]. \tag{8}$$

Since $\mathcal{U}(\Gamma_k)$ contains the unrestricted dimensions of γ_k, $k=1,2$, and $\Gamma_1 \supseteq \Gamma_2$ by assumption, the non-zero elements of $\gamma_1-\gamma_2$ can only appear in $\mathcal{U}(\Gamma_1)$ (otherwise the coefficients are fixed by design). Similarly, the non-zero elements of $\delta_1-\delta_2$ appear in $\mathcal{U}(\Delta_1)$. By using (7) and $A \perp\!\!\!\perp X$, $A^2=A$, we have

$$\begin{aligned}&E\left[(A-\pi)^2\epsilon_1\{(\gamma_1-\gamma_2)^T X + A(\delta_1-\delta_2)^T X\}\right]\\ =&(\gamma_1-\gamma_2)^T_{\mathcal{U}(\Gamma_1)}\left\{E\left[(1-2\pi)\epsilon_1 A X_{\mathcal{U}(\Gamma_1)}\right]+E\left[\pi^2\epsilon_1 X_{\mathcal{U}(\Gamma_1)}\right]\right\}\\ &+(\delta_1-\delta_2)^T_{\mathcal{U}(\Delta_1)}E\left[(1-\pi)^2\epsilon_1 A X_{\mathcal{U}(\Delta_1)}\right]=(\gamma_1-\gamma_2)^T_{\mathcal{U}(\Gamma_1)}E\left[(1-2\pi)\epsilon_1 A X_{\mathcal{U}(\Gamma_1)}\right],\end{aligned}$$

where the last equality follow from applying (7) to $\epsilon=\epsilon_1$ and $(\Gamma,\Delta)=(\Gamma_1,\Delta_1)$. Finally, $E\left[(1-2\pi)\epsilon_1 A X_{\mathcal{U}(\Gamma_1)}\right]=0$ if $\pi=1/2$ or $\mathcal{U}(\Delta_1)\supseteq\mathcal{U}(\Gamma_1)$ by (7). □

In words, Theorem 3.4 says that, when the expectation of the covariates is known, one linear regression-adjusted estimator is uniformly dominated by another if the two linear models are nested, the first estimator is obtained from the larger model, and the larger model includes an interaction term whenever the corresponding main effect is present; there is no such requirement for the smaller model. The conditions in (6) can be easily applied to obtain variance orderings among the examples in Section 2. We will discuss them in more detail after deriving a similar sufficient condition when the expectation of X is unknown.

3.2. Covariates with unknown expectation

In most practical situations, we do not know the expectation of the covariates and it is common to centre the covariates empirically before performing the linear regression. Let $\tilde\theta$ be the least squares estimator in (2) with X_i replaced by $X_i-\bar X$ where $\bar X=\sum_{i=1}^n X_i/n$, that is,

$$\tilde\theta=(\tilde\alpha,\tilde\beta,\tilde\gamma,\tilde\delta)=\underset{\gamma\in\Gamma,\delta\in\Delta}{\arg\min}\,\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\alpha-\beta A_i-\gamma^T(X_i-\bar X)-A_i\left\{\delta^T(X_i-\bar X)\right\}\right]^2. \tag{9}$$

We have $\tilde\beta=\hat\beta$ if no interaction term is included, i.e. if $\Delta=\{0\}^p$, because both (2) and (9) include an intercept term. More generally, by differentiating (9) with respect to α and β and following the proof of Lemma 3.1, it is straightforward to verify that $\tilde\beta=\hat\beta+\hat\delta^T\bar X$. Thus, $\tilde\beta$ is a reasonable estimator of $\beta_{\text{ATE}}=\beta+\delta^T E(X)$. Estimator $\tilde\theta$ is invariant to any shift transformation of the covariates. In other words, $\tilde\theta$ remains the same if we replace X_i by X_i+c, $i=1,\dots,n$, for any $c\in\mathbb{R}^p$. Therefore, the statistical properties of $\tilde\theta$ do not depend on $E(X)$ and, for simplifying the analysis, we shall assume $E(X)=0$ without loss of generality. The asymptotic variance of $\tilde\beta$ generally differs from that of $\hat\beta$ due to the variability in $\bar X$. The next result quantifies this difference and it is proved in Appendix A.4.

Table 1

Variance ordering of estimators of β_{ATE} in (1). V_k: variance of $\hat{\beta}_k$, $k = 1, 2$; $\tilde{V}_k$: variance of $\tilde{\beta}_k$, $k = 1, 2$.

Model 1	Model 2	$V_1 \leq V_2$	$\tilde{V}_1 \leq \tilde{V}_2$
$\sim 1 + A + X + A : X$	$\sim 1 + A$	True: Theorem 3.4	True: Theorem 3.6
$\sim 1 + A + X + A : X$	$\sim 1 + A + X$	True: Theorem 3.4	True: Theorem 3.6
$\sim 1 + A + X + A : X$	$\sim 1 + A + A : X$	True: Theorem 3.4	True: Theorem 3.6
$\sim 1 + A + X$	$\sim 1 + A$	Not always true: eq. (16)	
$\sim 1 + A + A : X$	$\sim 1 + A$	True: eq. (17)	Not always true: eq. (18)

Lemma 3.5. *Consider two estimators $\hat{\beta} = \hat{\beta}(\Gamma, \Delta)$ and $\tilde{\beta} = \tilde{\beta}(\Gamma, \Delta)$ for some user-specified (Γ, Δ). Under regularity conditions in Lemma 3.2 and as $n \to \infty$, we have*

$$n \left\{ var(\tilde{\beta}) - var(\hat{\beta}) \right\} \to \delta_s^T \Sigma (2\delta_f - \delta_s),$$

where $\delta_f = \delta_0(\mathbb{R}^p, \mathbb{R}^p)$ and $\delta_s = \delta_0(\Gamma, \Delta)$ are obtained by solving the population least squares problem (3) for the full model and a sub-model defined by $\Gamma \subseteq \mathbb{R}^p$ and $\Delta \subseteq \mathbb{R}^p$, respectively.

