Trading-off Bias and Variance in Stratified Experiments and in Staggered Adoption Designs, Under a Boundedness Condition on the Magnitude of the Treatment Effect

Clément de Chaisemartin

Trading-off Bias and Variance in Stratified Experiments and in Staggered Adoption Designs, Under a Boundedness Condition on the Magnitude of the Treatment Effect*

Clément de Chaisemartin †

First version: May 18th, 2021
This version: March 18, 2022

Abstract

I consider estimation of the average treatment effect (ATE), in a population composed of G groups, when one has unbiased and uncorrelated estimators of each group's conditional average treatment effect (CATE). These conditions are met in stratified randomized experiments. I assume that the outcome is homoscedastic, and that each CATE is bounded in absolute value by B standard deviations of the outcome, for some known B. I derive, across all linear combinations of the CATEs' estimators, the estimator of the ATE with the lowest worst-case mean-squared error. This minimax-linear estimator assigns a weight equal to group g's share in the population to the most precisely estimated CATEs, and a weight proportional to one over the estimator's variance to the least precisely estimated CATEs. I also derive the minimax-linear estimator when the CATEs' estimators are positively correlated, a condition that may be met by differences-in-differences estimators in staggered adoption designs.

Keywords: bias-variance trade-off, average treatment effect, mean-squared error, minimax-linear estimator, bounded normal mean model, stratified randomized experiments, differences-in-differences, staggered adoption designs, shrinkage.

JEL Codes: C21, C23

1 Introduction

I consider the estimation of the average treatment effect (ATE), in a population that can be divided into G groups. I assume that one has unbiased estimators of the conditional average treatment effect (CATE) in each group, with heterogeneous levels of statistical precision.

*This paper was previously circulated under the following title: "The Minimax Estimator of the Average Treatment Effect, among Linear Combinations of Estimators of Bounded Conditional Average Treatment Effects". I am very grateful to Timothy Armstrong, Xavier D'Haultfœuille, and Michal Kolesár for their helpful comments.

†de Chaisemartin: Sciences Po, Economics Department (email: clement.dechaisemartin@sciencespo.fr)

This situation often arises in stratified randomized controlled trials (RCTs), where the treatment probability may vary across strata, thus leading to more (resp. less) precisely estimated CATEs in strata where the treatment probability is close to (resp. far from) 1/2. In such instances, applied researchers typically regress the outcome on the treatment and strata fixed effects. This estimator downweights the least balanced strata, and it often has a lower variance than the unbiased propensity score estimator. But it is also biased if the CATEs vary across strata. Despite its ubiquitousness, it is unclear whether the fixed effects estimator still dominates the unbiased one when bias is accounted for. Another instance where one has unbiased estimators of CATEs with heterogeneous levels of statistical precision are staggered adoption designs, where units adopt a treatment at heterogeneous dates. Then, the long-run treatment effects arising at the end of the panel are often less precisely estimated than the short-run ones, because few units are still untreated and can be used as controls at the end of the panel. Then again, one may want to downweight those long-run effects to reduce variance, but doing so may create bias if short- and long-run effects differ.

To study this bias-variance trade-off, I derive the linear combination of CATEs' estimators with the lowest worst-case mean-squared error, the minimax-linear estimator.

To do so, I first assume that the estimators of the CATEs are uncorrelated, that the outcome is homoscedastic, and that each CATE is bounded in absolute value by B standard deviations of the outcome, for some known constant B. Under those assumptions, I show that the minimax-linear estimator is a weighted sum of the CATEs' estimators, with positive weights that sum to less than 1. The most precisely estimated CATEs receive a weight equal to the share of the population their group accounts for. The least precisely estimated CATEs receive a weight proportional to one over the estimator's variance, and shrunk towards zero. Given B, the optimal estimator is feasible: it only depends on known quantities.

The assumptions outlined in the previous paragraph may be applicable to stratified RCTs. There, CATE estimators are unbiased and uncorrelated by design. Moreover, normalizing by the outcome's standard deviation is a common practice, and applied researchers often have a good sense, based on the literature, of the effect size that a given intervention may produce. In such instances, they may be able to come up with plausible values of B. When the literature is silent as to which values of B are plausible upper bounds for the CATEs, researchers can conduct a sensitivity analysis by varying B. Finally, my approach also assumes homoscedasticity, but I show that with heteroscedasticity, the minimax-linear estimator still has a lower worst-case MSE than the unbiased one whenever the treated outcome's variance is larger than that of the untreated outcome. When the opposite holds, I show that the minimax-linear estimator still has a lower worst-case MSE, provided the ratio of the treated and untreated outcomes variances is not below a bound that can be readily computed from the data. In my application in Section 5, this bound is equal to 0.06, so the treated outcome's variance would have to be more than 94% smaller than the untreated outcome's for the

minimax-linear estimator to have a higher worst-case MSE than the unbiased one.

In stratified RCTs, my result implies that the minimax-linear estimator is actually "in between" the unbiased and fixed effects estimators. The unbiased estimator assigns to each CATE a weight equal to group g's share in the population, like the minimax-linear estimator does to the most precisely estimated CATEs, while the fixed effects estimator assigns to each CATE a weight proportional to one over its variance, like the minimax-linear estimator does to the least precisely estimated CATEs.

Then, I replace the assumption that the estimators are uncorrelated by the assumption that they are positively correlated, with covariances known up to the outcome's variance. Under this weaker assumption, I show that the minimax-linear estimator, across all linear combinations of the estimators with positive weights, is the solution of an easy-to-numerically-solve minimization problem, and is still feasible.

This second set-up may be applicable to the differences-in-differences (DID) estimators for staggered adoption designs proposed by Sun & Abraham (2020), Callaway & Sant'Anna (2020), and de Chaisemartin & D'Haultfœuille (2020). There, the target parameter is the average treatment effect on the treated (ATT), and the CATEs are the average treatment effect at period t among units that started receiving the treatment at period k. If potential outcomes without treatment are iid, across units and over time, and if the treatment effects and treatments are non-stochastic, the unbiased estimators of those CATEs proposed by Sun & Abraham (2020), Callaway & Sant'Anna (2020), and de Chaisemartin & D'Haultfœuille (2020) are positively correlated, with covariances known up to the outcome's variance. I consider a numerical example with 50 units (e.g.: the 50 US states) and five periods, where 10 units respectively become treated at periods 2, 3, 4, and 5, and 10 units remain untreated. I set $B = 0.75$, meaning that CATEs should all be lower than 75% of the outcome's standard deviation. I find that the minimax-linear estimator downweights some CATEs at period 5. Those are the least precisely estimated CATEs, because there are few units that are still untreated and can be used as controls at that period. If outcomes are indeed iid, the minimax-linear estimator has substantially lower variance and worst-case MSE than the unbiased one. The minimax-linear estimator still has substantially lower variance and worst-case MSE if each unit's outcome follows an AR(1), provided the AR(1) coefficient is not too close to 1.

Finally, I consider a number of extensions. I show that assuming that CATEs are positive and bounded, rather than just assuming that they are bounded, does not change the minimax-linear estimator. I also characterize the minimax-linear estimator with heteroscedasticity. Its formula remains the same as under homoscedasticity, but it is not feasible anymore: it depends on the estimators' variances, that are typically unknown. A feasible estimator can easily be computed, by replacing those variances by their estimators.

I use my results to revisit Behaghel et al. (2017), who use a stratified RCT to measure the effect of a boarding school for disadvantaged students in France on students' math test scores.

Their RCT has 363 students and 14 strata. The treatment probability varies substantially across strata, but no stratum has a treatment probability, say, lower than 0.1 or higher than 0.9. Based on a review of several papers that have estimated CATEs of similar interventions (see Curto & Fryer Jr, 2014, Dobbie & Fryer Jr, 2011, Angrist et al., 2010, and Abdulkadiroğlu et al., 2011), I argue in Section 5 that 50% of a standard deviation of students' test scores is a plausible upper bound for CATEs in this application. With that value of B, the robust standard error of the minimax-linear estimator is 9% smaller than that of the unbiased estimator, and its worst-case MSE under homoscedasticity is 6% smaller. In the subsample of male students, its robust standard error and worst-case MSE are 15% and 12% smaller.

This paper is related to the pioneering work of Crump et al. (2009), who consider matching estimators when the treatment probability can be close to zero or one conditional on some values of the covariates. My estimator can also be used in such instances, provided the covariates take a finite number of values. Crump et al. (2009) propose to redefine the target parameter as the ATE in the subpopulation whose ATE can be estimated most precisely. In other words, Crump et al. (2009) minimize variance, but do not control bias with respect to the original target parameter (the ATE in the full population). Instead, I propose to use the minimax-linear estimator. This avoids changing the target parameter. But this requires taking a stand on how large CATEs may be, by specifying an upper bound B for them. Below, I show that when $B \to +\infty$, meaning that one does not restrict the magnitude of the CATEs, the minimax-linear estimator converges towards the unbiased one, and the trade-off between bias and variance becomes trivial. Finally, note that the main result in Crump et al. (2009) is derived under homoscedasticity, like the main results in this paper.

This paper is also related to a vast literature in statistics and econometrics, that has studied linear- and affine-minimax estimators. The setting I consider can be cast as a bounded normal mean model, where realizations of G normally distributed random variables are used to estimate a linear combination of their means, which are assumed to be bounded.[1] Donoho (1994), who considers a more general setup than the bounded normal mean model, characterizes the risk of the affine-minimax estimator, and shows that it cannot be more than 25% larger than that of the minimax estimator. Armstrong & Kolesár (2018) consider a similar set-up as Donoho (1994), and show how to construct optimal confidence intervals. My paper makes the following contributions. First, it is the first to apply the bounded normal mean model to derive the minimax-linear estimator of the ATE, under boundedness conditions on the CATEs. Thus, it complements a growing econometrics literature that has applied the set-up in Donoho (1994) to other estimation problems.[2] Second, the closed-form expression of the minimax-linear estimator that I derive when CATEs are uncorrelated is, to my knowledge,

[1] I do not assume that CATEs' estimators are normally distributed, but as noted by Armstrong & Kolesár (2021*a*), this distributional assumption is not of essence to derive the minimax-linear estimator.

[2] Examples include: ATE estimation under uncounfoundedness when the mean outcome conditional on the covariates is a Lipschitz function with a bounded Lipschitz constant (see Armstrong & Kolesár, 2021*a*); sensitivity analysis in locally misspecified GMM models (see Armstrong & Kolesár, 2021*b*); DID estimation with

new. In their Section 4.2., Ibragimov & Khas'minskii (1985) derive a closed-form expression of the minimax-linear estimator when $G = 1$. My result generalizes theirs to the case where $G > 1$. The question this paper is concerned with, namely trading-off bias and variance when averaging several CATE estimators, only arises when $G > 1$, so this extension is of essence to the problem at hand.[3] This closed-form expression unveils a connection between the minimax-linear estimator and two commonly-used estimators, and may thus improve our understanding of the former. Third, bounding the CATEs by B standard deviations of the outcome, rather than by a constant B, is natural given that applied researchers often normalize their outcome by its standard deviation, and allows me to propose minimax-linear estimators that are feasible given B. In particular, computing those estimators does not require estimating the outcome's variance in a first step, unlike other estimators that have been proposed in this literature (see e.g. Armstrong & Kolesár, 2021*a*). Fourth, the realization that with correlated CATE estimators, the minimax-linear estimator with positive weights can be approximated via a simple numerical method is also, to my knowledge, new.

