Estimating Treatment Effects Under Bounded Heterogeneity*

Soonwoo Kwon† and Liyang Sun‡

March 2026

Abstract

Specifications that impose constant treatment effects are common but biased, while fully flexible alternatives can be imprecise or infeasible. Under a bound on treatment effect heterogeneity, we propose a generalized ridge estimator, `regulaTE`, that yields heterogeneity-aware confidence intervals (CIs). The ridge penalty is chosen to optimally trade off worst-case bias and variance in a Gaussian homoskedastic setting; the resulting CIs remain tight more generally and are valid even under lack of overlap. Varying the bound enables sensitivity analysis to departures from constant effects, which we illustrate in leading empirical applications of unconfoundedness and staggered adoption designs.

`Key words:` heterogeneous treatment effects, limited overlap, ridge regression

`JEL classification codes:` C12, C21, C51.

*We thank Isaiah Andrews, Tim Armstrong, Clément de Chaisemartin, Ben Deaner, Avi Feller, Phillip Heiler, Peter Hull, Jing Kong, Matt Masten, Claudia Noack, Tymon Słoczyński, Jon Roth and conference and seminar participants at Emory University, Erasmus School of Economics, Econometrics Society World Congress, University of Glasgow Adam Smith Business School, UC Berkeley - Statistics, Stanford University, SEA, CEMFI Workshop on Applied Microeconometrics and Panel Data, LMU-NYU Workshop on Advances in Policy Analysis, and ASSA for helpful comments and discussion. Liyang Sun gratefully acknowledges support from the Economic and Social Research Council (new investigator grant UKRI607). Yubin Kim, Justin Lee and Tian Xie provided excellent research assistance. An R library implementing the `regulaTE` estimator proposed in this paper is available online at https://github.com/lsun20/regulaTEr.

†Department of Economics, Brown University, Email: soonwoo_kwon@brown.edu

‡Department of Economics, University College London and CEMFI, Email: liyang.sun@ucl.ac.uk

1 Introduction

In many empirically relevant settings for causal inference, correctly specifying the model requires allowing for fully flexible treatment effect heterogeneity as a function of covariates. However, estimating such rich models often leads to imprecise estimates, and can become infeasible under limited overlap in covariate distributions. Consequently, estimating the fully flexible model may not be desirable in empirical applications, particularly when researchers are primarily interested in the average treatment effect (ATE) rather than the full distribution of heterogeneous effects. Consider the unconfoundedness design, where the treatment is assumed to be as good as randomly assigned conditional on the covariates. Many applied studies estimate a "short" regression of the form that assumes constant effects

$$Y_i = D_i\beta_{\text{short}} + X_i'\gamma_{\text{short}} + u_i,$$

and interpret the estimated regression coefficient $\hat{\beta}_{\text{short}}$ as an estimate of the ATE (e.g., Martinez-Bravo, 2014; Favara and Imbs, 2015; Michalopoulos and Papaioannou, 2016; Bobonis et al., 2016; Xu, 2018).

There is growing recognition that treatment effect heterogeneity affects the interpretation of the "short" regression. In an influential study, Angrist (1998) shows that under unconfoundedness and discrete covariates, the "short" regression estimates a convex average of treatment effects associated with each covariate value, and therefore can differ from estimates based on matching that does not restrict treatment effect heterogeneity. Słoczyński (2022) further shows that this convex average corresponds to a weighted average of treatment effects on the treated and untreated, with counter-intuitive weights. Similar concerns have been raised in recent studies of the "short" regression for staggered adoption designs (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021).

However, revisiting Angrist (1998), the textbook (Angrist and Pischke, 2009, Chapter 3.3.1) asserts that if treatment effect heterogeneity is limited, the "short" regression estimate may still be preferable to unbiased but noisier semi- and nonparametric estimators of the average effect. First, when treatment effects are constant, the "short" regression is unbiased and most efficient under homoskedasticity. Second, Angrist and Pischke (2009) informally

argue that the bias of the "short" regression remains small when treatment effects do not vary too much, and the resulting efficiency gain can still justify its use. An additional, and often overlooked, consideration is that in settings without overlap, semi- and nonparametric estimators may cease to be well defined, whereas the short regression remains feasible because it extrapolates treatment effect estimates to covariate values without overlap. Consequently, practitioners' intuition is that the confidence intervals (CIs) based on the "short" regression should have coverage close to the nominal level, making the "short" regression a pragmatic choice in applied work.