Let $\hat{\theta}_k$ and $\tilde{\theta}_k$ be the solution to (2) and (9), respectively, for the choice $(\Gamma, \Delta) = (\Gamma_k, \Delta_k)$, $k = 1, 2$. Let $\tilde{V}_k$ be the asymptotic variance of $\tilde{\beta}_k$, that is, $\tilde{V}_k = \lim_{n \to \infty} n \text{var}(\tilde{\beta}_k)$. Our main goal is to derive conditions on (Γ_1, Δ_1) and (Γ_2, Δ_2) such that $\tilde{V}_1$ and $\tilde{V}_2$ admit a deterministic ordering. It seems natural to require that the models are nested: $\Gamma_1 \supseteq \Gamma_2$ and $\Delta_1 \supseteq \Delta_2$.

Table 1 provides a list of uniform dominance relationships in some basic comparisons. We use the `R` convention to denote linear models: explanatory variables in the regression are joined by $+$, 1 stands for the intercept term, and $A : X$ stands for the treatment-covariate interaction. Using Table 1, we conjecture that in order for $\tilde{V}_1 \leqslant \tilde{V}_2$, the third condition in (6) needs to be modified which is verified in the next theorem.

Theorem 3.6. *Consider two estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$ obtained from solving (9) with $(\Gamma, \Delta) = (\Gamma_1, \Delta_1)$ and (Γ_2, Δ_2), respectively, and suppose $(\Gamma_1, \Delta_1) \neq (\Gamma_2, \Delta_2)$. Then $\tilde{\beta}_1$ uniformly dominates $\tilde{\beta}_2$ if*

$$\Gamma_1 \supseteq \Gamma_2, \ \ \Delta_1 \supseteq \Delta_2, \ \ \textit{and}\ \mathcal{U}(\Gamma_1) = \mathcal{U}(\Delta_1); \tag{10}$$

Proof. Let $d_\gamma = \gamma_1 - \gamma_2$ and $d_\delta = \delta_1 - \delta_2$. By applying (8) and Lemma 3.5 with the first model treated as the full model,

$$\begin{aligned}
\tilde{V}_2 - \tilde{V}_1 &= (V_2 - V_1) + (\tilde{V}_2 - V_2) - (\tilde{V}_1 - V_1) \\
&= \frac{1}{\pi^2(1-\pi)^2} E\left[(A - \pi)^2 \{ d_\gamma^T X + A d_\delta^T X \}^2 \right] + \delta_2^T \Sigma (2\delta_1 - \delta_2) - \delta_1^T \Sigma \delta_1 \\
&= \frac{E\left[A(A-\pi)^2 \{ d_\gamma^T X + d_\delta^T X \}^2 \right] + E\left[(1-A)(A-\pi)^2 \{ d_\gamma^T X \}^2 \right]}{\pi^2(1-\pi)^2} - d_\delta^T \Sigma d_\delta \\
&= \frac{1}{\pi} (d_\gamma + d_\delta)^T \Sigma (d_\gamma + d_\delta) + \frac{1}{1-\pi} d_\gamma^T \Sigma d_\gamma - d_\delta^T \Sigma d_\delta
\end{aligned}$$

$$= \frac{1}{\pi(1-\pi)}\left\{d_\gamma^T\Sigma d_\gamma + (1-\pi)^2 d_\delta^T\Sigma d_\delta + 2(1-\pi)d_\gamma^T\Sigma d_\delta\right\}$$
$$= \frac{1}{\pi(1-\pi)}\{d_\gamma + (1-\pi)d_\delta\}^T\Sigma\{d_\gamma + (1-\pi)d_\delta\} \geqslant 0, \tag{11}$$

which completes the proof. □

By verifying the conditions in (10), we have the following results concerning the examples in Section 2.

Corollary 3.7. *The ANHECOVA estimator is asymptotically more efficient than the ANOVA and ANCOVA estimators. There is no guaranteed variance ordering between ANOVA and ANCOVA.*

Corollary 3.8. *The lagged-dependent-variable regression estimator is more efficient than the difference-in-differences estimator.*

Results established in Theorems 3.4 and 3.6 can be used to provide practical efficiency gains. Consider the asymptotic variances V_k, $\tilde{V}_k$ of estimators $\hat{\beta}_k$, $\tilde{\beta}_k$, $k = 1, 2$. The relative efficiency of these estimators can be defined as follows

$$\mathrm{e}(\hat{\beta}_1, \hat{\beta}_2) = V_2/V_1, \quad \tilde{\mathrm{e}}(\tilde{\beta}_1, \tilde{\beta}_2) = \tilde{V}_2/\tilde{V}_1. \tag{12}$$

In Section 5 we present the results of the empirical study in which we analyse numerically the efficiency of ANECOVA versus ANOVA and ANCOVA.