The remainder of the paper is organized as follows. Section 2 presents the paper's main results. Section 3 presents some extensions. Section 4 presents some numerical examples. Section 5 presents an empirical application.

2 Main results: feasible minimax-linear estimators

Throughout the paper, I consider the following set-up.

Definition 2.1 *(Set-up) One is interested in estimating an unknown parameter τ, equal to a weighted average of unknown parameters $(\tau_g)_{1\leq g\leq G}$, with weights $(p_g)_{1\leq g\leq G}$ that are known, positive, and sum to 1:*

$$\tau = \sum_{g=1}^{G} p_g \tau_g. \tag{2.1}$$

One observes random variables $(\widehat{\tau}_g)_{1\leq g\leq G}$ such that $E(\widehat{\tau}_g) = \tau_g$ for all g.

In the applications I consider, τ is an ATE in a population that can be divided into G groups, τ_g is the CATE in group g, p_g is the share of the population group g accounts for, and $\widehat{\tau}_g$ is an unbiased estimator of τ_g. The more abstract set-up in Definition 2.1 is useful to connect this paper with the bounded normal mean model (see, e.g., Donoho, 1994).

bounded departures from parallel trends (see Rambachan & Roth, 2019); estimation in regression discontinuity designs with bounded second derivatives of the mean of the potential outcomes conditional on the running variable (see Armstrong & Kolesár, 2018, Imbens & Wager, 2019, and Noack & Rothe, 2019).

[3]Relatedly, Berry (1990) studies the minimax estimator in a multivariate bounded normal mean model. However, the normality assumption is of essence to derive the minimax estimator, so his closed-form expression differs from mine. Vidakovic (1993) studies the minimax-affine estimator in a multivariate bounded normal mean model, assuming that the ℓ_2 norm of the vector of means are bounded. Instead, I assume that its ℓ_∞ norm is bounded, arguably a more natural assumption for the applications I consider.

2.1 Minimax-linear estimator with uncorrelated and homoscedastic CATE estimators

In this section, I make the following assumption.

Assumption 1 *For all $g \in \{1, ..., G\}$:*

1. *For all $g' \neq g$, $cov\left(\widehat{\tau}_g, \widehat{\tau}_{g'}\right) = 0$.*
2. *There is a strictly positive unknown real number σ and positive known real numbers $(v_g)_{1 \leq g \leq G}$ such that $V\left(\widehat{\tau}_g\right) \leq \sigma^2 v_g$. The upper bound in the previous inequality is sharp.*
3. *There is a strictly positive known real number B such that $|\tau_g| \leq B\sigma$.*

Assumption 1 requires that the estimators $\widehat{\tau}_g$ be unbiased, uncorrelated across g, and that their variances can be bounded by the product of an unknown real number σ^2 and known real numbers v_g. It also requires that the CATEs be bounded in absolute value by $B\sigma$.

This assumption may for instance be applicable to stratified completely randomized controlled trials (RCTs), see Section 9.3.2 in Imbens & Rubin (2015). There, groups are equal to the experimental strata. τ_g is the CATE in stratum g, and $\widehat{\tau}_g$ is just the difference between the average outcome of treated and control units in that stratum. Under the assignment mechanism in Section 9.3.2 in Imbens & Rubin (2015), $\widehat{\tau}_g$ is unbiased for τ_g, and $cov\left(\widehat{\tau}_g, \widehat{\tau}_{g'}\right) = 0$ for all $g' \neq g$. If one further assumes that the outcome is homoscedastic,

$$V\left(\widehat{\tau}_g\right) \leq \sigma^2 \left(\frac{1}{n_{0,g}} + \frac{1}{n_{1,g}}\right),$$

where $n_{0,g}$ and $n_{1,g}$ respectively denote the number of treated and control units in stratum g. The upper bound is sharp: it is reached if the treatment effect is homogeneous in stratum g, or if the units in stratum g are randomly drawn from a super population. In stratified RCTs, Point 3 assumes that the CATEs are all bounded in absolute value by B standard deviations of the outcome. Normalizing by the outcome's standard deviation is a common practice in applied research. Based on the literature, researchers often have a pretty good sense of the effect sizes, in percent of the outcome's standard deviation, that the intervention they consider may realistically produce. In such instances, they may be able to come up with a plausible value for B. When the literature is silent as to which value of B may be a plausible upper bound for the CATEs, researchers can conduct a sensitivity analysis by varying B.

Beyond stratified RCTs, there are other instances where this setup is applicable, including for instance treatment effect estimation under uncounfoundedness, when treatment is independent of potential outcomes conditional on covariates taking a finite number of values.

For any $1 \times G$ deterministic vector $\boldsymbol{w} = (w_1, ..., w_G)$, let

$$\widehat{\tau}(\boldsymbol{w}) = \sum_{g=1}^{G} w_g \widehat{\tau}_g. \tag{2.2}$$

$\widehat{\tau}(\boldsymbol{w})$ is a linear combination of the estimators $\widehat{\tau}_g$. Lemma 2.1 gives its worst-case MSE.

Lemma 2.1 *(Worst-case MSE of $\widehat{\tau}(\boldsymbol{w})$)*
If Assumption 1 holds, then for any $1 \times G$ deterministic vector $\boldsymbol{w} = (w_1, ..., w_G)$

$$E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2\right) \leq \overline{MSE}(\boldsymbol{w}) \equiv \sigma^2 \left(\sum_{g=1}^{G} w_g^2 v_g + B^2 \left(\sum_{g=1}^{G} |w_g - p_g| \right)^2 \right).$$

The upper bound in the previous display is sharp: it is attained if $\tau_g = \sigma B \left(1\{w_g \geq p_g\} - 1\{w_g < p_g\}\right)$ and $V(\widehat{\tau}_g) = \sigma^2 v_g$.

Without loss of generality, assume that

$$p_1 v_1 \leq p_2 v_2 \leq ... \leq p_G v_G.$$

Let $\overline{g} = \min\{g \in \{1, ..., G\} : \frac{1}{\frac{1}{B^2} + \sum_{g'=g}^{G} \frac{1}{v_{g'}}} \sum_{g'=g}^{G} p_{g'} \leq p_g v_g\}$. $\overline{g}$ is well defined, because $\frac{1}{\frac{1}{B^2} + \frac{1}{v_G}} p_G \leq p_G v_G$. For any $h \in \{1, ..., G\}$, let $\boldsymbol{w}_h$ be such that

$$\begin{aligned} w_{g,h} &= p_g \text{ for all } g < h \\ w_{g,h} &= \frac{1}{v_g} \frac{1}{\frac{1}{B^2} + \sum_{g'=h}^{G} \frac{1}{v_{g'}}} \sum_{g'=h}^{G} p_{g'} \text{ for all } g \geq h. \end{aligned} \tag{2.3}$$

Finally, let

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathbb{R}^G}{\operatorname{argmin}} \, \overline{MSE}(\boldsymbol{w}).$$

It follows from Lemma 2.1 that $\widehat{\tau}(\boldsymbol{w}^*)$ is the minimax-linear estimator of τ.

Theorem 2.2 *(Minimax-linear estimator of τ, with bounded CATEs)*
If Assumption 1 holds, then $\widehat{\tau}(\boldsymbol{w}^*) = \widehat{\tau}(\boldsymbol{w}_{h^*})$, where $h^* = \underset{h \in \{\overline{g}, ..., G\}}{\operatorname{argmin}} \, \overline{MSE}(\boldsymbol{w}_h)$.

Theorem 2.2 shows that under Assumption 1, the minimax-linear estimator is a weighted sum of the $\widehat{\tau}_g$s, with positive weights, that sum to less than 1. For a precisely estimated $\widehat{\tau}_g$ (one with a low value of $p_g v_g$), the optimal weight is just p_g. On the other hand, for an imprecisely estimated $\widehat{\tau}_g$ (one with a high value of $p_g v_g$), the optimal weight is proportional to one over v_g, the non-constant part of its variance. For an imprecisely estimated $\widehat{\tau}_g$, the optimal weight is also shrunk towards zero, where the shrinkage depends on how close to zero B is.

Importantly, the minimax-linear estimator in Theorem 2.2 is feasible: given B, the optimal weights $\boldsymbol{w}^*$ only depend on the known quantities $(v_g)_{1 \leq g \leq G}$ and $(p_g)_{1 \leq g \leq G}$. For instance, in a stratified RCT, $v_g = 1/n_{0,g} + 1/n_{1,g}$ and $p_g = n_g/n$, where $n_g = n_{0,g} + n_{1,g}$ and $n = \sum_{g=1}^{G} n_g$. Accordingly, the weights $\boldsymbol{w}^*$ are not stochastic, and

$$V(\widehat{\tau}(\boldsymbol{w}^*)) \leq \sigma^2 \sum_{g=1}^{G} (w_g^*)^2 \left(\sigma_{0,g}^2 / n_{0,g} + \sigma_{1,g}^2 / n_{1,g} \right), \tag{2.4}$$

where $\sigma^2_{0,g}$ and $\sigma^2_{1,g}$ respectively denote the variances of the untreated and treated outcomes in stratum g. The right hand side in the previous display can easily be estimated, thus giving rise to a conservative variance estimator $\widehat{V}(\widehat{\tau}(\boldsymbol{w}^*))$. If the sample size is large enough for $\widehat{\tau}(\boldsymbol{w}^*)/V(\widehat{\tau}(\boldsymbol{w}^*))$ to be approximately normally distributed,[4] one may use $\widehat{\tau}(\boldsymbol{w}^*)$ and $\widehat{V}(\widehat{\tau}(\boldsymbol{w}^*))$ to construct conservative confidence intervals for $\sum_{g=1}^G w_g^* \tau_g$. Deriving conservative confidence intervals for τ would require accounting for the estimator's bias. I refer the reader to Armstrong & Kolesár (2018) for confidence intervals trading-off bias and variance optimally.