In this paper, we formalize the argument of Angrist and Pischke (2009). Under a restriction on treatment effect heterogeneity, specifically a bound on the variance of conditional average treatment effect heterogeneity (VCATE), we first show how to bias-correct the "short" regression CI such that the resulting CI has correct coverage and is heterogeneity-aware. This approach is feasible even when the average effect is not point identified due to a complete lack of overlap, since the bound on treatment effect heterogeneity provides partial identifying information on the treatment effect for the sample with non-overlap.[1] We denote this bound by C^2. As C^2 increases from zero, the associated bias-corrected "short" regression CI achieves correct coverage under each restriction, allowing one to assess the sensitivity of the "short" regression estimates to violations of constant effects.

However, as the "short" regression estimator does not adjust to the bound, bias correction can yield an excessively wide heterogeneity-aware CI. To provide a more informative sensitivity analysis, we propose a generalized ridge regression, `regulaTE`, which coincides with the "short" regression when heterogeneity is absent but otherwise can depart from the "short" regression if its bias is large. We propose selecting the ridge penalty to minimize the heterogeneity-aware CI length under a normal, homoskedastic setting, which facilitates rapid computation. We show in simulations and empirical illustrations that under more general error distributions, the resulting CI remains tighter than the bias-corrected "short" regression CI. We provide an implementation of sensitivity analysis based on `regulaTE` in

[1] When overlap fails completely, the ATE is not point identified. The VCATE bound provides partial identifying information that restricts the range of treatment effects in the non-overlapping sample, enabling valid inference even in this setting.

our accompanying R package.[2]

An important input for our method is C^2, the bound on VCATE, which we primarily view as a sensitivity parameter rather than a quantity to be estimated. In practice, the researcher examines how the CIs change as C^2 varies over a user-specified range, thereby assessing how sensitive the empirical conclusions are to deviations from constant treatment effects. A useful summary is the "breakdown value" of C^2—the smallest value at which the ATE becomes statistically insignificant—following Kline and Santos (2013) and Masten and Poirier (2020). Because C^2 has a clear interpretation as a bound on VCATE, which is frequently used in variance decomposition and welfare analyses (Kline et al., 2020; Sanchez-Becerra, 2023; de Chaisemartin and Deeb, 2024) and whose magnitude applied researchers often have intuition about, the resulting sensitivity analysis is straightforward to interpret. We conduct sensitivity analysis based on our proposed `regulaTE` in two leading empirical applications under unconfoundedness and staggered adoption. We show how subgroup analyses reported in the original empirical studies can be used to inform a plausible range of bounds for the sensitivity analysis. We emphasize, however, that choosing C^2 in a data-driven manner generally leads either to invalid confidence intervals or to excessively wide intervals, reflecting the impossibility results of Armstrong and Kolesár (2018).

Related literature. Since the seminal work of Angrist (1998), a growing body of literature has characterized the bias of the "short" regression estimator under treatment effect heterogeneity, including Gibbons, Serrato, and Urbancic (2019), Słoczyński (2022), Goldsmith-Pinkham, Hull, and Kolesár (2024), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021). We show how to bias-correct the "short" regression under a restriction on treatment effect heterogeneity, and propose a sensitivity analysis to evaluate the impact of treatment effect heterogeneity on robust estimation of the average effect.

Several papers have proposed estimators for the average effect under restrictions on treatment effects. In the context of the unconfoundedness design, Lechner (2008) assumes outcomes are bounded and derives the Manski bound for the ATE. Lee and Weidner (2021) improves the Manski bound by combining it with the IPW estimator using reference propensity

[2] An R package implementing the `regulaTE` estimator is available online at https://github.com/lsun20/regulaTEr.

scores. Athey, Imbens, and Wager (2018) assume sparsity of effects, Armstrong and Kolesár (2021) assume smoothness of effects with respect to covariates, and de Chaisemartin (2024) impose bounds on effects. In the context of difference-in-differences, Manski and Pepper (2018) impose bounds on the variation in outcomes over time and across states. In contrast to these papers, our method is tied directly to a measure of treatment effect heterogeneity by restricting the variance of treatment effects. This does not entail a functional form assumption on the heterogeneity, mimicking the common belief that effects are broadly similar.

There are semi- and nonparametric estimators, such as the inverse propensity score weighted estimator, that remain unbiased for the average treatment effect even under unrestricted treatment effect heterogeneity. However, these estimators rely on strong overlap for consistency and asymptotic normality. When propensity scores approach 0 or 1, they can become highly variable, making the conventional normal approximation unreliable, as demonstrated in Rothe (2017) or Heiler and Kazak (2021). To address extreme values of propensity scores, Crump, Hotz, Imbens, and Mitnik (2009) propose trimming the inverse propensity score weighted estimator by removing observations with estimated propensity scores outside the range $(0.1, 0.9)$. Since this approach only estimates the average effect for a subpopulation with strong overlap, several papers have characterized the resulting bias and proposed bias correction methods, e.g., Ma and Wang (2020) and Sasaki and Ura (2022), under certain smoothness conditions. Our method provides an alternative route for dealing with limited overlap, which includes settings with complete lack of overlap: rather than focusing on a subpopulation with strong overlap, we restrict treatment effect heterogeneity and use this restriction to infer the worst-case heterogeneity for covariate cells with limited overlap.