Remark 2. The condition in (10) might be further weakened when $\pi = 1/2$. In particular, the difference in (11) is exactly 0 if in addition, $\Gamma_1 = \Gamma_2$ and $\pi = 1/2$. To show this, by differentiating (3) with respect to $\gamma_{\mathcal{U}(\Gamma_1)}$, we have $E\left[X_{\mathcal{U}(\Gamma_1)}\{Y - \alpha - \beta A - \gamma_k^T X - A(\delta_k^T X)\}\right] = 0,\ k = 1, 2$. By subtracting the two equations, we obtain

$$E\left[X_{\mathcal{U}(\Gamma_1)}(X^T d_\gamma + AX^T d_\delta)\right] = 0. \tag{13}$$

Because $\Gamma_1 = \Gamma_2$ and $\Delta_1 \supseteq \Delta_2$, we have $\gamma_{1j} = \gamma_{2j}$ and $\delta_{1j} = \delta_{2j}$ for $j \notin \mathcal{U}(\Gamma_1) = \mathcal{U}(\Delta_1)$. Thus $d_{\gamma,j} = d_{\delta,j} = 0$ when $j \notin \mathcal{U}(\Gamma_1)$. Together with equation (13), this shows that $\Sigma(d_\gamma + \pi d_\delta) = 0$. Therefore, if we have $\pi = 1/2$ in addition, the difference $\tilde{V}_2 - \tilde{V}_1 = 0$. In other words, when $\pi = 1/2$, adding or removing (more precisely, unrestricting or restricting) an interaction term AX_j when corresponding the main effect X_j is already present in the model does not change the asymptotic variance of $\tilde{\beta}$.

4. Variance ordering beyond linear regression

4.1. Stratified randomization experiments

One can establish the uniform dominance in case of more sophisticated randomization schemes. In particular, our derivations extend to stratified randomization experiments with units grouped into B strata which is in alignment with the results of [11]. The authors considered two asymptotic regimes with the number of strata B or with their sample sizes increasing as the total number of units n increases. Let S be a categorical variable with levels indicating whether unit i belongs to stratum s, $s \in \{1, 2, \dots, B\}$. In this setting, for the uniform dominance to hold, the treatment assignment should be completely randomized within strata with allocation probability equal across strata, and the estimator should be obtained using weighted regression including centred variables A, S, X, as well as interactions of A with S and A with X.

4.2. General assignment mechanisms

In contrast, the results on the uniform dominance do not usually extend to a more general assignment mechanism which depends on X with $\pi(x) = p(A = 1 \mid X = x)$. Within this framework, we can identify β_{ATE} for all Γ, Δ, but $\hat{\beta}$ in (2), (3) and (9) do not converge to β_{ATE}. More specifically, consider a population version of the least squares optimization problem in (3) and a general assignment mechanism which depends on X through $\pi(x)$. After setting up the first order conditions to find an estimate for β_{ATE} (cf., the proof of Lemma A.1), we can derive the following expression:

$$\beta = - \frac{\{E(AX) - \pi(x)E(X)\}^T \gamma + \{E(A^2X) - \pi(x)E(AX)\}^T \delta + \pi(x)E(Y) - E(AY)}{\{E(A^2) - \pi(x)E(A)\}}.$$

Due to the lack of $A \perp\!\!\!\perp X$, terms $E(AX)$ and $E(A^2X)$ cannot be further simplified which implies that $\hat{\beta}$ does not converge to β_{ATE}.

Nevertheless, there exist at least two ways to recover (1) within a setting with a general treatment assignment mechanism. First, one can make more stringent assumptions on the structure of covariates used in randomization and their relation to covariates used in the model-assisted regression in (2), (3) and (9), cf. [23] who proved some results on the variance ordering of covariate-adjusted estimators in the setting with discrete covariates used in the covariate-based randomization. Second, one can use weighted estimators [17, 18] to recover (1). In particular, let $\sigma^2(x) = \pi(x)\{1 - \pi(x)\}$, $A_c = A - \pi(x)$ and $X_c = X - E\{X|\pi(x)\}$. Consider the population version of the constrained linear regression problem for θ_w

$$\theta_w = (\alpha_w, \beta_w, \gamma_w, \delta_w) = \underset{\gamma_w \in \Gamma, \delta_w \in \Delta}{\arg\min}\, E\left[\frac{\{Y - \alpha_w - \beta_w A_c - \gamma_w^T X - A_c(\delta^T X_c)\}^2}{\sigma^2(X)}\right]. \tag{14}$$

By differentiating (14) with respect to θ_w and solving it for α_w, β_w, γ_w, δ_w we can conclude that β_w converges to β_{ATE}. In addition, by using almost identical arguments as in the proof of Lemma 3.3, see Appendix A.3, we derive that $n \times var(\hat{\beta}_w) \to E\left[\sigma^{-4}(X)\{A - \pi(X)\}^2 \epsilon_w^2\right]$ where $\epsilon_w = \epsilon_w(\theta_w) = Y - \alpha_w -$

$\beta_w A_c - \gamma^T X_c - A_c(\delta^T X_c)$. However, similarly as above, due to the lack of $A \perp\!\!\!\perp X$ or any further simplifications of the structure of X and/or its relation with A, the uniform dominance results as in Theorem 3.4 or Theorem 3.6 are not available. These conclusions are supported by the numerical results in Section 5.