One always has $w_G^* < p_G$, so $\widehat{\tau}(\boldsymbol{w}^*)$ never coincides with the unbiased estimator $\widehat{\tau}(\boldsymbol{p})$. In stratified RCTs, $\widehat{\tau}(\boldsymbol{w}^*)$ is somewhere "in between" the unbiased and strata fixed effects estimators. Let $\boldsymbol{p} = (p_1, ..., p_G)$. The unbiased estimator is equal to $\widehat{\tau}(\boldsymbol{p})$. Let $\widehat{\beta}_{fe}$ be the coefficient of D_{ig} in the regression of Y_{ig} on a constant, D_{ig} and strata fixed effects. Let

$$\boldsymbol{w}_{fe} = \left(\frac{\left(\frac{1}{n_{0,1}} + \frac{1}{n_{1,1}}\right)^{-1}}{\sum_{g=1}^G \left(\frac{1}{n_{0,g}} + \frac{1}{n_{1,g}}\right)^{-1}}, ..., \frac{\left(\frac{1}{n_{0,G}} + \frac{1}{n_{1,G}}\right)^{-1}}{\sum_{g=1}^G \left(\frac{1}{n_{0,g}} + \frac{1}{n_{1,g}}\right)^{-1}} \right).$$

It follows from, e.g., (3.3.7) in Angrist & Pischke (2008), that $\widehat{\tau}(\boldsymbol{w}_{fe}) = \widehat{\beta}_{fe}$. $\widehat{\tau}(\boldsymbol{w}^*)$ assigns to precisely estimated $\widehat{\tau}_g$s the same weights as the unbiased estimator, but it assigns to imprecisely estimated $\widehat{\tau}_g$s weights proportional to those used by the strata fixed effects estimator, shrunk towards zero. Note that in a stratified RCT, $p_g v_g = 1/(n p_{d,g}(1 - p_{d,g}))$, where $p_{d,g} = n_{1,g}/n_g$ is the proportion of treated units in strata g. Accordingly, precisely (resp. imprecisely) estimated $\widehat{\tau}_g$s are those for which $p_{d,g}$ is close to (resp. far from) 1/2.

Corollary 2.3 below shows that when $B \to +\infty$, meaning that one does not restrict the magnitude of the CATEs, $\widehat{\tau}(\boldsymbol{w}^*)$ converges towards $\widehat{\tau}(\boldsymbol{p})$, the unbiased estimator.

Corollary 2.3 *(Minimax-linear estimator of τ when $B \to +\infty$)* $\lim_{B \to +\infty} \boldsymbol{w}^* = (p_g)_{g \in \{1,...,G\}}$.

When the bias that can arise from downweighting one of the $\widehat{\tau}_g$s goes to infinity, bias dominates variance and the unbiased estimator becomes optimal.

Operationally, to find the minimax-linear estimator, one just needs to compute $\overline{g}$, and then evaluate $\overline{MSE}(\boldsymbol{w})$ at $\boldsymbol{w}_h$ for $h \in \{\overline{g}, ..., G\}$. The following lemma shows that to compute $\overline{g}$, one just needs to evaluate the inequalities $\frac{1}{\sum_{g'=g}^G \frac{1}{v_{g'}}} \sum_{g'=g}^G p_{g'} \leq p_g v_g$ for $g = G-1$, $g = G-2$, etc., until one finds a first value where the inequality fails. $\overline{g}$ is equal to that value plus one.

Lemma 2.2

$$\frac{1}{\sum_{g'=g}^G \frac{1}{v_{g'}}} \sum_{g'=g}^G p_{g'} \leq p_g v_g \Rightarrow \frac{1}{\sum_{g'=g+1}^G \frac{1}{v_{g'}}} \sum_{g'=g+1}^G p_{g'} \leq p_{g+1} v_{g+1}.$$

Finally, I give sufficient conditions under which the minimax-linear estimator still has lower worst-case MSE than the unbiased one with heteroscedasticity. As this is not of essence, below I omit that CATEs' variances can sometimes only be conservatively estimated.

[4] See Li & Ding (2017) for a central limit theorem for completely randomized experiments.

Assumption 2 *1. For all $g \in \{1, ..., G\}$, there is a strictly positive unknown real number σ, unknown real numbers $(h_g)_{1 \leq g \leq G}$ such that $h_g \geq 1$ for all g, and positive known real numbers $(v_{0,g}, v_{1,g})_{1 \leq g \leq G}$, such that $V(\widehat{\tau}_g) = \sigma^2(v_{0,g} + h_g v_{1,g})$.*

2. For all $g \in \{1, ..., G\}$, there is a strictly positive unknown real number σ, an unknown real number h such that $0 \leq h \leq 1$, and positive known real numbers $(v_{0,g}, v_{1,g})_{1 \leq g \leq G}$, such that $V(\widehat{\tau}_g) = \sigma^2(v_{0,g} + h v_{1,g})$.

In a stratified RCT, Point 1 of Assumption 2 holds if the untreated outcome's variance does not vary across strata, as is the case when in each stratum researchers standardize their outcome by its standard deviation among the stratum's control group, and if in each stratum the variance of the treated outcome is larger than that of the untreated one. Then, σ^2 is the untreated outcome's variance, while h_g is the ratio of the treated and untreated outcomes' variances in strata g. Point 2 instead holds if the variance of the treated outcome is lower than that of the untreated one, and if heteroscedasticity is constant across strata.

Corollary 2.4 *(Ordering of $\widehat{\tau}(\boldsymbol{w}^*)$'s and $\widehat{\tau}(\boldsymbol{p})$'s worst-case MSE with heteroskedasticity)*

1. *If Points 1 and 3 of Assumption 1 and Point 1 of Assumption 2 hold, the worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ is lower than that of $\widehat{\tau}(\boldsymbol{p})$.*

2. *If Points 1 and 3 of Assumption 1 and Point 2 of Assumption 2 hold, and if*

$$h \geq \frac{B^2 \left(\sum_{g=1}^{G} |w_g^* - p_g|\right)^2 - \sum_{g=1}^{G}((p_g)^2 - (w_g^*)^2)v_{0,g}}{\sum_{g=1}^{G}((p_g)^2 - (w_g^*)^2)v_{1,g}},$$

the worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ is lower than that of $\widehat{\tau}(\boldsymbol{p})$.

In stratified RCTs, Point 1 of Corollary 2.4 implies that the worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ is lower than that of $\widehat{\tau}(\boldsymbol{p})$, if the variance of the outcome with treatment is larger than that of the outcome without treatment. If heteroscedasticity goes in the other direction and does not vary across strata, Point 2 shows that the worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ is still lower than that of $\widehat{\tau}(\boldsymbol{p})$ if the ratio of the treated and untreated outcomes' variances is greater than a lower bound which only depends on the design and can be readily computed. In the application in Section 5, this lower bound is equal to 0.06 so the worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ can only be higher than that of of $\widehat{\tau}(\boldsymbol{p})$ under implausible amounts of heteroscedasticity. Overall, it seems that $\widehat{\tau}(\boldsymbol{w}^*)$'s worst-case MSE is still lower than $\widehat{\tau}(\boldsymbol{p})$'s under realistic heteroscedasticity.[5]

[5]In stratified RCTs, Assumption 2 and Point 3 of Assumption 1 bound the CATEs by B% of the untreated outcome's variance. If one uses instead the treated outcome's variance as the numeraire, the conclusions of Corollary 2.4 revert: $\widehat{\tau}(\boldsymbol{w}^*)$'s worst-case MSE is always lower than $\widehat{\tau}(\boldsymbol{p})$'s if the untreated outcome's variance is larger than that of the treated outcome, while $\widehat{\tau}(\boldsymbol{w}^*)$'s worst-case MSE is lower than $\widehat{\tau}(\boldsymbol{p})$'s under a bound on heteroscedasticity if the opposite holds. Assumption 2 follows the common practice in applied work of standardizing the outcome by its variance in the control group.

2.2 Minimax-linear estimator with correlated homoscedastic CATE estimators

In this section, I replace Point 1 of Assumption 1 by the following assumption.

Assumption 3 *There are known, positive real numbers* $(c_{g,g'})_{1\leq g\neq g'\leq G}$ *such that for all* $g' \neq g$, $c_{g,g'} = c_{g',g}$ *and cov* $(\widehat{\tau}_g, \widehat{\tau}_{g'}) = \sigma^2 c_{g,g'}$.

Assumption 3 allows for covariances between the $\widehat{\tau}_g$s, but requires that their covariances be equal to the product of the outcome's variance σ^2, and known real numbers $c_{g,g'}$.

This framework may for instance be applicable to differences-in-differences (DID) estimators in staggered adoption designs. Assume one has a panel of N units observed over T periods. The design is staggered, meaning that every unit's treatment $D_{i,t}$ is weakly increasing over time: $D_{i,t} \geq D_{i,t-1}$. Let t_i be the first date at which unit i is treated. For every $k \in \{1, ..., T\}$, let $N_k = \sum_{i=1}^{N} 1\{t_i = k\}$ be the number of units that start receiving the treatment at period k. For every $t \geq k$, let $\tau_{k,t}$ denote the average effect of the treatment at period t among units that started receiving it at period k. τ, the average treatment effect on the treated (ATT), can be decomposed as follows:

$$\tau = \sum_{k:N_k>0} \sum_{t=k}^{T} \frac{N_k}{N_1} \tau_{k,t}, \tag{2.5}$$

where $N_1 = \sum_{i,t} D_{i,t}$ is the number of treated units. Therefore, (2.1) holds, with a group g being a pair (k, t).

Callaway & Sant'Anna (2020), Sun & Abraham (2020), and de Chaisemartin & D'Haultfœuille (2020) propose estimators of $\tau_{k,t}$ that can all be written as

$$\widehat{\tau}_{k,t} = \frac{1}{N_k} \sum_{i:t_i=k} (Y_{i,t} - Y_{i,k-1}) - \frac{1}{N_{C_t}} \sum_{j\in C_t} (Y_{j,t} - Y_{j,k-1}),$$

where C_t is a set of control units at t and N_{C_t} is the number of units in that set. To simplify, I assume below that $C_T \neq \emptyset$: there are never-treated units. Then, C_t are the never treated units in Sun & Abraham (2020), the not-yet treated units in de Chaisemartin & D'Haultfœuille (2020), while Callaway & Sant'Anna (2020) consider both the never- and not-yet treated units. In all cases, $C_{t+1} \subseteq C_t$. The estimators $\widehat{\tau}_{k,t}$ are unbiased for $\tau_{k,t}$ under a parallel trends assumption (see Callaway & Sant'Anna, 2020).

Assume that the potential outcomes without being ever treated $Y_{i,t}(0)$ are independent across (i, t) and homoscedastic with variance σ^2, that the treatment effects are not stochastic, and that the treatments are non-stochastic. Those are essentially the assumptions of the Gauss-Markov Theorem (Borusyak et al., 2021). Then, for any $1 \leq k \leq t \leq T$,

$$V(\widehat{\tau}_{k,t}) = 2\sigma^2 \left(\frac{1}{N_k} + \frac{1}{N_{C_t}} \right), \tag{2.6}$$

and for any $1 \leq k \leq t \leq T$, $1 \leq k' \leq t' \leq T$, $k < k'$, and $t < t'$,

$$\begin{aligned} cov(\widehat{\tau}_{k,t}, \widehat{\tau}_{k,t'}) &= \sigma^2 \left(\frac{1}{N_k} + \frac{1}{N_{C_t}} \right) \\ cov(\widehat{\tau}_{k,t}, \widehat{\tau}_{k',t}) &= \sigma^2 \frac{1}{N_{C_t}} \\ cov(\widehat{\tau}_{k,t}, \widehat{\tau}_{k',t'}) &= 0, \end{aligned} \tag{2.7}$$

so Point 2 of Assumption 1 and Assumption 3 hold.