The remainder of the paper is organized as follows. Section 2 introduces the setup and formalizes the restriction on treatment effect heterogeneity. Section 3 develops the proposed inference method, establishes its optimality, and illustrates its finite-sample performance in calibrated simulations. Section 4 applies the method to empirical settings and extends it to staggered adoption. Section 5 concludes.

2 Setup and treatment effect heterogeneity

Consider the unconfoundedness setting where we observe a random sample of n units, and each unit is characterized by a pair of potential outcomes $(Y_i(0), Y_i(1))$ under no treatment and treatment, respectively; a set of (possibly including flexible transformations of) covariates $X_i \in \mathbb{R}^{k+1}$ that includes a constant; and a treatment indicator $D_i \in \{0, 1\}$. We assume unconfoundedness (also known as selection on observables): $Y_i(0), Y_i(1) \perp D_i \mid X_i$. As common in empirical practice (Wooldridge, 2010), we further assume that $\mathbb{E}[Y_i(1) \mid X_i]$ and $\mathbb{E}[Y_i(0) \mid X_i]$ are both linear in X_i.[3] Denote $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$ to be the covariate-specific average treatment effect (CATE).

Unless stated otherwise, we condition on the realized values $\{x_i, d_i\}_{i=1}^n$ of the covariates and treatment $\{X_i, D_i\}_{i=1}^n$, where lowercase letters denote realizations. All probability statements are therefore taken with respect to the conditional distribution of $\{Y_i(0), Y_i(1)\}_{i=1}^n$. Under these assumptions, the DGP for the observed outcome Y_i can be written as

$$Y_i = d_i\tau(x_i) + x_i'\gamma + \varepsilon_i, \tag{1}$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of mutually independent noise terms that are mean zero. This fixed-design framework follows, for example, Rothe (2017) and Armstrong and Kolesár (2021), to investigate finite-sample properties. As illustrated in Section 4.2, we also allow the covariates that parameterize treatment effect heterogeneity to differ from the confounders x_i.

Our goal is to estimate a weighted average of the CATEs with known weights, where the weights reflect a predetermined target population. The leading example is the sample ATE, $\beta = \beta^{\text{ATE}} = \mathbb{E}_n[\tau(x_i)] = \frac{1}{n}\sum_{i=1}^n \tau(x_i)$ as defined in Imbens and Wooldridge (2009). Here $\mathbb{E}_n$ denotes the "empirical" mean (i.e., $\mathbb{E}_n a_i = n^{-1}\sum_{i=1}^n a_i$ for any fixed or random sequence

[3] In principle, such linearity assumptions are not required, and our approach to estimating average effects under bounded treatment effect heterogeneity can be extended to more general settings. For example, assuming linearity only of $\mathbb{E}[Y_i(0) \mid X_i]$ leads to a related ridge-type estimation procedure. One could also adopt a fully nonparametric approach by imposing a bound on treatment effect heterogeneity in addition to the smoothness assumptions of Armstrong and Kolesár (2021). Nevertheless, given the prevalence of linear models in empirical practice, and the intuitive interpretation and computational efficiency afforded by this assumption, we view this as a useful trade-off between generality and practicality.

$\{a_i\}_{i=1}^n$). Under the linearity assumption, we can write $\tau(x_i) = \alpha + x_{i,-1}'\delta$, where $x_{i,-1}$ denotes the vector of covariates excluding the constant. Letting $\widetilde{x}_{i,-1} = x_{i,-1} - \mathbb{E}_n[x_{i,-1}]$ denote the demeaned covariates, deviations of the CATE from the ATE are given by $\tau(x_i) - \beta^{\text{ATE}} = \widetilde{x}_{i,-1}'\delta$. Therefore, the DGP (1) can be reparameterized as the *long regression* as in Imbens and Wooldridge (2009):

$$Y_i = d_i\beta_{\text{long}}^{\text{ATE}} + (d_i\tilde{x}_{i,-1})'\delta + x_i'\gamma + \varepsilon_i, \text{ where } \widetilde{x}_{i,-1} = x_{i,-1} - \mathbb{E}_n[x_{i,-1}]. \tag{2}$$

Other leading target parameters include the sample average treatment effect on the treated (ATT) and the sample average treatment effect on the untreated (ATU). Since the long regression (2) handles all three targets through appropriate demeaning of the covariates, we focus on the ATE for simplicity.