4.3. Generalized linear models

The result in Lemma 3.1 is tightly related to the properties of orthogonal projections and does not carry over to a wider class of generalized linear model, even if the link function we use is collapsible [a link function is collapsible if including independent regressors does not change the population regression coefficients, see 6, 1, for more detials]. The most common collapsible link functions in this setting are identity and log function; the latter is the canonical link for Poisson regression and is frequently applied. Yet, if the link is collapsible but not the identity, including the treatment-covariate interaction may identify a different estimand. To exemplify this problem, let $\theta_P = (\alpha_P, \beta_P, \gamma_P, \delta_P)$ be coefficients within the framework of Poisson regression. Consider the population versions of the first order conditions to find estimates of α_P and β_P

$$E\left[Y - \exp\{\alpha_P + \beta_P A + \gamma_P^T X + A(\delta_P^T X)\}\right] = 0,$$
$$E\left(A\left[Y - \exp\{\alpha_P + \beta_P A + \gamma_P^T X + A(\delta_P^T X)\}\right]\right) = 0.$$

By solving these for $E(Y|A=1)$ and $E(A|A=0)$, we have

$$E(Y|A=1) = \exp(\alpha_P + \beta_P), E\left[\exp\{(\gamma_P + \delta_P)^T X\}\right],$$
$$E(Y|A=0) = \exp(\alpha_P)E\{\exp(\gamma_P^T X)\},$$

from where it follows that

$$\frac{E(Y|A=1)}{E(Y|A=0)} = \exp(\beta_P)\frac{E\left[\exp\{(\gamma_P + \delta_P)^T X\}\right]}{E\{\exp(\gamma_P^T X)\}},$$

that is, even after applying the canonical link function to the left-and the right-hand side of the expression above, coefficient β_P does not identify β_{ATE}, cf. results in Section 5. Thus, an attempt to seek the uniform dominance by comparing the asymptotic variance of $\hat{\beta}$ seems to be an ill-posed research question in this setting.

5. Simulation study

We carry out numerical simulation study to verify our theoretical developments in previous sections and explore scenarios under which the uniform dominance does not hold (see discussion in Section 6). We consider scenarios with potential outcomes generated from normal and Poisson distribution. In all scenarios, the covariate is centred and normally distributed $X = X_{nc} - \bar{X}$, $X_{nc} \sim N(2, 1)$

whereas $\bar{X}$ is a sample mean of observations; the observed outcome is $Y = AY(1) + (1 - A)Y(0)$. For linear regression, errors are normally distributed $e_1, e_0 \sim N(0, 1)$. We consider the sample size of $n = 1000$. Below we describe simulation scenarios.

Scenario 1 Treated outcomes: $Y(1) \sim 5 + 2.5AX + e_1$, untreated outcomes: $Y(0) \sim 3 + X + e_0$, the treatment assigned using a Bernoulli trial $A \sim Bernoulli(\pi)$, $\beta_{ATE} = 2$, estimation method: linear regression.

Scenario 2 Treated outcomes: $Y(1) \sim Poisson(\mu_1)$, untreated outcomes: $Y(0) \sim Poisson(\mu_0)$, $\mu_1 = \exp(5.5 + 0.7A + 0.5X + 0.8XA)$, $\mu_0 = \exp(4.5 + 0.6A + 0.4X + 0.7XA)$, the treatment assigned using a Bernoulli trial $A \sim Bernoulli(\pi)$, $\beta_{ATE} = 1.2759$, estimation method: Poisson regression with the canonical link function.

Scenario 3 Treated outcomes: $Y(1) \sim 7 + X + e_1$, untreated outcomes: $Y(0) \sim 2 - X + X^2 + e_0$, the treatment assigned using a general assignment mechanism $A \sim Bernoulli(\pi(X))$, $\pi(X) = \exp(4 - 2X)/(1 + \exp(4 - 2X))$, $\beta_{ATE} = 4$, estimation method: linear regression with weights $w(X) = (\pi(X)(1 - \pi(X))^{-1}$.

To study the performance of the estimators of β_{ATE}, we calculated an average bias and a standard deviation over 1000 Monte Carlo replications. On top of that, for the estimators under Scenario 1, we calculated the relative efficiency defined in (12). Finally, under the Poisson log model, true values of β_{ATE} was approximated by the difference of the large sample average ($n = 10^7$) of potential outcomes. Figure 1 shows results of simulations under Scenario 1 over different values of π. The estimates of β_{ATE} are almost unbiased assuming any model (top-right panel). On the other hand, the standard deviation of $\hat{\beta}$ is the smallest for the model with both X and AX, nevertheless adjusting for covariate only is not beneficial for higher values of π. The relative efficiency with respect to the ANECOVA estimator is uniformly the lowest for ANOVA estimator (between 0.36 and 0.42), and the highest for the model with both A and AX (between 0.53 and 0.61). Nevertheless, the sample size gains from using ANECOVA are still substantial. In the bottom panel of Figure 1 we plotted the inverse of the relative efficiency which provides a rough estimate of the multiplicative factor by which we would need to multiply the sample sizes of ANOVA or ANCOVA to obtain the efficiency of ANECOVA. For example, to obtain the same efficiency as the ANECOVA fitted using 5000 observations, we would need to collect 12942 observations for the ANOVA fit and 11656 for the ANCOVA fit. When it comes to the results under the Poisson log model in Figure 2, $\hat{\beta}$ suffers from a significant bias as it does not recover the estimand of interest, cf. discussion in Section 4.3.

Table 2 displays numerical performance of $\hat{\beta}$ under Scenarios 3 for which our theory does not hold. We can see that the estimators do not suffer from the excessive bias, which is an indication that the weighted regression estimators recover β_{ATE}. At the same time, we can see that the uniform dominance does not hold which is in alignment with the discussion in Section 4.2.