Theorem 2.5 *(Minimax-linear estimator of τ, with bounded CATEs, correlations, and homoscedasticity)*
If Points 2 and 3 of Assumption 1 and Assumption 3 hold, then for any $1 \times G$ deterministic vector $\boldsymbol{w} = (w_1, ..., w_G)$

$$E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2 \right) \leq \overline{MSE}_2(\boldsymbol{w}) \equiv \sigma^2 \left(\sum_{g=1}^{G} \left(w_g^2 v_g + \sum_{g' \neq g}^{G} w_g w_{g'} c_{g,g'} \right) + B^2 \left(\sum_{g=1}^{G} |w_g - p_g| \right)^2 \right).$$

The upper bound in the previous display is sharp: it is attained if $\tau_g = B\left(1\{w_g \geq p_g\} - 1\{w_g < p_g\}\right)$ and $V(\widehat{\tau}_g) = \sigma^2 v_g$. Minimizing $\overline{MSE}_2(\boldsymbol{w})$ across all $\boldsymbol{w}$ such that $w_g \geq 0$ for all g is equivalent to minimizing

$$\sum_{g=1}^{G} \left(w_g^2 v_g + \sum_{g' \neq g}^{G} w_g w_{g'} c_{g,g'} \right) + B^2 \left(\sum_{g=1}^{G} (p_g - w_g) \right)^2, \tag{2.8}$$

subject to $0 \leq w_g \leq p_g$.

While it does not have a closed-form solution, the minimization problem in (2.8) is easy to solve numerically. Given B, this problem is feasible, as it only depends on known quantities. Accordingly, the optimal weights are non-stochastic, and the variance of the minimax-linear estimator can easily be estimated if one has an estimator of the variance-covariance matrix of $(\widehat{\tau}_g)_{g \in \{1,...,G\}}$. In the DID staggered design example, the numbers $(v_g)_{g \in \{1,...,G\}}$ and $(c_{g,g'})_{1 \leq g \neq g' \leq G}$ depend on the design, so the optimal weights are non-stochastic conditional on the design. Note that the estimator in Theorem 2.5 is minimax across all linear combinations of the $\widehat{\tau}_g$s with positive weights. Extending that result to allow for negative weights may be achieved by developing an algorithm similar to the LASSO-LAR algorithm (see Efron et al., 2004 and Rosset & Zhu, 2007) to minimize $\overline{MSE}_2(\boldsymbol{w})$, as Armstrong & Kolesár (2021*b*) do in the overidentified GMM setup they consider. This extension is left for future work.

3 Extensions

3.1 Feasible minimax-linear estimators with positive CATEs

In this section, I consider again the setup of Section 2.1, but I add the following assumption:

Assumption 4 *For all $g \in \{1, ..., G\}$, $0 \leq \tau_g$.*

Assumption 4 is applicable when CATEs are known to all be positive, on top of being bounded. It is plausible when the treatment can ex-ante be assumed to not be detrimental. Theorem 3.1 below still holds if one instead assumes that CATEs are all negative. Let

$$\overline{MSE}^{+}(\boldsymbol{w}) = \sigma^2 \left(\sum_{g=1}^{G} w_g^2 v_g + B^2 \max \left(\left(\sum_{g=1}^{G} (w_g - p_g) 1\{w_g > p_g\} \right)^2, \left(\sum_{g=1}^{G} (w_g - p_g) 1\{w_g < p_g\} \right)^2 \right) \right).$$

Theorem 3.1 *(Minimax-linear estimator of τ, with bounded and positive CATEs) If Assumptions 1 and 4 hold, then for any $1 \times G$ deterministic vector $\boldsymbol{w} = (w_1, ..., w_G)$*

$$E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2 \right) \leq \overline{MSE}^{+}(\boldsymbol{w}).$$

The upper bound in the previous display is sharp. $\widehat{\tau}(\boldsymbol{w}_{h^*}) = \underset{\boldsymbol{w} \in \mathbb{R}^G}{argmin}\ \overline{MSE}^{+}(\boldsymbol{w})$.

Perhaps surprisingly, Theorem 3.1 shows that the minimax-linear estimator is the same when one assumes positive and bounded CATEs and when one only assumes bounded CATEs. Assuming $\tau_g \geq 0$ can only change the worst-case squared bias of $\widehat{\tau}(\boldsymbol{w})$, not its variance. Accordingly, the conclusion of Theorem 3.1 still holds with correlated CATE estimators: there as well, assuming positive CATEs does not change the minimax-linear estimator.

Rather than boundedness and sign restrictions, there are alternative restrictions on CATEs one may wish to consider. For instance, one may wish to assume that each CATE is no more than $B\sigma$ away from the ATE: $|\tau_g - \tau| \leq B\sigma$. Deriving the minimax-linear estimator under such restrictions is more complicated than under the restrictions I consider in this paper: as τ is a weighted average of the τ_gs, deriving a closed form of a sharp upper bound of the squared bias of $\widehat{\tau}(\boldsymbol{w})$ is not straightforward. This extension is left for future work.

3.2 Infeasible minimax-linear estimators without homoscedasticity

A result similar to Theorem 2.2 still holds without the homoscedasticity assumption, and under a modified version of Point 3 in Assumption 1:

Assumption 5 *For all $g \in \{1, ..., G\}$: there is a strictly positive known real number B such that $|\tau_g| \leq B$.*

Without loss of generality, assume that

$$p_1 V(\widehat{\tau}_1) \leq p_2 V(\widehat{\tau}_2) \leq ... \leq p_G V(\widehat{\tau}_G).$$

Let $\overline{g}_2 = \min\{g \in \{1,...,G\} : \frac{1}{\frac{1}{B^2}+\sum_{g'=g}^{G}\frac{1}{V(\widehat{\tau}_{g'})}} \sum_{g'=g}^{G} p_{g'} \leq p_g V(\widehat{\tau}_g)\}$. For any $h \in \{1,...,G\}$, let $\boldsymbol{w}_{h,2}$ be such that

$$\begin{aligned} w_{g,h,2} &= p_g \text{ for all } g < h \\ w_{g,h,2} &= \frac{1}{V(\widehat{\tau}_g)} \frac{1}{\frac{1}{B^2} + \sum_{g'=h}^{G} \frac{1}{V(\widehat{\tau}_g)}} \sum_{g'=h}^{G} p_{g'} \text{ for all } g \geq h. \end{aligned} \tag{3.1}$$

Finally, let

$$\overline{MSE}_3(\boldsymbol{w}) = \sum_{g=1}^{G} w_g^2 V(\widehat{\tau}_g) + B^2 \left(\sum_{g=1}^{G} |w_g - p_g| \right)^2$$

and

$$\boldsymbol{w}_2^* = \underset{\boldsymbol{w} \in \mathbb{R}^G}{\operatorname{argmin}} \, \overline{MSE}_3(\boldsymbol{w}).$$

Theorem 3.2 *(Minimax-linear estimator of τ, with bounded CATEs and heteroskedasticity) If Point 1 of Assumption 1 and Assumption 5 hold, then for any $1 \times G$ deterministic vector $\boldsymbol{w} = (w_1, ..., w_G)$,*

$$E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2 \right) \leq \overline{MSE}_3(\boldsymbol{w}).$$

The upper bound in the previous display is sharp: it is attained if $\tau_g = B\left(1\{w_g \geq p_g\} - 1\{w_g < p_g\}\right)$. $\widehat{\tau}(\boldsymbol{w}_2^*) = \widehat{\tau}(\boldsymbol{w}_{h_2^*})$, where $h_2^* = \underset{h \in \{\overline{g}_2,...,G\}}{argmin} \, \overline{MSE}_3(\boldsymbol{w}_{h,2})$.

Theorem 3.2 shows that without the homoscedasticity assumption, the minimax-linear estimator is still a weighted sum of the $\widehat{\tau}_g$, with positive weights, that sum to less than 1, and where the most precisely estimated CATEs receive a weight equal to the share of the population their group accounts for, while the least precisely estimated CATEs receive a weight proportional to one over the estimator's variance.

While the minimax-linear estimator in Theorem 2.2 is feasible, that in Theorem 3.2 is infeasible, as it depends on the variances of the $\widehat{\tau}_g$s, that are unknown. In most instances, it is possible to estimate those variances,[6] to then form a feasible estimator $\widehat{\tau}\left(\widehat{\boldsymbol{w}_2^*}\right)$ proxying for $\widehat{\tau}\left(\boldsymbol{w}_2^*\right)$. Studying the properties of $\widehat{\tau}\left(\widehat{\boldsymbol{w}_2^*}\right)$ is left for future work.

Finally, one can also relax the assumption that the $\widehat{\tau}_g$s are uncorrelated, or that their correlations has the specific expression in Assumption 3.

Theorem 3.3 *(Minimax-linear estimator of τ, with bounded CATEs, heteroskedasticity, and correlations)*
If Assumption 5 holds, then for any $1 \times G$ deterministic vector $\boldsymbol{w} = (w_1, ..., w_G)$

$$E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2 \right) \leq \overline{MSE}_4(\boldsymbol{w}) \equiv \sum_{g=1}^{G} \left(w_g^2 V(\widehat{\tau}_g) + \sum_{g' \neq g}^{G} w_g w_{g'} cov(\widehat{\tau}_g, \widehat{\tau}_{g'}) \right) + B^2 \left(\sum_{g=1}^{G} |w_g - p_g| \right)^2.$$

[6] In stratified RCTs with non-stochastic potential outcomes, it is only possible to estimate upper bounds of the variances, but that does not affect the result in Theorem 3.2, as those upper bounds are sharp.

The upper bound in the previous display is sharp: it is attained if $\tau_g = B\left(1\{w_g \geq p_g\} - 1\{w_g < p_g\}\right)$. *If* $cov(\widehat{\tau}_g, \widehat{\tau}_{g'}) \geq 0$ *for all* (g, g')*, then minimizing* $\overline{MSE}_4(\boldsymbol{w})$ *across all* $\boldsymbol{w}$ *such that* $w_g \geq 0$ *for all* g *is equivalent to minimizing*

$$\sum_{g=1}^{G}\left(w_g^2 V(\widehat{\tau}_g) + \sum_{g' \neq g}^{G} w_g w_{g'} cov(\widehat{\tau}_g, \widehat{\tau}_{g'})\right) + B^2\left(\sum_{g=1}^{G}(p_g - w_g)\right)^2, \tag{3.2}$$

subject to $0 \leq w_g \leq p_g$.