We define the short regression as the specification that omits the interaction terms between treatment and covariates:

$$Y_i = d_i\beta_{\text{short}} + x_i'\gamma_{\text{short}} + u_i, \tag{3}$$

which is the most commonly used specification in empirical research.

2.1 Challenges under treatment effect heterogeneity

Even when treatment effects are fully heterogeneous, the long regression (2) remains correctly specified for identifying average effects. However, estimating it can be practically challenging: the resulting estimates may be highly imprecise, and the model may not even be estimable under limited overlap. First, when the covariates x_i are mixed, containing both continuous covariates and discrete covariates, the discrete covariates typically enter the long regression as fixed effects capturing confounding from, for example, location effects. If all units in a given location are treated, contributing to lack of overlap, then the interaction between treatment and the centered location fixed effect becomes multicollinear with the treatment indicator and the location fixed effect itself, rendering the long regression $\hat{\beta}_{\text{long}}$ undefined.

Second, if the covariates $\{x_i\}_{i=1}^n$ are generated by saturating discrete covariates, when

overlap fails entirely for a particular covariate value, the long regression $\hat{\beta}_{\text{long}}$ is again not well-defined. The long regression coefficient estimator is algebraically equivalent to the following IPW estimator:

$$\hat{\beta}^{\text{ATE}}_{\text{long}} = \mathbb{E}_n \left[\frac{d_i - p(x_i)}{p(x_i)\,(1 - p(x_i))} Y_i \right]$$

where $p(x_i)$ denotes the empirical propensity score for covariate value x_i. Therefore, if $p(x_i) \in \{0, 1\}$ for some x_i due to lack of overlap, then the long regression $\hat{\beta}_{\text{long}}$ is not well-defined. Furthermore, because $p(x_i)(1 - p(x_i))$ appears in the denominator, the variability of $\hat{\beta}^{\text{ATE}}_{\text{long}}$ is heavily influenced by observations for which $p(x_i)$ is close to zero or one.

In contrast, the short regression estimator is less affected by limited overlap and remains well-defined even under complete lack of overlap at some covariate values. To see this, by the Frisch-Waugh-Lovell (FWL) theorem, the short regression estimator in (3) can be written as

$$\hat{\beta}_{\text{short}} = \frac{\mathbb{E}_n \left[(d_i - p(x_i))\, Y_i \right]}{\mathbb{E}_n \left[(d_i - p(x_i))\, d_i \right]}.$$

This representation highlights that the short regression places little weight on observations with extreme propensity scores. Moreover, it remains well-defined even when overlap fails for some covariate values, since such observations have no influence on the estimator. However, by omitting interactions between d_i and $\tilde{x}_{i,-1}$, the short regression is generally misspecified in the presence of treatment effect heterogeneity. As a result, it may suffer from omitted variable bias, with the bias measured relative to the ATE. From Angrist (1998), the short regression estimand is:

$$\beta_{\text{short}} = \mathbb{E}[\hat{\beta}_{\text{short}}] = \mathbb{E}_n \left[\frac{p(x_i)\,(1 - p(x_i))}{\mathbb{E}_n \left[p(x_i)\,(1 - p(x_i)) \right]} \tau(x_i) \right]. \tag{4}$$

This estimand coincides with β only if one of the following holds: (i) $\tau(x_i)$ is constant, (ii) $p(x_i)$ is constant, or (iii) $p(x_i)(1-p(x_i))$ is orthogonal to $\tau(x_i)$. In general, the short regression recovers a weighted average of heterogeneous treatment effects, with weights proportional to the conditional variance of treatment assignment. As discussed in Poirier and Słoczyński (2024), the subpopulation for which the short-regression estimand corresponds to a true average treatment effect can be small, limiting the internal validity of the estimator.

Nevertheless, when treatment effect heterogeneity is limited, the short regression remains nearly unbiased and typically yields more precise estimates. This point is made informally by Angrist and Pischke (2009, Chapter 3.3.1): "Of course, the difference in weighting schemes is of little importance if $\tau(x)$ does not vary across cells (though weighting still affects the statistical efficiency of estimators)." This observation helps explain why the short regression is substantially more common than the long regression in empirical practice.

2.2 Bounding the treatment effect heterogeneity via VCATE

To formalize the belief that treatment effects are broadly similar across covariates, we impose an upper bound on the (sample) variance of the CATE $\tau(x_i)$, the VCATE, which we define as $\mathbb{E}_n[(\tau(x_i) - \mathbb{E}_n[\tau(x_i)])^2]$.