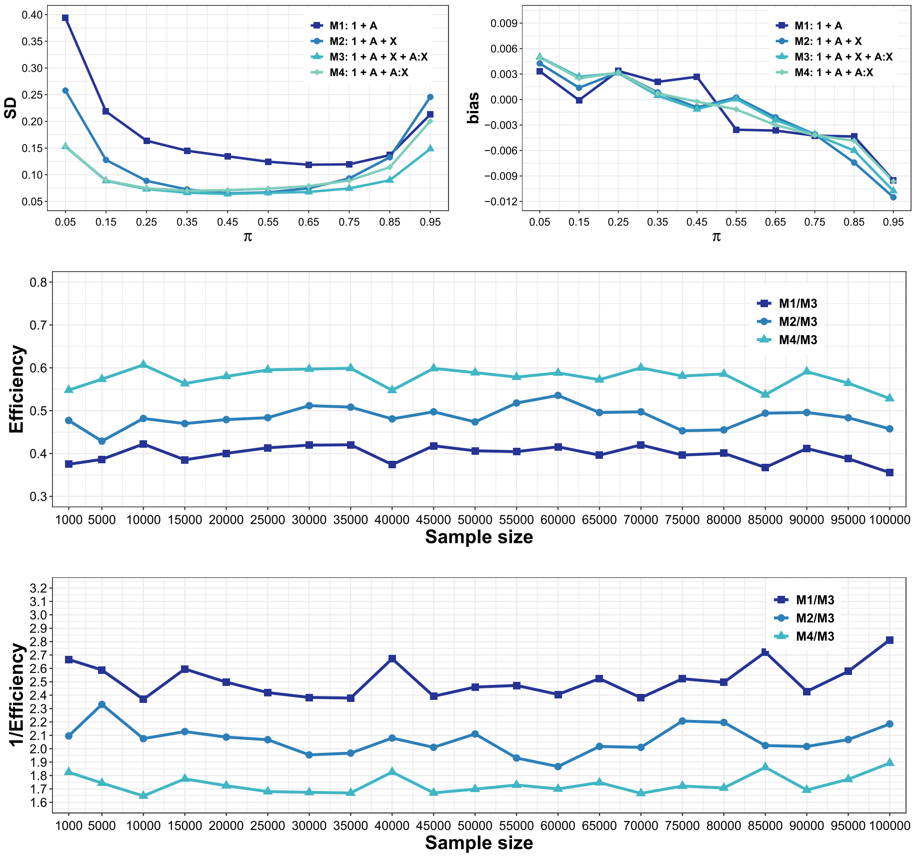

Fig 1. *Performance of estimators of β_{ATE} over different values of π under Scenario 1 (top panels), relative efficiency of estimators over different sample sizes (middle panel), and the inverse of relative efficiency (bottom panel), SD: standard deviation.*

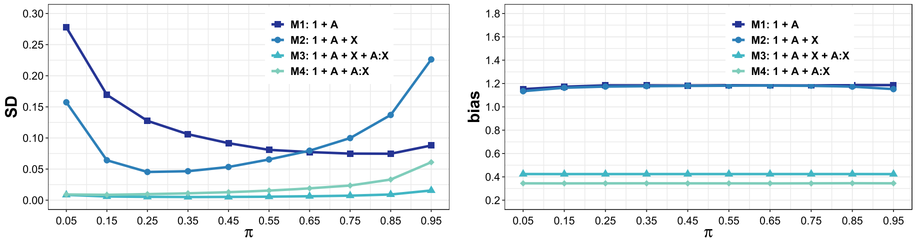

Fig 2. *Performance of estimators of β_{ATE} over different values of π under Scenario 2, SD: standard deviation.*

Table 2
Performance of $\hat{\beta}$. SD, standard deviation.

Scenario 3		
$\beta_{\text{ATE}} = 4$, $logit\{\pi(X)\} = 4 - 2X$		
Model	bias	SD
$\sim 1 + A$	-0.007	0.159
$\sim 1 + A + X$	0.004	0.232
$\sim 1 + A + X + A : X$	0.037	0.185

6. Discussion

Linear regression model is still widely used to estimate the average treatment effect in hope to increase the precision of the estimator. We re-established and generalized previous results on linear-regression adjusted estimators under possible model misspecification by providing a simplified and more accessible proof of uniform dominance. Yet, our proof has a geometric element that exploits the linearity of the regression adjustment, and this cannot be extended to other settings. Thus, the phenomenon of the efficiency gain seems to be limited to the estimation problems which fit into the linear framework and to the treatment assignment mechanisms which do not depend on X.

Appendix A: Additional proofs

A.1. Proof of Lemma 3.1

Proof. Observe that α and β are always unrestricted in (3). By taking partial derivatives with respect to α and β, we obtain

$$E\{Y - \alpha - \beta A - \gamma^T X - A(\delta^T X)\} = 0 \text{ and } E\{A(Y - \alpha - \beta A - \gamma^T X - A(\delta^T X))\} = 0.$$

By multiplying the first equation by π and subtracting the second equation, we obtain

$$\begin{aligned}&\pi E(Y) - E(AY) + \{E(A^2) - \pi E(A)\}\beta \\ &+ \{E(AX) - \pi E(X)\}^T \gamma + \{E(A^2 X) - \pi E(AX)\}^T \delta = 0.\end{aligned}$$