The minimization problem in (3.2) is easy to solve numerically. This problem is not feasible, as it depends on the variance-covariance matrix of $(\widehat{\tau}_g)_{1 \leq g \leq G}$, which is typically unknown. But a feasible estimator can be computed, by replacing those quantities by estimators.

4 Numerical examples

4.1 Stratified RCTs

In this section, I compute $\boldsymbol{w}^*$ in stratified RCTs, and compare the performance of $\widehat{\tau}(\boldsymbol{w}^*)$ to that of the unbiased and fixed effects estimators $\widehat{\tau}(\boldsymbol{p})$ and $\widehat{\tau}(\boldsymbol{w}_{fe})$. I first consider an RCT with two strata of 100 units ($N = 200$). The first strata is perfectly balanced: half of units are treated. The second strata may be imbalanced, and I vary its treatment probability p_{imb} from 0.5 to 0.99. In each design, I compute $\boldsymbol{w}^*$ assuming $B = 0.75$ and $B = 0.5$. Then, I repeat the same exercise but considering an RCT with two strata of 50 units ($N = 100$). This gives rise to four sets of estimators, for which results are shown on the four panels of Figure 1 below. On each panel, the blue and brown lines respectively show w_1^* and w_2^*, the weights assigned to the balanced and imbalanced strata by the minimax-linear estimator $\widehat{\tau}(\boldsymbol{w}^*)$, as a function of p_{imb}. The green and orange lines respectively show the ratio of the standard error and worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ and $\widehat{\tau}(\boldsymbol{p})$, under the homoscedasticity condition in Point 2 of Assumption 1. Finally, the black and red lines respectively show the ratio of the standard error and worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ and $\widehat{\tau}(\boldsymbol{w}_{fe})$, under that same assumption.

In all panels, w_1^* starts below 0.5, and increases with p_{imb}, until it reaches 0.5. w_2^* on the other hand is decreasing with p_{imb}: the more imbalanced the second stratum is, the more the minimax-linear estimator downweights it. Comparing the four panels, one can see that w_2^* is increasing in B. The larger B is, the less the minimax-linear estimator downweights the imbalanced stratum, as downweighting it may result in more bias. w_2^* is also increasing in N. The larger N is, the lower the variances of the $\widehat{\tau}_g$ estimators, and the less variance matters in the bias-variance trade-off. By construction, $\widehat{\tau}(\boldsymbol{w}^*)$ always has lower worst-case MSE and standard error than $\widehat{\tau}(\boldsymbol{p})$. The magnitude of the difference depends on p_{imb}, B, and N. For instance, if $B = 1$ and $N = 200$, $se(\widehat{\tau}(\boldsymbol{w}^*))/se(\widehat{\tau}(\boldsymbol{p})) \leq 0.9$ if p_{imb} is greater than 0.89. If $B = 0.5$ and $N = 200$ (resp. $B = 1$ and $N = 100$, $B = 0.5$ and $N = 100$)

$se(\widehat{\tau}(\boldsymbol{w}^*))/se(\widehat{\tau}(\boldsymbol{p})) \leq 0.9$ if $p_{imb} \geq 0.77$ (resp. $p_{imb} \geq 0.80$, $p_{imb} \geq 0.5$). $\widehat{\tau}(\boldsymbol{w}^*)$ also always has lower worst-case MSE than $\widehat{\tau}(\boldsymbol{w}_{fe})$, and this difference is large for values of p_{imb} greater than 0.75. Perhaps surprisingly, $\widehat{\tau}(\boldsymbol{w}^*)$ also almost always has lower standard error than $\widehat{\tau}(\boldsymbol{w}_{fe})$. This is due to the shrinkage embedded in $\widehat{\tau}(\boldsymbol{w}^*)$.

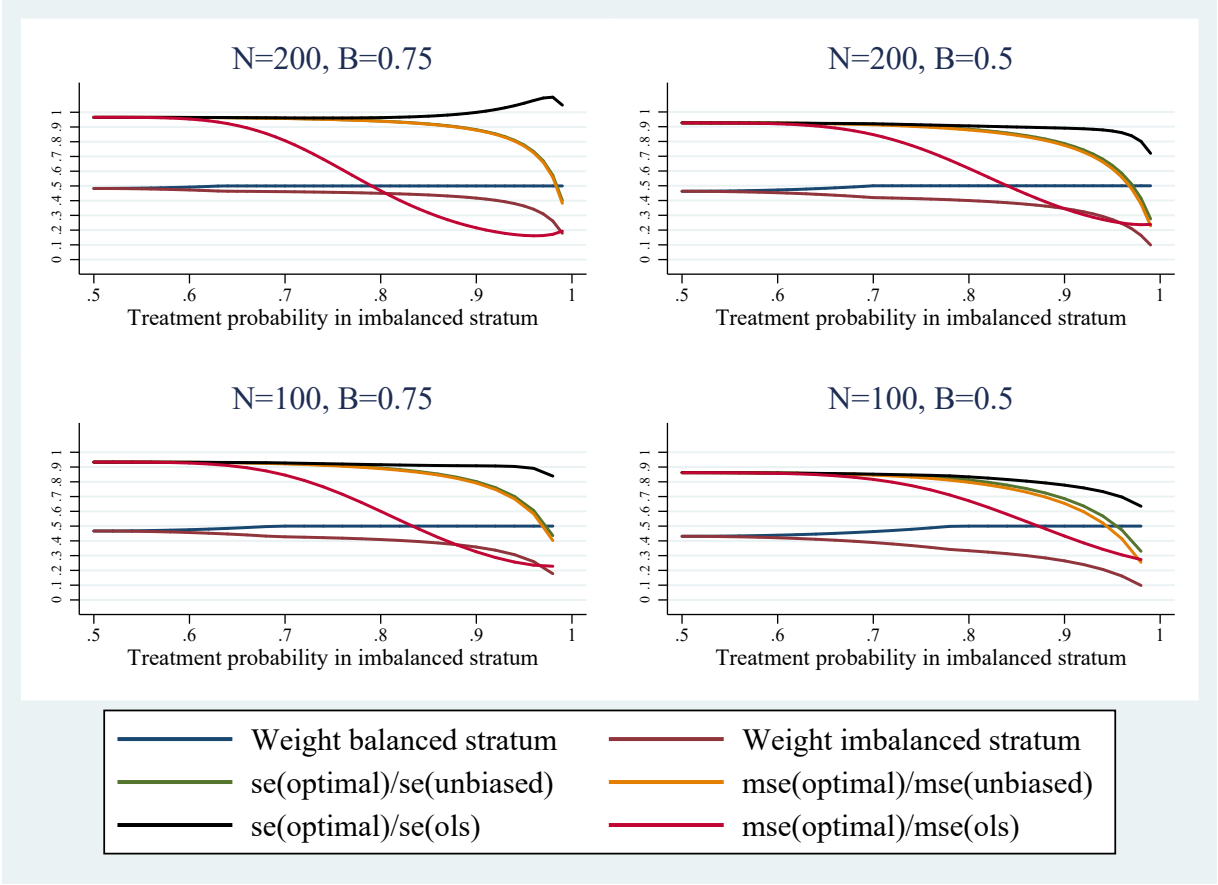

Figure 1: $\widehat{\tau}(\boldsymbol{w}^*)$, $\widehat{\tau}(\boldsymbol{p})$, and $\widehat{\tau}(\boldsymbol{w}_{fe})$ in stratified RCTs

Notes: This figure compares the performance of the minimax-linear estimator $\widehat{\tau}(\boldsymbol{w}^*)$ to that of the unbiased and fixed effects estimators $\widehat{\tau}(\boldsymbol{p})$ and $\widehat{\tau}(\boldsymbol{w}_{fe})$ in stratified RCTs with two strata. On each panel, the blue and brown lines respectively show w_1^* and w_2^*, the weights assigned to the balanced and imbalanced strata by $\widehat{\tau}(\boldsymbol{w}^*)$, as a function of the treatment probability in the imbalanced strata. The green and orange lines respectively show the ratio of the standard error and worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ and $\widehat{\tau}(\boldsymbol{p})$, under the homoscedasticity condition in Point 2 of Assumption 1. Finally, the black and red lines respectively show the ratio of the standard error and worst-case MSE of $\widehat{\tau}(\boldsymbol{w}^*)$ and $\widehat{\tau}(\boldsymbol{w}_{fe})$, under that same assumption. In the top left panel, each strata has 100 units ($N = 200$) and $B = 0.75$. In the top right panel, each strata has 100 units ($N = 200$) and $B = 0.5$. In the bottom left panel, each strata has 50 units ($N = 100$) and $B = 0.75$. In the bottom right panel, each strata has 50 units ($N = 100$) and $B = 0.5$.

4.2 Difference-in-difference estimators with a staggered adoption design

In this section, I consider another numerical example: a staggered adoption design with 50 units and five time periods, where 10 units get treated in periods 2, 3, 4, and 5, while 10 units remain never treated. For $(k,t) \in \{2,...,5\} \times \{k,...,5\}$, let $\widehat{\tau}_{k,t} = \frac{1}{10}\sum_{i:t_i=k}(Y_{i,t} - Y_{i,k-1}) - \frac{1}{10(6-t)}\sum_{i:t_i>t}(Y_{i,t} - Y_{i,k-1})$ be an estimator of the average treatment effect at period t in the cohort that started receiving the treatment at period k. This estimator is unbiased under a parallel trends assumption (see Callaway & Sant'Anna, 2020).

If one assumes that units' potential outcomes without treatment $Y_{i,t}(0)$ are independent across (i,t) and homoscedastic with variance σ^2, that the treatment effects are not stochastic, and that the treatments are non-stochastic, one can use (2.7) to derive the variance-covariance matrix of $(\widehat{\tau}_{k,t})_{(k,t)\in\{2,...,5\}\times\{k,...,5\}}$. Then, for $B = 0.75$, the solution to the minimization problem in (2.8) is

$$\boldsymbol{w}^* = (0.1, 0.1, 0.1, 0.0148, 0.1, 0.1, 0.0565, 0.1, 0.1, 0.1).$$

The minimax-linear estimator downweights $\widehat{\tau}_{2,5}$ and $\widehat{\tau}_{3,5}$, two effects estimated at the last period of the panel, when few units can be used as controls. Doing so leads to a substantial gain under Assumption 3: $se(\widehat{\tau}(\boldsymbol{w}^*))/se(\widehat{\tau}(\boldsymbol{p})) = 0.83$ and $mse(\widehat{\tau}(\boldsymbol{w}^*))/mse(\widehat{\tau}(\boldsymbol{p})) = 0.82$.