While any restriction that imposes that the vector of heterogeneous effects $\{\tau(x_i) - \mathbb{E}_n[\tau(x_i)]\}_{i=1}^n$ lies in a bounded set would, in principle, suffice to control the bias induced by misspecification of the short regression or other linear-in-outcome estimators, we focus on bounds on VCATE for several reasons. First, because variance is the most widely used and canonical measure of dispersion, VCATE provides an intuitive measure of treatment effect heterogeneity that captures how dispersed treatment effects are across covariate values, rather than imposing potentially opaque geometric restrictions on the entire vector of effects.[4] This makes the restriction easy to interpret and to vary in sensitivity analyses. In contrast, many alternative convex restrictions, such as ℓ_p balls, do not correspond to commonly interpreted notions of treatment effect heterogeneity and are therefore harder to calibrate in empirical applications.

Second, VCATE is a well-studied object in the treatment effects literature, and as a result researchers often have a clear sense of how to reason about its magnitude.[5] VCATE arises naturally in variance decompositions and distributional analyses, and recent work has developed estimators and inference procedures for VCATE under a range of identifying assumptions. This existing body of work makes VCATE a familiar and interpretable sensitivity

[4]For example, Levy et al. (2021) and Sanchez-Becerra (2023) introduce VCATE as leading measures of treatment effect heterogeneity.

[5]See Kline et al. (2020), Sanchez-Becerra (2023), and de Chaisemartin and Deeb (2024) for recent work on estimation and inference for VCATE.

parameter: researchers can draw on prior results, empirical benchmarks, and domain knowledge to assess whether a given bound on VCATE is plausible in their setting. Moreover, Popoviciu's inequality gives a simple (yet conservative) bound

$$\mathbb{E}_n[(\tau(x_i) - \mathbb{E}_n[\tau(x_i)])^2] \leq \frac{1}{4}(\max_i \tau(x_i) - \min_i \tau(x_i))^2,$$

providing a simple reference point for calibration.

Note that, under our maintained specification, VCATE can be expressed as a simple quadratic form in the heterogeneity coefficient δ:

$$\mathbb{E}_n\left[(\tau(x_i) - \mathbb{E}_n[\tau(x_i)])^2\right] = \mathbb{E}_n\left[(\tilde{x}'_{i,-1}\delta)^2\right] = \delta'\,\mathbb{E}_n[\tilde{x}_{i,-1}\tilde{x}'_{i,-1}]\,\delta. \tag{5}$$

For notational convenience, define $V_x := \mathbb{E}_n[\tilde{x}_{i,-1}\tilde{x}'_{i,-1}]$, which is the sample analogue of $\mathrm{Var}(X_{i,-1})$. We assume that V_x is invertible, which is a minimal condition ensuring that the short regression is well defined. Accordingly, imposing an upper bound on VCATE of the form $\mathbb{E}_n[(\tau(x_i) - \mathbb{E}_n[\tau(x_i)])^2] \leq C^2$ is equivalent to imposing the weighted quadratic constraint $\delta'\,V_x\,\delta \leq C^2$ on the parameter vector δ. Throughout the paper, we report and plot results on the standard-deviation scale C rather than the variance scale C^2, because C shares the units of the outcome.

Although we do not pursue this direction here, the approach we introduce can be extended straightforwardly to more general restrictions that impose δ to lie in a bounded, convex set. For example, one could replace the quadratic constraint above with coordinatewise bounds of the form $\max_\ell |\delta_\ell| \leq C$ or other convex constraints on δ by following the general procedure described in Armstrong et al. (2023).

3 Heterogeneity-aware confidence intervals

In this section, we develop inference methods that explicitly account for treatment effect heterogeneity. We begin by showing how to construct valid confidence intervals based on estimators that are linear in outcomes, while allowing for limited or even complete lack of overlap.

3.1 Proposed method

Building on (2), we consider the regression model

$$Y_i = d_i\beta + x_i'\gamma + d_i\widetilde{x}_{i,-1}'\delta + \varepsilon_i, \tag{6}$$

where Y_i is the outcome, d_i is the (non-random) treatment indicator, and $x_i = (1, x_{i,-1}')'$ is the vector of non-random covariates including a constant. Define $D = (d_1, \ldots, d_n)'$ as the (realized) treatment vector and $X = (x_1, \ldots, x_n)'$ the matrix of (realized) covariates. We assume $X'X$ is invertible. To introduce our proposal cleanly, we assume $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ throughout this subsection, an assumption we relax in Section 3.2. That is, treatment effect heterogeneity is fully captured by the covariates x_i under this specification.