Finally, by using the assumption that the treatment is randomized i.e. $A \perp\!\!\!\perp X$ and $E(A) = E(A^2) = \pi$, we find that

$$\begin{aligned}\beta &= -\frac{\{\pi E(Y) - E(AY)\} + (\pi - \pi^2)\delta^T E(X)}{\pi - \pi^2} \\ &= -\frac{\pi\{\pi E(Y \mid A = 1) + (1 - \pi)E(Y \mid A = 0)\} - \pi E(Y \mid A = 1)}{\pi - \pi^2} - \delta^T E(X) \\ &= \beta_{\text{ATE}} - \delta^T E(X)\end{aligned}$$

as desired. □

A.2. Proof of Lemma 3.2

Proof. To be able to use the results of [21], we need to check that our assumptions are sufficient to evoke regularity conditions cited by the author. Since Y and Z have bounded forth moments, there exist η and H such that $E(|\epsilon(\theta)^2|^{\eta+1}) < H$ and $E(|Z_j Z_k|^{\eta+1}) < H$, $j, k = 1, \ldots, p$. By Hölder inequality, this implies that $E(\epsilon(\theta)^2 |Z_j Z_k|^{\eta+1}) < H$ is also uniformly bounded. Furthermore, we assumed that $E(ZZ^t)$ is positive definite, that is, $E(ZZ^t)$ is nonsingular and $det E(ZZ^t) > \eta > 0$. The same is valid for $E(\epsilon(\theta)^2 ZZ^t)$. Thus, Assumptions 2 and 3 of [21] are satisfied and we can use the same steps as the author to prove the consistency and the asymptotic normality of $\hat{\theta}$. □

A.3. Proof of Lemma 3.3

Proof. First, consider the unrestricted case where $\Gamma = \Delta = \mathbb{R}^p$. To use Lemma 3.2, we simply need to compute $E(ZZ^T)$ and $E(ZZ^T \epsilon^2)$ for $Z = (1, A, X^T, AX^T)^T$. Employing $A \perp\!\!\!\perp X$, $E(Z) = 0$, $A^2 = A$, and $E(XX^T) = \Sigma$, we have

$$E(ZZ^T) = \begin{pmatrix} 1 & \pi & 0 & 0 \\ \pi & \pi & 0 & 0 \\ 0 & 0 & \Sigma & \pi\Sigma \\ 0 & 0 & \pi\Sigma & \pi\Sigma \end{pmatrix}.$$

Using properties of block diagonal matrices, it follows that

$$n\mathrm{var}(\hat{\beta}) \to \left[\begin{pmatrix} 1 & \pi \\ \pi & \pi \end{pmatrix}^{-1} \begin{pmatrix} E(\epsilon^2) & E(A\epsilon^2) \\ E(A\epsilon^2) & E(A\epsilon^2) \end{pmatrix} \begin{pmatrix} 1 & \pi \\ \pi & \pi \end{pmatrix}^{-1} \right]_{(22)} = \frac{E\{(A-\pi)^2\epsilon^2\}}{\pi^2(1-\pi)^2}.$$

Here, $[\cdot]_{(22)}$ means the entry on the second row and second column of the matrix.

If some dimensions of Γ or Δ are singletons, we can simply remove the corresponding entries in Z. By a similar calculation, the same formula holds and the asymptotic variance of $\hat{\beta}$ only differs in the regression error ϵ, which depends on $\theta = \theta(\Gamma, \Delta)$. □

A.4. Proof of Lemma 3.5

Proof. We fix Γ and Δ and suppress the dependence of $\hat{\beta}$ and $\tilde{\beta}$ on (Γ, Δ). We decompose $\tilde{\beta}$ as

$$\tilde{\beta} = \hat{\beta} + \hat{\delta}^T \bar{X} = \hat{\beta} + \delta_s^T \bar{X} + (\hat{\delta} - \delta_s)^T \bar{X}.$$

Due to the assumption $E(X) = 0$, we have $\hat{\delta} - \delta_s = O_p(n^{-1/2})$ and $\bar{X} = O_p(n^{-1/2})$. Hence, the last term on the right hand side is negligible. Therefore,

$$n\left\{\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta})\right\} \to n\mathrm{var}(\delta_s^T \bar{X}) + 2n\mathrm{cov}(\hat{\beta}, \delta_s^T \bar{X}).$$

Let Z be the unrestricted variables in the linear regression. By applying the sandwich variance formula for the following set of estimating equations

$$\begin{aligned} E(X-\mu) &= 0, \\ E[Z\epsilon(\theta)] &= 0, \end{aligned}$$

we obtain

$$\begin{aligned} & n\mathrm{cov}\left(\hat{\theta}, \bar{X}\right) \\ & \to \left\{ \begin{pmatrix} -I & 0 \\ 0 & -E(ZZ^T) \end{pmatrix}^{-1} \begin{pmatrix} \Sigma & E(XZ^T\epsilon) \\ E(ZX^T\epsilon) & E(ZZ^T\epsilon^2) \end{pmatrix} \begin{pmatrix} -I & 0 \\ 0 & -E(ZZ^T) \end{pmatrix}^{-T} \right\}_{(2)} \\ & = \left\{E(ZZ^T)\right\}^{-1} E(ZX^T\epsilon), \end{aligned}$$

where ϵ is $\epsilon(\theta) = \epsilon(\theta_f) + (\gamma_f - \gamma_s)^T X + A(\delta_f - \delta_s)^T X$. It follows that

$$\begin{aligned} n\mathrm{cov}\left(\hat{\beta}, \bar{X}\right) &\to \left[\left\{E(ZZ^T)\right\}^{-1} E(ZX^T\epsilon)\right]_{(22)} \\ &= \left[\begin{pmatrix} 1 & \pi & 0 \\ \pi & \pi & 0 \\ 0 & 0 & * \end{pmatrix}^{-1} \begin{pmatrix} (\gamma_f - \gamma_s)^T\Sigma + \pi(\delta_f - \delta_s)^T\Sigma \\ \pi(\gamma_f - \gamma_s)^T\Sigma + \pi(\delta_f - \delta_s)^T\Sigma \\ * \end{pmatrix} \right]_{(22)} \\ &= (\delta_f - \delta_s)^T\Sigma, \end{aligned}$$