It may often be implausible that the outcomes of the same unit are uncorrelated over time. To assess if results are sensitive to that assumption, I assume instead that $V(Y_{i,t}(0)) = 1$ for all (i,t) and $cov(Y_{i,t}(0), Y_{i,t'}(0)) = \rho^{t'-t}$ for all i and $1 \leq t \leq t' \leq T$. Those conditions for instance hold if $Y_{i,t}$ follows a stationnary AR(1) model with parameter ρ: $Y_{i,1}(0) = \varepsilon_{i,1}$, and $Y_{i,t}(0) = \rho Y_{i,t-1}(0) + \sqrt{1-\rho^2}\varepsilon_{i,t}$, with $\varepsilon_{i,t}$ i.i.d. with mean 0 and variance 1. Under those assumptions,

$$V(\widehat{\tau}_{k,t}) = 2\left(1-\rho^{t-k+1}\right)\left(\frac{1}{N_k} + \frac{1}{N_{C_t}}\right),$$

and for any $1 \leq k \leq t \leq T$, $1 \leq k' \leq t' \leq T$, $k < k'$, and $t < t'$,

$$\begin{aligned}
cov(\widehat{\tau}_{k,t}, \widehat{\tau}_{k,t'}) &= \left(\rho^{t'-t} - \rho^{t-k+1} - \rho^{t'-k+1} + 1\right)\left(\frac{1}{N_k} + \frac{1}{N_{C_t}}\right) \\
cov(\widehat{\tau}_{k,t}, \widehat{\tau}_{k',t}) &= \left(1 - \rho^{t-k+1} - \rho^{t-k'+1} + \rho^{k'-k}\right)\frac{1}{N_{C_t}} \\
cov(\widehat{\tau}_{k,t}, \widehat{\tau}_{k',t'}) &= 1\{k' \leq t\}\left(\rho^{t'-t} - \rho^{t-k'+1} - \rho^{t'-k+1} + \rho^{k'-k}\right)\frac{1}{N_{C_t}}.
\end{aligned}$$

With $\rho = 0.50$, $se(\widehat{\tau}(\boldsymbol{w}^*))/se(\widehat{\tau}(\boldsymbol{p})) = 0.80$ and $mse(\widehat{\tau}(\boldsymbol{w}^*))/mse(\widehat{\tau}(\boldsymbol{p})) = 0.78$. Results are similar with $\rho = 0.25$ and $\rho = 0.75$. With $\rho = 0.9$, $se(\widehat{\tau}(\boldsymbol{w}^*))/se(\widehat{\tau}(\boldsymbol{p})) = 0.76$ and $mse(\widehat{\tau}(\boldsymbol{w}^*))/mse(\widehat{\tau}(\boldsymbol{p})) = 1.05$. The weights $\boldsymbol{w}^*$ are computed assuming independent outcomes, but in this numerical example the precision gains attached to using $\widehat{\tau}(\boldsymbol{w}^*)$ are even larger with positively correlated outcomes. On the other hand, the worst-case MSE gains attached to using $\widehat{\tau}(\boldsymbol{w}^*)$ persist with moderately positively correlated outcomes, but disappear with highly positively correlated outcomes.

Overall, $\widehat{\tau}(\boldsymbol{w}^*)$ may lead to substantial precision and MSE gains in staggered adoption designs with a number of units of the same order as the number of US states.

5 Application

In this section, I use the data from Behaghel et al. (2017) to illustrate the results in the paper. The authors conducted a stratified RCT to estimate the effect of a boarding school for disadvantaged students in France. The boarding school's pedagogy is similar to that of "No Excuse" charter schools in the US. It has capacity constraints at the gender $\times$ grade level, and in 2009 and 2010, the school had more applicants than seats in 14 gender $\times$ grade strata. In each stratum, seats were randomly offered to some applicants. Thereafter, treatment is defined as receiving an offer to enter the school. This is not the same thing as entering the school, an issue I shall return to. Two years after the randomization, 363 applicants out of the 395 that participated in a lottery took a standardized maths test. The number of treatment and control applicants per strata is shown in Table 1 below. The treatment probability varies across strata, but there is no strata where it is higher than 0.9 or lower than 0.1.

Table 1: Number of treatment and control applicants in each strata

Strata	Control applicants	Treated applicants
8th grade, males, 2009	11	15
8th grade, females, 2009	15	3
9th grade, females, 2009	8	22
10th grade, males, 2009	5	22
10th grade, females, 2009	36	27
6th grade, males, 2010	6	9
7th grade, males, 2010	9	8
7th grade, females, 2010	5	10
8th grade, males, 2010	5	19
9th grade, males, 2010	8	6
9th grade, females, 2010	3	13
10th grade, males, 2010	12	16
10th grade, females, 2010	39	24
11th grade, females, 2010	3	4

Notes: Number of treated and control applicants in the 14 strata in Behaghel et al. (2017).

I use the data from Behaghel et al. (2017) to compute the minimax-linear estimator $\widehat{\tau}(\boldsymbol{w}^*)$. Here, τ_g is the intention-to-treat effect of receiving an offer on students' test scores two years after the lottery. At that point, the first-stage effect of receiving an offer on the number of

years spent in the school is equal to 1.01 (standard error=0.16),[7] so the τ_gs can be interpreted as effects of having spent one year in the boarding school. The first-stage effects may vary across strata, which would complicate this interpretation. However, I cannot reject that the first-stages are equal for males and females (t-stat=-1.24), for 2009 and 2010 applicants (t-stat=-1.06), and for middle- and high-school applicants (t-stat=0.53), so the first-stage effect indeed seems to be constant across strata.[8]

Based on the literature, 0.5σ is a plausible upper bound for the CATEs. The paper studying the closest intervention is Curto & Fryer Jr (2014), who study a "No Excuse" charter boarding school in Washington DC. In their full sample, they find that one year spent in the school increases students' math test scores by 0.23σ. They also estimate CATEs in eight subgroups of students: males/females, students benefiting/not benefiting from the free lunch program, students in/not in special education, and students above/below the median at baseline. The estimated effects in those 8 subgroups are included between 0.04 and 0.36σ, and in 7 of the 8 subgroups one can reject at the 90% level that the effect is greater than 0.5σ, the only exception being the special education group that only has 30 students. Results from Angrist et al. (2010), Dobbie & Fryer Jr (2011), and Abdulkadiroğlu et al. (2011), three papers studying successful non-boarding "No Excuse" charter schools in New-York and Boston, also suggest that 0.5σ is a plausible upper bound. Together, these papers estimate 14 CATEs of spending one year in those schools on students' math test scores. All estimates are included between 0.18 and 0.36σ, and one can reject at the 90% level that 13 of the 14 CATEs are greater than 0.5σ. One may argue that the intervention in Behaghel et al. (2017) takes place in France, a different setting from that in those papers. However, effects of 0.5σ are very rarely found for educational interventions in high-income countries (see Fryer Jr, 2017). As a robustness check, I will show that results do not change much with $B = 0.6$.

Results are shown in Table 2. In the full sample (Panel A), $\widehat{\tau}(\boldsymbol{w}^*)$ is 13% smaller than $\widehat{\tau}(\boldsymbol{p})$. Its robust standard error, estimated following (2.4), is 9% smaller. Assuming homoskedascity, we can compute the worst-case MSE of both estimators, expressed in percentage points of the outcome's variance. That of $\widehat{\tau}(\boldsymbol{w}^*)$ is 6% smaller. As shown in Point 1 of Corollary 2.4, with heteroscedasticity $\widehat{\tau}(\boldsymbol{w}^*)$'s worst-case MSE is still lower than $\widehat{\tau}(\boldsymbol{p})$'s if the treated outcome's variance is higher than that of the untreated outcome. This may be a plausible assumption in this context, as quantile treatment effects of this intervention are higher at the top than at the bottom of the distribution of test scores (see Behaghel et al., 2017). I still compute the lower bound in Point 2 of Corollary 2.4, and find that it is equal to 0.06: under Point 2 of Assumption 2, the variance of the treated outcome would have to be more than 94% lower

[7] Numbers differ from those in Behaghel et al. (2017), because they use the doubly-reweighted-ever-offer (DREO) estimators proposed by de Chaisemartin & Behaghel (2020). As the DREO estimators of the treatment effect in each strata do not satisfy Point 2 of Assumption 1, I use instead the initial offer estimators. Moreover, Behaghel et al. (2017) have control variables in their specification, while I do not have controls.

[8] Strata are too small to directly test if the first-stage is the same in all of them.

than that of the untreated outcome for $\widehat{\tau}(\boldsymbol{w}^*)$'s worst-case MSE to be higher than $\widehat{\tau}(\boldsymbol{p})$'s.

Those precision and MSE gains are achieved by downweighting the least balanced strata. The weights assigned by $\widehat{\tau}(\boldsymbol{w}^*)$ to 8th-grade-females-2009 and 10th-grade-males-2009 applicants are roughly half as large as the proportions these strata account for in the sample. The weights assigned by $\widehat{\tau}(\boldsymbol{w}^*)$ to all the other strata are equal to their shares in the population.

$\widehat{\tau}(\boldsymbol{w}^*)$'s robust standard error is also 3% smaller than that of $\widehat{\tau}(\boldsymbol{w}_{fe})$, while its worst-case MSE is 33% smaller. When one looks at female and male students separately, the precision and MSE gains attached to using $\widehat{\tau}(\boldsymbol{w}^*)$ rather than $\widehat{\tau}(\boldsymbol{p})$ tend to be larger. For instance, among boys $\widehat{\tau}(\boldsymbol{w}^*)$'s robust standard error and worst-case MSE are respectively 15 and 12% smaller than that of $\widehat{\tau}(\boldsymbol{p})$. Finally, results are similar with $B = 0.6$. Then, in the full sample $\widehat{\tau}(\boldsymbol{w}^*)$'s robust standard error and worst-case MSE are respectively 7 and 4% smaller than that of $\widehat{\tau}(\boldsymbol{p})$, and 2 and 41% smaller than that of $\widehat{\tau}(\boldsymbol{w}_{fe})$.

Table 2: $\widehat{\tau}(\boldsymbol{w}^*)$, $\widehat{\tau}(\boldsymbol{p})$, and $\widehat{\tau}(\boldsymbol{w}_{fe})$ in Behaghel et al. (2017)

	$\widehat{\tau}(\boldsymbol{w}^*)$	$\widehat{\tau}(\boldsymbol{p})$	$\widehat{\tau}(\boldsymbol{w}_{fe})$
Panel A: Full sample (N=363)			
Point estimate	0.097	0.111	0.101
Robust standard error	0.127	0.139	0.130
Worst-case MSE	0.012	0.013	0.019
Panel B: Females (N=212)			
Point estimate	0.177	0.190	0.180
Robust standard error	0.146	0.155	0.160
Worst-case MSE	0.021	0.022	0.026
Panel C: Males (N=151)			
Point estimate	-0.009	0.000	-0.010
Robust standard error	0.216	0.253	0.220
Worst-case MSE	0.028	0.032	0.039

Notes: This table shows $\widehat{\tau}(\boldsymbol{w}^*)$, $\widehat{\tau}(\boldsymbol{p})$, and $\widehat{\tau}(\boldsymbol{w}_{fe})$ in Behaghel et al. (2017). $\widehat{\tau}(\boldsymbol{w}^*)$ is computed with $B = 0.5$. The treatment is defined as being offered a seat in the boarding school. The outcome is students' standardized maths test scores two years after the lottery. The table shows the point estimates, their robust standard errors, and their worst-case MSE computed assuming homoscedasticity and expressed in percentage points of the outcome's variance. In the first panel, results are shown for the full sample. In the second (resp. third) panel, results are shown for female (resp. male) students.