In a fixed-design setting, estimating the ATE under a VCATE bound is equivalent to conducting inference on β in (6) subject to the constraint $\delta' V_x \delta \leq C^2$ as explained in (5). Equivalently, the restriction on the parameter space is $(\beta, \gamma', \delta')' \in \Theta_C$, where

$$\Theta_C := \{(\beta, \gamma', \delta')' \in \mathbb{R}^{2+2k} : \ \delta' V_x \delta \leq C^2\}.$$

Let $Y = (Y_1, \ldots, Y_n)'$. We focus on linear estimators of the form $\hat{\beta}_a = a'Y$, where $a \in \mathbb{R}^n$ is non-random. Since $(d_i, x_i')'$ are treated as fixed, the weights a may depend on them; this class includes typical estimators such as difference-in-means, the short regression, and the long regression (when defined).

The variance of the estimator $\hat{\beta}_a = a'Y$ is $\mathrm{Var}(\hat{\beta}_a) = \sigma^2 a'a$ under homoskedasticity. From (6), the worst-case bias of $\hat{\beta}_a$ over Θ_C is

$$\overline{\mathrm{bias}}_{a,C} = \sup_{(\beta,\gamma',\delta') \in \Theta_C} \left\{ a'(D\beta + X\gamma + (D \circ \tilde{X}_{-1})\delta) - \beta \right\}, \tag{7}$$

where $D \circ \tilde{X}_{-1}$ is the matrix with ith row equal to $d_i \tilde{x}_{i,-1}'$. Hence, given the weights a, the worst-case bias of the corresponding estimator can be calculated explicitly. Since Θ_C leaves β and γ unrestricted, it is necessary that the weights satisfy $a'D = 1$ and $a'X = 0$ to achieve finite worst-case bias. Under these conditions, the remaining bias $a'(D \circ \tilde{X}_{-1})\delta$ is linear

in δ, and since Θ_C is centrosymmetric in δ (i.e., $\delta \in \Theta_C$ implies $-\delta \in \Theta_C$), the one-sided supremum in (7) equals $\sup_{\Theta_C} |\mathbb{E}[\hat{\beta}_a] - \beta|$.

A fixed-length confidence interval (FLCI) is constructed as

$$\hat{\beta}_a \pm \chi_{a,C}, \qquad \chi_{a,C} = \mathrm{Var}(\hat{\beta}_a)^{1/2}\, \mathrm{cv}_\alpha\Big(\overline{\mathrm{bias}}_{a,C}\,/\mathrm{Var}(\hat{\beta}_a)^{1/2}\Big), \tag{8}$$

where $\mathrm{cv}_\alpha(B)$ denotes the $(1-\alpha)$ quantile of the folded normal distribution $|N(B,1)|$. The critical value therefore reflects the potential worst-case bias of $\hat{\beta}_a$. Importantly, the validity of the confidence interval does not rely on point identification of β. In cases of complete lack of overlap, β is set identified rather than point identified under a VCATE bound; as shown in Lemma 1, an immediate consequence of Armstrong and Kolesár (2018), the FLCI attains correct coverage uniformly over the identified set.

Lemma 1. *Under the assumption that VCATE is bounded by C^2 and the error terms in (6) are Gaussian,*

$$\inf_{(\beta,\gamma',\delta')\in\Theta_C} \Pr\Big(\beta \in [\hat{\beta}_a - \chi_{a,C},\ \hat{\beta}_a + \chi_{a,C}]\Big) \geq 1-\alpha.$$

Proof. Under the stated assumptions, $(\hat{\beta}_a - \beta)/\mathrm{Var}(\hat{\beta}_a)^{1/2} \sim N(b,1)$, where the bias term b satisfies $|b| \leq \overline{\mathrm{bias}}_{a,C}\,/\mathrm{Var}(\hat{\beta}_a)^{1/2}$. The result follows by construction of $\chi_{a,C}$. □

Write $P_X = X(X'X)^{-1}X'$ and $H_X = I - P_X$. As shown in Appendix Lemma A.1, the worst-case bias of the short regression estimator is

$$\overline{\mathrm{bias}}_{a_s,C} = C\sqrt{a_s'(D\circ\tilde{X}_{-1})V_x^{-1}(D\circ\tilde{X}_{-1})'a_s}, \qquad a_s = (D'H_XD)^{-1}H_XD.$$

Because the short-regression weights a_s do not adjust as C increases, bias-corrected short-regression confidence intervals can become excessively long. Since the length of any FLCI of the form $\hat{\beta}_a \pm \chi_{a,C}$ equals $2\chi_{a,C}$ and increases in both $\overline{\mathrm{bias}}_{a,C}$ and $\mathrm{Var}(\hat{\beta}_a)$, improving performance for $C > 0$ requires a better bias–variance trade-off.