where $*$ represent some unspecified matrices that are not important for deriving the quantities of interest. Therefore,

$$\begin{aligned} n\left\{\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta})\right\} &\to n\mathrm{var}(\delta_s^T\bar{X}) + 2n\mathrm{cov}(\hat{\beta}, \delta_s^T\bar{X}) \\ &= \delta_s^T\Sigma\delta_s + 2(\delta_f - \delta_s)^T\Sigma\delta_s \\ &= \delta_s^T\Sigma(2\delta_f - \delta_s). \end{aligned}$$

□

A.5. Variance orderings in Table 1

In this section, we provide an alternative, simpler proof to derive variance ordering in Table 1. These conclusions can be derived from Theorems 3.4 and 3.6. Let $\theta_f = \theta_0(\mathbb{R}^p, \mathbb{R}^p)$ denote the full model parameters. From the first order condition, we have

$$\gamma_f = \Sigma^{-1}E(XY \mid A=0), \quad \delta_f = \Sigma^{-1}\{E(XY \mid A=1) - E(XY \mid A=0)\}. \tag{15}$$

Let $V_f = \lim_{n\to\infty} n\mathrm{var}(\hat{\beta}_f)$ be the asymptotic variance of $\hat{\beta}_f$.

(a) We compare variances of $\hat{\beta}$ in Model 1 with $Z_1 = (1, A, X^T)^T$ and in Model 2 with $Z_1 = (1, A)^T$. Consider $\Gamma_1 = \mathbb{R}^p$, $\Delta_1 = 0$, $\Gamma_2 = 0$ and $\Delta_2 = 0$, that is, $\hat{\beta}_1$ is the ANCOVA and $\hat{\beta}_2$ is the ANOVA estimator in Example 1 in the main document. In this case, only the third condition $\mathcal{U}(\Delta_1) \supseteq \mathcal{U}(\Gamma_1)$ in Theorem 1 is not satisfied. We show that $V_1 > V_2$ when $\pi E(XY \mid A = 0) = (\pi - 1)E(XY \mid A = 1)$, $\pi \neq \frac{1}{2}$, and $E(XY \mid A = 1) \neq E(XY \mid A = 0)$.

By definition, $\delta_1 = \delta_2 = \gamma_2 = 0$, and $\gamma_1 = \Sigma^{-1}E(XY) = \gamma_f + \delta_f \pi$. Then, by applying derivations in Theorem 1, we have

$$
\begin{aligned}
\pi^2(1-\pi)^2(V_2 - V_f) &= E\big[(A-\pi)^2\{\gamma_f^T X + A\delta_f^T X\}^2\big], \\
\pi^2(1-\pi)^2(V_1 - V_f) &= E\big[(A-\pi)^2\{(\gamma_f - \gamma_1)^T X + A\delta_f^T X\}^2\big], \\
&= E\big[(A-\pi)^2(\gamma_1^T X)^2\big] + E\big[(A-\pi)^2(\gamma_f^T X + A\delta_f^T X)^2\big] \\
&\quad - 2E\big[(A-\pi)^2\gamma_1^T X(\gamma_f^T X + A\delta_f^T X)\big].
\end{aligned}
$$

Hence, using the fact that $A \perp\!\!\!\perp X$,

$$
\begin{aligned}
&\pi^2(1-\pi)^2(V_1 - V_2) \\
=&E\big[(A-\pi)^2(\gamma_1^T X)^2\big] - 2E\big[(A-\pi)^2\gamma_1^T X(\gamma_f^T X + A\delta_f^T X)\big] \\
=&E\big[(A-\pi)^2(\gamma_1^T X)(\gamma_1 - 2\gamma_f - 2A\delta_f)^T X\big] \\
=&E\big[(A-\pi)^2(\gamma_f^T X + \pi\delta_f^T X)(\pi\delta_f^T X - \gamma_f^T X - 2A\delta_f^T X)\big] \\
=&E\big[A(1-\pi)^2(\gamma_f^T X + \pi\delta_f^T X)(\pi\delta_f^T X - \gamma_f^T X - 2\delta_f^T X)\big] \\
&\quad + E\big[(1-A)\pi^2(\gamma_f^T X + \pi\delta_f^T X)(-\gamma_f^T X + \pi\delta_f^T X)\big] \\
=&\pi(1-\pi)^2(\gamma_f + \pi\delta_f)^T\Sigma(\pi\delta_f - \gamma_f - 2\delta_f) + (1-\pi)\pi^2(\gamma_f + \pi\delta_f)^T\Sigma(\pi\delta_f - \gamma_f) \\
=&\pi(1-\pi)(\gamma_f + \pi\delta_f)^T\Sigma\{(3\pi - 2)\delta_f - \gamma_f\}.
\end{aligned}
$$

When $\gamma_f = (\pi - 1)\delta_f, \pi \neq \frac{1}{2}$ and $\delta_f \neq 0$, we have

$$
\pi^2(1-\pi)^2(V_1 - V_2) = \pi(1-\pi)(2\pi - 1)^2 E(\delta_f^2) > 0. \tag{16}
$$

Under this scenario, $(\tilde{V}_1, \tilde{V}_2) = (V_1, V_2)$. We can thus proceed in the same way to prove $\tilde{V}_1 > \tilde{V}_2$.