6 Conclusion

This paper considers situations where one wants to estimate the ATE in a population that can be divided into G groups, and one has unbiased estimators of the CATE in each group

with heterogeneous levels of statistical precision. To trade-off bias and variance, I derive the minimax estimator of the ATE, across all linear combinations of CATEs' estimators, assuming that CATEs are bounded by B standard deviations of the outcome and homoscedasticity.

References

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J. & Pathak, P. A. (2011), 'Accountability and flexibility in public schools: Evidence from boston's charters and pilots', *The Quarterly Journal of Economics* **126**(2), 699–748.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A. & Walters, C. R. (2010), 'Inputs and impacts in charter schools: Kipp lynn', *American Economic Review* **100**(2), 239–43.

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

Armstrong, T. B. & Kolesár, M. (2018), 'Optimal inference in a class of regression models', *Econometrica* **86**(2), 655–683.

Armstrong, T. B. & Kolesár, M. (2021*a*), 'Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness', *Econometrica* **89**(3), 1141–1177.

Armstrong, T. B. & Kolesár, M. (2021*b*), 'Sensitivity analysis using approximate moment condition models', *Quantitative Economics* **12**(1), 77–108.

Behaghel, L., De Chaisemartin, C. & Gurgand, M. (2017), 'Ready for boarding? the effects of a boarding school for disadvantaged students', *American Economic Journal: Applied Economics* **9**(1), 140–164.

Berry, J. C. (1990), 'Minimax estimation of a bounded normal mean vector', *Journal of Multivariate Analysis* **35**(1), 130–139.

Borusyak, K., Jaravel, X. & Spiess, J. (2021), Revisiting event study designs: Robust and efficient estimation. Working Paper.

Callaway, B. & Sant'Anna, P. H. (2020), 'Difference-in-differences with multiple time periods', *Journal of Econometrics* .

Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. (2009), 'Dealing with limited overlap in estimation of average treatment effects', *Biometrika* **96**(1), 187–199.

Curto, V. E. & Fryer Jr, R. G. (2014), 'The potential of urban boarding schools for the poor: Evidence from seed', *Journal of Labor Economics* **32**(1), 65–93.

de Chaisemartin, C. & Behaghel, L. (2020), 'Estimating the effect of treatments allocated by randomized waiting lists', *Econometrica* **88**(4), 1453–1477.

de Chaisemartin, C. & D'Haultfœuille, X. (2020), 'Difference-in-differences estimators of intertemporal treatment effects', *Available at SSRN 3731856* .

Dobbie, W. & Fryer Jr, R. G. (2011), 'Are high-quality schools enough to increase achievement among the poor? evidence from the harlem children's zone', *American Economic Journal: Applied Economics* **3**(3), 158–87.

Donoho, D. L. (1994), 'Statistical estimation and optimal recovery', *The Annals of Statistics* pp. 238–270.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *The Annals of statistics* **32**(2), 407–499.

Fryer Jr, R. G. (2017), The production of human capital in developed countries: Evidence from 196 randomized field experiments, *in* 'Handbook of economic field experiments', Vol. 2, Elsevier, pp. 95–322.

Ibragimov, I. A. & Khas'minskii, R. Z. (1985), 'On nonparametric estimation of the value of a linear functional in gaussian white noise', *Theory of Probability & Its Applications* **29**(1), 18–32.

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

Imbens, G. & Wager, S. (2019), 'Optimized regression discontinuity designs', *Review of Economics and Statistics* **101**(2), 264–278.

Li, X. & Ding, P. (2017), 'General forms of finite population central limit theorems with applications to causal inference', *Journal of the American Statistical Association* **112**(520), 1759–1769.

Noack, C. & Rothe, C. (2019), 'Bias-aware inference in fuzzy regression discontinuity designs', *arXiv preprint arXiv:1906.04631* .

Rambachan, A. & Roth, J. (2019), 'An honest approach to parallel trends', *Unpublished manuscript, Harvard University.[99]* .

Rosset, S. & Zhu, J. (2007), 'Piecewise linear regularized solution paths', *The Annals of Statistics* pp. 1012–1030.

Sun, L. & Abraham, S. (2020), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *Journal of Econometrics* .

Vidakovic, B. (1993), 'On the efficiency of affine minimax rules in estimating a bounded multivariate normal mean: On the efficiency of affine minimax rules', *Communications in Statistics-Simulation and Computation* **22**(3), 655–669.

Proofs

Proof of Lemma 2.1

$$\begin{aligned}
E\left(\left(\widehat{\tau}(\boldsymbol{w})-\tau\right)^2\right) =& V(\widehat{\tau}(\boldsymbol{w}))+\left(E\left(\widehat{\tau}(\boldsymbol{w})\right)-\tau\right)^2 \\
=& \sum_{g=1}^{G} w_g^2 V(\widehat{\tau}_g)+\left(\sum_{g=1}^{G}(w_g-p_g)\tau_g\right)^2 \\
\leq& \sigma^2\sum_{g=1}^{G} w_g^2 v_g+\left(\sum_{g=1}^{G}(w_g-p_g)\tau_g\right)^2 \\
\leq& \sigma^2\sum_{g=1}^{G} w_g^2 v_g+\left(\sum_{g=1}^{G}|w_g-p_g||\tau_g|\right)^2 \\
\leq& \sigma^2\left(\sum_{g=1}^{G} w_g^2 v_g+B^2\left(\sum_{g=1}^{G}|w_g-p_g|\right)^2\right).
\end{aligned}$$

The first equality follows from the fact that an estimator's MSE is the sum of its variance and squared bias. The second equality follows from the fact $\boldsymbol{w}$ is deterministic, from Equations (2.2) and (2.1), and from Definition 2.1 and Point 1 of Assumption 1. The first inequality follows from Point 2 of Assumption 1. The second inequality follows from the fact that for any real number a, $a^2=|a|^2$, from the triangle inequality, and from the fact that $x\mapsto x^2$ is increasing on $\mathbb{R}^+$. The third inequality follows from Point 3 of Assumption 1.

The sharpness of the upper bound follows from plugging $\tau_g=\sigma B\left(1\{w_g\geq p_g\}-1\{w_g<p_g\}\right)$ and $V(\widehat{\tau}_g)=\sigma^2 v_g$ into the second equality in the previous display.

Proof of Theorem 2.2

First, assume that $\boldsymbol{w}^*$ has at least one coordinate that is strictly larger than the corresponding coordinate of $(p_1,...,p_G)$. Without loss of generality, assume that $w_1^*>p_1$. One has $\overline{MSE}(\boldsymbol{w}^*)>\overline{MSE}(p_1,w_2^*,...,w_G^*)$, a contradiction. Therefore, each coordinate of $\boldsymbol{w}^*$ is at most as large as the corresponding coordinate of $(p_1,...,p_G)$. Accordingly, finding the minimax-linear estimator is equivalent to minimizing $\overline{MSE}(\boldsymbol{w})$ with respect to $\boldsymbol{w}$, across all $\boldsymbol{w}=(w_1,...,w_G)$ such that $w_g\leq p_g$ for all $g\in\{1,...,G\}$.

If $w_g\leq p_g$ for all $g\in\{1,...,G\}$,

$$\overline{MSE}(\boldsymbol{w})=\sigma^2\left(\sum_{g=1}^{G} w_g^2 v_g+B^2\left(\sum_{g=1}^{G}(p_g-w_g)\right)^2\right).$$

Therefore, $\boldsymbol{w}^*$ is the minimizer of

$$\sum_{g=1}^{G} w_g^2 v_g + B^2 \left(\sum_{g=1}^{G} (p_g - w_g) \right)^2,$$

subject to

$$w_g - p_g \leq 0 \text{ for all } g.$$

The objective function is convex, and the inequality constraints are continuously differentiable and concave. Therefore, the necessary conditions for optimality are also sufficient.

The Lagrangian of this problem is

$$L(\boldsymbol{w}, \boldsymbol{\mu}) = \sum_{g=1}^{G} w_g^2 v_g + B^2 \left(\sum_{g=1}^{G} (p_g - w_g) \right)^2 + \sum_{g=1}^{G} 2\mu_g (w_g - p_g).$$

The Karush-Kuhn-Tucker necessary conditions for optimality are

$$\begin{aligned} & w_g^* v_g - B^2 \left(1 - \sum_{g=1}^{G} w_g^* \right) + \mu_g = 0 \\ & w_g^* \leq p_g \\ & \mu_g \geq 0 \\ & \mu_g (w_g^* - p_g) = 0. \end{aligned} \tag{6.1}$$

Those conditions are equivalent to

$$\begin{aligned} w_g^* &= \min \left(\frac{1}{v_g} B^2 \left(1 - \sum_{g=1}^{G} w_g^* \right), p_g \right) \\ \mu_g &= \max \left(0, B^2 \left(1 - \sum_{g=1}^{G} w_g^* \right) - p_g v_g \right). \end{aligned} \tag{6.2}$$

One has that

$$\begin{aligned} & \frac{1}{v_g} B^2 \left(1 - \sum_{g=1}^{G} w_g^* \right) < p_g \\ \Leftrightarrow & B^2 \left(1 - \sum_{g=1}^{G} w_g^* \right) < p_g v_g. \end{aligned}$$

Together with (6.2), the previous display implies that

$$w_g^* < p_g \Rightarrow w_{g+1}^* < p_{g+1}. \tag{6.3}$$

Let $g^* = \min\{g \in \{1, ..., G\} : w_g^* < p_g\}$, with the convention that $g^* = G + 1$ if the set is empty. It follows from Equations (6.2) and (6.3) that

$$\begin{aligned} & w_g^* = p_g \text{ for all } g < g^* \\ & w_g^* = \frac{1}{v_g} B^2 \left(1 - \sum_{g=1}^{G} w_g^* \right) \text{ for all } g \geq g^*. \end{aligned} \tag{6.4}$$

(6.4) implies that

$$\sum_{g=g^*}^{G} w_g^* = \frac{B^2 \sum_{g=g^*}^{G} \frac{1}{v_g}}{1 + B^2 \sum_{g=g^*}^{G} \frac{1}{v_g}} \sum_{g=g^*}^{G} p_g.$$

Plugging this equation into (6.4) yields

$$\begin{aligned} w_g^* &= p_g \text{ for all } g < g^* \\ w_g^* &= \frac{1}{v_g} \frac{1}{\frac{1}{B^2} + \sum_{g'=g^*}^{G} \frac{1}{v_{g'}}} \sum_{g'=g^*}^{G} p_{g'} \text{ for all } g \geq g^*. \end{aligned} \tag{6.5}$$

Finally, assume that $g^* < \overline{g}$. Then, $w_{g^*}^* > p_{g^*}$, a contradiction. Some algebra shows that

$$\begin{aligned} &\overline{MSE}(\boldsymbol{p}) - \overline{MSE}\left(p_1, ..., p_{G-1}, \frac{\frac{1}{v_G}}{\frac{1}{B^2} + \frac{1}{v_G}} p_G\right) \\ =& p_G^2 v_G - \left(p_G^2 \left(\frac{\frac{1}{v_G}}{\frac{1}{B^2} + \frac{1}{v_G}}\right)^2 v_G + B^2 p_G^2 \left(\frac{\frac{1}{B^2}}{\frac{1}{v_G} + \frac{1}{B^2}}\right)^2\right) \\ =& \frac{p_G^2}{\left(\frac{1}{B^2} + \frac{1}{v_G}\right)^2} \left(\frac{v_G}{B^4} + \frac{1}{B^2}\right) > 0. \end{aligned}$$

Therefore,

$$g^* \in \{\overline{g}, ..., G\} \tag{6.6}$$

The result follows from Equations (6.5) and (6.6).