To achieve a better bias–variance trade-off, we propose solving the generalized ridge least-squares problem

$$\min_{\beta,\gamma,\delta} \|Y - D\beta - X\gamma - (D\circ\tilde{X}_{-1})\delta\|^2/n + \lambda\,\delta'V_x\delta \tag{9}$$

where $\lambda > 0$ is chosen as discussed later in Section 3.1.1. Let $\hat{\beta}_\lambda$ denote the resulting coefficient estimator for β. Since $\delta' V_x \delta$ equals the sample VCATE as shown in (5), the penalty directly shrinks the heterogeneity coefficients δ toward zero—that is, toward the constant-effects model—with the degree of shrinkage governed by λ. It is not *a priori* obvious that such a penalty on the outcome regression should yield optimal inference for β, since penalized regression targets a prediction error objective rather than the bias–variance trade-off specific to β.[6] However, the quadratic geometry of the VCATE restriction implies that $\hat{\beta}_\lambda$ lies on the bias–variance frontier for every $\lambda > 0$: no linear estimator achieves simultaneously smaller variance and smaller worst-case bias. Moreover, all linear estimators that lie on the frontier can be written in this form. Before we formalize this optimality result in Theorem 1 in Section 3.1.1, we provide some intuition.

Lemma A.2 shows that the generalized ridge regression coefficient estimator $\hat{\beta}_\lambda$ can be written as

$$\hat{\beta}_\lambda = a_\lambda' Y = \frac{\tilde{D}_\lambda' Y}{\tilde{D}_\lambda' D}, \tag{10}$$

where the weights $a_\lambda := \frac{\tilde{D}_\lambda}{\tilde{D}_\lambda' D}$ and the residuals $\tilde{D}_\lambda := D - X\pi_{1,\lambda} - (D \circ \tilde{X}_{-1})\pi_{2,\lambda}$ are from a penalized propensity score regression:

$$\min_{\pi_1, \pi_2} \|D - X\pi_1 - (D \circ \tilde{X}_{-1})\pi_2\|^2/n + \lambda \pi_2' V_x \pi_2. \tag{11}$$

For $\lambda > 0$ define the shrinkage matrix

$$S_\lambda := H_X(D \circ \tilde{X}_{-1}) \left((D \circ \tilde{X}_{-1})' H_X (D \circ \tilde{X}_{-1}) + \lambda V_x\right)^{-1} (D \circ \tilde{X}_{-1})' H_X,$$

which maps any vector $v \in \mathbb{R}^n$ to the fitted values from the ridge regression of $H_X v$ on $H_X(D \circ \tilde{X}_{-1})$ with penalty matrix λV_x. With this notation, the residualized treatment used

[6] Under an ℓ_1 bound on δ, for instance, the LASSO-penalized outcome regression does not produce optimal inference for β. The coincidence between the ridge penalty on the outcome regression and the optimal inference procedure for β is specific to the quadratic VCATE constraint, which has been pointed out in Li (1982) and Armstrong et al. (2023).

by the generalized ridge estimator can be written as

$$\tilde{D}_\lambda = H_X D - H_X(D \circ \tilde{X}_{-1})\pi_{2,\lambda} = (I - S_\lambda)H_X D, \tag{12}$$

and therefore

$$\hat{\beta}_\lambda = \frac{\tilde{D}_\lambda' Y}{\tilde{D}_\lambda' D} = \frac{((I - S_\lambda)H_X D)'Y}{((I - S_\lambda)H_X D)'D}.$$

Hence, the effect of λ on $\hat{\beta}_\lambda$ operates entirely through the shrinkage operator S_λ.

When $\lambda = 0$ and $(D \circ \tilde{X}_{-1})'H_X(D \circ \tilde{X}_{-1})$ is invertible (i.e., the long regression is well-defined),

$$S_0 = H_X(D \circ \tilde{X}_{-1})((D \circ \tilde{X}_{-1})'H_X(D \circ \tilde{X}_{-1}))^{-1}(D \circ \tilde{X}_{-1})'H_X$$

is the orthogonal projection onto $\text{span}(H_X(D \circ \tilde{X}_{-1}))$, and $(I - S_0)H_X D$ equals the residual from regressing D on $(X, D \circ \tilde{X}_{-1})$. Consequently, $\hat{\beta}_0$ coincides with the long-regression coefficient on D.[7] On the other hand, as $\lambda \to \infty$, it is clear that $\tilde{D}_\lambda \to H_X D$ and $\hat{\beta}_\lambda$ converges to the short regression estimator.