(b) We compare variances of $\hat{\beta}$ in Model 1 with $Z_1 = (1, A, AX^T)^T$ and in Model 2 with $Z_1 = (1, A)^T$. First we shall prove $V_2 \geqslant V_1$. Consider $\Gamma_1 = 0$, $\Delta_1 = \mathbb{R}^p$, $\Gamma_2 = 0$ and $\Delta_2 = 0$. In this case we have $\gamma_1 = \gamma_2 = \delta_2 = 0$, $\delta_1 = \Sigma^{-1}E(XY|A = 1)$ and

$$
\pi^2(1-\pi)^2(V_2 - V_1) = E\left\{(A-\pi)^2 2\epsilon_1 A\delta_1^T X\right\} + E\left\{(A-\pi)^2(A\delta_1^T X)^2\right\} \geqslant 0. \tag{17}
$$

The first term on the right hand side in (17) is 0 by applying the sufficient condition in Theorem 1, $A \perp\!\!\!\perp X$ and $A^2 = A$.

Now we prove that $\tilde{V}_1 > \tilde{V}_2$ for some cases. Let γ_f and δ_f be as defined in equation (15), $\gamma_1 = \gamma_2 = \delta_2 = 0$ and $\delta_1 = \Sigma^{-1}E(XY|A = 1) = \delta_f + \gamma_f \neq 0$. In addition, let $\Omega_l = E(XY|A = l)$ where $l = 0, 1$. When $\Omega_0 = -1/2\Omega_1$, we have

$$
\tilde{V}_1 - \tilde{V}_2 \to V_1 + \delta_1^T\Sigma(2\delta_f - \delta_1) - V_2
$$

$$
\begin{aligned}
&= \frac{E\{(A-\pi)^2(A\delta_1^T X)^2\}}{\pi^2(1-\pi)^2} + \delta_1^T\Sigma(2\delta_f - \delta_1)\\
&= \frac{E\{(\gamma_f^T X + \delta_f^T X)^2\}}{\pi} + (\delta_f + \gamma_f)^T\Sigma(\delta_f - \gamma_f)\\
&= \frac{(\delta_f + \gamma_f)^T\Sigma(\delta_f + \gamma_f) + \pi(\delta_f + \gamma_f)^T\Sigma(\delta_f - \gamma_f)}{\pi}\\
&= \frac{\Omega_1^T\Sigma^{-1}\Omega_1 + \pi\Omega_1^T\Sigma^{-1}(\Omega_1 - 2\Omega_0)}{\pi} = \frac{\Omega_1^T\Sigma^{-1}\Omega_1 + 2\pi\Omega_1^T\Sigma^{-1}\Omega_1}{\pi} > 0.
\end{aligned}
\tag{18}
$$

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

The first author gratefully acknowledge support from the Swiss National Science Foundation, project P2GEP2-195898.

Supplementary Material

R code

(doi: 10.1214/24-EJS2233SUPP; .zip). Supplement published online [13] includes the R code to reproduce the simulations in Section 5.

References

[1] Daniel, R., Zhang, J., and Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom. J.*, 63(3):528–557. MR4226593

[2] Ding, P. and Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Polit. Anal.*, 27(4):605–615.

[3] Fisher, R. A. (1928). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 4th edition.

[4] Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.*, 2(1):176–196. MR2415599

[5] Freedman, D. A. (2008b). On regression adjustments to experimental data. *Adv. Appl. Math.*, 40(2):180–193. MR2388610

[6] Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Stat. Sci.*, 14(1):29–46.

[7] Guo, K. and Basse, G. (2023). The generalized Oaxaca-blinder estimator. *J. Am. Stat. Assoc.*, 118(541):524–536. MR4571139

[8] Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, 1st edition. MR3309951

[9] Koch, G. G., Tangen, C. M., Jung, J.-W., and Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Stat. Med.*, 17(15-16):1863–1892.

[10] Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.*, 7(1):295–318. MR3086420

[11] Liu, H. and Yang, Y. (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107(4):935–948. MR4186497

[12] Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Ann. Agric. Sci.*, 10:1–51. (Translated to English and edited by Dabrowska, D. M. and Speed, T. P. (1990) *Stat. Sci.*, 5(4):465–480). MR1092986

[13] Reluga, K., Ye T., and Zhao Q. (2024). Supplement to "A unified analysis of regression adjustment in randomized experiments". DOI: 10.1214/24-EJS2233SUPP.

[14] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66(5):688–701.

[15] Rubin, D. B. and van der Laan, M. J. (2011). Targeted ANCOVA estimator in RCTs. In *Targeted Learning*, 201–215. Springer, New York, NY, 1st edition. MR2867124

[16] Schochet, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *J. Stat. Plan. Inference*, 140(1):246–259. MR2568136

[17] Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Stat. Soc., A: Stat. Soc.*, 174(2):369–386. MR2898850

[18] Tao, Y. and Fu, H. (2019). Doubly robust estimation of the weighted average treatment effect for a target population. *Stat. Med.*, 38(3):315–325. MR3897162

[19] Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Stat. Med.*, 27(23):4658–4677. MR2528575

[20] Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., and Rosenblum, M. (2023). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *J. Am. Stat. Assoc.*, 118(542):1152–1163. MR4595484

[21] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

MR0575027

[22] Yang, L. and Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *Am. Stat.*, 55(4):314–321. MR1943328

[23] Ye, T., Shao, J., and Zhao, Q. (2023). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *J. Am. Stat. Assoc.*, 118(544):2370-2382. MR4681589