Proof of Corollary 2.3

$\boldsymbol{w}^*$ belongs to $(\boldsymbol{w}_h)_{h \in \{1,...,G\}}$. For every h, for all $g \geq h$,

$$\lim_{B \to +\infty} w_{g,h} = \frac{1}{v_g} \frac{1}{\sum_{g'=h}^{G} \frac{1}{v_{g'}}} \sum_{g'=h}^{G} p_{g'}.$$

Therefore, for every h,

$$\lim_{B \to +\infty} \sum_{g=1}^{G} w_{g,h} = 1,$$

which then implies that

$$\lim_{B \to +\infty} \sum_{g=1}^{G} w_g^* = 1.$$

As $w_g^* \leq p_g$ and $\sum_{g=1}^{G} p_g = 1$, this implies that for every g,

$$\lim_{B \to +\infty} w_g^* = p_g.$$

Proof of Lemma 2.2

Assume that

$$\frac{1}{\sum_{g'=g}^{G}\frac{1}{v_{g'}}}\sum_{g'=g}^{G}p_{g'}\leq p_g v_g.$$

Then,

$$\begin{aligned}
&p_{g+1}v_{g+1}\sum_{g'=g+1}^{G}\frac{1}{v_{g'}}\\
=&p_{g+1}v_{g+1}\sum_{g'=g}^{G}\frac{1}{v_{g'}}-p_{g+1}\frac{v_{g+1}}{v_g}\\
=&p_g v_g\sum_{g'=g}^{G}\frac{1}{v_{g'}}+(p_{g+1}v_{g+1}-p_g v_g)\sum_{g'=g}^{G}\frac{1}{v_{g'}}-p_{g+1}\frac{v_{g+1}}{v_g}\\
\geq&p_g v_g\sum_{g'=g}^{G}\frac{1}{v_{g'}}+p_{g+1}\frac{v_{g+1}}{v_g}-p_g-p_{g+1}\frac{v_{g+1}}{v_g}\\
\geq&\sum_{g'=g+1}^{G}p_{g'}.
\end{aligned}$$

6.1 Proof of Corollary 2.4

Proof of Point 1

Assume that Points 1 and 3 of Assumption 1 hold. If Point 1 of Assumption 2 were to hold with $h_g = 1$ for all g, then Point 2 of Assumption 1 would hold with $v_g = v_{0,g} + v_{1,g}$, and $\widehat{\tau}(\boldsymbol{w}^*)$ would be minimax-linear. Accordingly, its worst-case MSE under that DGP has to be lower than that of $\widehat{\tau}(\boldsymbol{p})$, which implies that

$$\sigma^2 B^2\left(\sum_{g=1}^{G}|w_g^*-p_g|\right)^2\leq\sigma^2\sum_{g=1}^{G}((p_g)^2-(w_g^*)^2)(v_{0,g}+v_{1,g}). \tag{6.7}$$

As for all g, $v_{1,g}\geq 0$, $(p_g)^2-(w_g^*)^2\geq 0$, and $h_g\geq 1$ under Point 1 of Assumption 2,

$$\sigma^2\sum_{g=1}^{G}((p_g)^2-(w_g^*)^2)(v_{0,g}+v_{1,g})\leq\sigma^2\sum_{g=1}^{G}((p_g)^2-(w_g^*)^2)(v_{0,g}+h_g v_{1,g}). \tag{6.8}$$

Combining (6.7) and (6.8) and rearranging proves the result.

Proof of Point 2

If Points 1 and 3 of Assumption 1 and Point 2 of Assumption 2 hold, the worst-case MSEs of $\widehat{\tau}(\boldsymbol{w}^*)$ and $\widehat{\tau}(\boldsymbol{p})$ are respectively equal to

$$\sigma^2\left(\sum_{g=1}^{G}(w_g^*)^2(v_{0,g}+hv_{1,g})+B^2\left(\sum_{g=1}^{G}|w_g^*-p_g|\right)^2\right)$$

and

$$\sigma^2 \sum_{g=1}^{G} (p_g)^2 (v_{0,g} + h v_{1,g}).$$

Taking the difference between the two preceding displays, setting that difference lower than 0 and rearranging yields the result.

Proof of Theorem 2.5

That $E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2\right) \leq \overline{MSE}_2(\boldsymbol{w})$ follows from the same steps as the proof of Lemma 2.1.

As $c_{g,g'} \geq 0$ for all (g, g'), if $w_g \geq 0$ for all g, then using a reasoning similar to that in the proof of Theorem 2.2, one can show that the minimizer of $\overline{MSE}_2(\boldsymbol{w})$ must be such that each of its coordinates are lower than the corresponding coordinate of $\boldsymbol{p}$. Therefore, this minimization problem is equivalent to that in (2.8).

Proof of Theorem 3.1

$0 \leq \tau_g \leq B\sigma$ implies that

$$B\sigma(w_g - p_g)1\{w_g < p_g\} \leq (w_g - p_g)\tau_g \leq B\sigma(w_g - p_g)1\{w_g > p_g\},$$

which in turn implies that

$$B\sigma \sum_{g=1}^{G} (w_g - p_g)1\{w_g < p_g\} \leq \sum_{g=1}^{G} (w_g - p_g)\tau_g \leq B\sigma \sum_{g=1}^{G} (w_g - p_g)1\{w_g > p_g\}. \tag{6.9}$$

Then, reasoning as in the proof of Lemma 2.1 yields

$$E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2\right) \leq \overline{MSE}^{+}(\boldsymbol{w}).$$

When

$$\left| \sum_{g=1}^{G} (w_g - p_g)1\{w_g > p_g\} \right| \geq \left| \sum_{g=1}^{G} (w_g - p_g)1\{w_g < p_g\} \right|,$$

the sharpness of the upper bound follows from plugging $\tau_g = \sigma B 1\{w_g \geq p_g\}$ into (6.9). When

$$\left| \sum_{g=1}^{G} (w_g - p_g)1\{w_g > p_g\} \right| < \left| \sum_{g=1}^{G} (w_g - p_g)1\{w_g < p_g\} \right|,$$

the sharpness of the upper bound follows from from plugging $\tau_g = \sigma B 1\{w_g < p_g\}$ into (6.9).

Finally, assume that w^*, the argmin of $\overline{MSE}^{+}(\boldsymbol{w})$, has at least one coordinate that is strictly larger than the corresponding coordinate of $(p_1, ..., p_G)$. Without loss of generality, assume that $w_1^* > p_1$. One has

$$\left(\sum_{g=1}^{G} (w_g^* - p_g)1\{w_g^* > p_g\} \right)^2 > \left(\sum_{g=2}^{G} (w_g^* - p_g)1\{w_g^* > p_g\} \right)^2$$

and

$$\left(\sum_{g=1}^{G}(w_g^* - p_g)1\{w_g^* < p_g\}\right)^2 = \left(\sum_{g=2}^{G}(w_g^* - p_g)1\{w_g^* < p_g\}\right)^2,$$

so the squared bias of $\widehat{\tau}(\boldsymbol{w}^*)$ is at least as large as that of $\widehat{\tau}(p_1, w_2^*, ..., w_G^*)$, while its variance is strictly larger. Then, $\overline{MSE}^+(\boldsymbol{w}^*) > \overline{MSE}^+(p_1, w_2^*, ..., w_G^*)$, a contradiction. Therefore, each coordinate of $\boldsymbol{w}^*$ is at most as large as the corresponding coordinate of $(p_1, ..., p_G)$. Accordingly, finding the minimax-linear estimator is equivalent to minimizing $\overline{MSE}^+(\boldsymbol{w})$ with respect to $\boldsymbol{w}$, across all $\boldsymbol{w} = (w_1, ..., w_G)$ such that $w_g \leq p_g$ for all $g \in \{1, ..., G\}$. On that set,

$$\begin{aligned}
\overline{MSE}^+(\boldsymbol{w}) =& \sigma^2\left(\sum_{g=1}^{G} w_g^2 v_g + B^2\left(\sum_{g=1}^{G}(w_g - p_g)1\{w_g < p_g\}\right)^2\right) \\
=& \sigma^2\left(\sum_{g=1}^{G} w_g^2 v_g + B^2\left(\sum_{g=1}^{G}(w_g - p_g)1\{w_g \leq p_g\}\right)^2\right) \\
=& \sigma^2\left(\sum_{g=1}^{G} w_g^2 v_g + B^2\left(\sum_{g=1}^{G}(w_g - p_g)\right)^2\right) \\
=& \overline{MSE}(\boldsymbol{w}).
\end{aligned}$$

This completes the proof.

Proof of Theorem 3.2

That $E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2\right) \leq \overline{MSE}_3(\boldsymbol{w})$ follows from the same steps as the proof of Lemma 2.1. That $\overline{MSE}_3(\boldsymbol{w})$ is minimized at $\boldsymbol{w}_{h_2^*}$ follows from the same steps as the proof of Theorem 2.2.

Proof of Theorem 3.3

That $E\left((\widehat{\tau}(\boldsymbol{w}) - \tau)^2\right) \leq \overline{MSE}_4(\boldsymbol{w})$ follows from the same steps as the proof of Lemma 2.1. If $cov(\widehat{\tau}_g, \widehat{\tau}_{g'}) \geq 0$ for all (g, g') and $w_g \geq 0$ for all g, then using a reasoning similar to that in the proof of Theorem 2.2, one can show that the minimizer of $\overline{MSE}_4(\boldsymbol{w})$ must be such that each of its coordinates are lower than the corresponding coordinate of $\boldsymbol{p}$. Therefore, this minimization problem is equivalent to that in (3.2).