To gain more intuition on our estimator, for fixed $\lambda > 0$, suppose X is generated from saturating discrete covariates. Lemma A.3 shows the weights in (10) can be written as $a_{\lambda,i} = \frac{(d_i - p(x_i))(1 - \widetilde{x}_{i,-1}'\pi_{2,\lambda})}{\mathbb{E}_n\left[(d_i - p(x_i))(1 - \widetilde{x}_{i,-1}'\pi_{2,\lambda})d_i\right]}$. The estimand of $\hat{\beta}_\lambda$ is therefore a weighted average of CATEs:

$$\beta_\lambda = \mathbb{E}[\hat{\beta}_\lambda] = \mathbb{E}_n\left[\frac{(p(x_i)(1 - p(x_i)))(1 - \widetilde{x}_{i,-1}'\pi_{2,\lambda})}{\mathbb{E}_n\left[(p(x_i)(1 - p(x_i)))(1 - \widetilde{x}_{i,-1}'\pi_{2,\lambda})\right]}\tau(x_i)\right], \tag{13}$$

As shown in Lemma A.4, expression (13) further simplifies to

$$\beta_\lambda = \mathbb{E}_n\left[\frac{p(x_i)(1 - p(x_i))/(p(x_i)(1 - p(x_i)) + \lambda)}{\mathbb{E}_n\left[p(x_i)(1 - p(x_i))/(p(x_i)(1 - p(x_i)) + \lambda)\right]}\tau(x_i)\right]. \tag{14}$$

The weight on each unit's CATE is proportional to $p(x_i)(1 - p(x_i))/(p(x_i)(1 - p(x_i)) + \lambda) \in [0, 1]$, which equals $1/(1 + \lambda/[p(x_i)(1 - p(x_i))])$. This factor is close to one for well-overlapped cells where $p(x)(1 - p(x))$ is large relative to λ, and close to zero for poorly overlapped cells

[7] When the long regression is not well-defined due to lack of overlap, in the case of discrete covariates, it is possible to analytically characterize the limit of $\hat{\beta}_\lambda$ as $\lambda \to 0$, which is the long regression estimator $\hat{\beta}_{\text{long}}$ restricted to the trimmed subsample with overlap as shown in Lemma A.6.

where $p(x)(1-p(x))$ is small relative to λ. As $\lambda \to 0$, the shrinkage factor approaches one for all cells and the weights converge to cell shares, recovering the ATE. As $\lambda \to \infty$, the weights converge to the short regression weights in (4), proportional to $p(x)(1-p(x))$. For intermediate λ, the ridge estimand overweights well-overlapped cells relative to the ATE, but less so than the short regression: cells with $p(x)(1-p(x)) \gg \lambda$ receive weight close to their population share, while cells with $p(x)(1-p(x)) \ll \lambda$ are heavily downweighted. Note that the weights are always non-negative and sum to one. Later in Section 3.4, we illustrate these weights in the context of Angrist (1998) (see Figure 1).

3.1.1 Choosing the penalty parameter

It remains to choose the penalty parameter λ. We select it to minimize the half-length of the resulting FLCI in the homoskedastic Gaussian setting. By the same argument as Appendix Lemma A.1, the worst-case bias of $\hat{\beta}_\lambda$ takes the form $\overline{\text{bias}}_{a_\lambda,C} = C \cdot \sqrt{a_\lambda'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'a_\lambda}$. Let λ_C^* denote the penalty parameter value that minimizes the half-length of the corresponding fixed-length CI:

$$\lambda_C^* := \text{argmin}_\lambda \ \sigma\|a_\lambda\| \cdot \text{cv}_\alpha\left(\overline{\text{bias}}_{a_\lambda,C}/(\sigma\|a_\lambda\|)\right), \tag{15}$$

and construct the corresponding CI $\hat{\beta}_{\lambda_C^*} \pm \chi_{a_{\lambda_C^*},C}$. We refer to this inference procedure as `regulaTE`, for its ability to regularize treatment effect heterogeneity.

The `regulaTE` CI is in fact the shortest possible fixed-length CI based on any linear estimator in the homoskedastic Gaussian setting. We formalize this in the following theorem.

Theorem 1. *The generalized ridge regression estimator $\hat{\beta}_\lambda$ solves the bias-variance trade-off problem:*

$$\min_{a \in \mathbb{R}^n} a'a \quad s.t. \quad \sup_{(\beta,\gamma',\delta') \in \Theta_C} a'(D\beta + X\gamma + (D \circ \tilde{X}_{-1})\delta) - \beta \leq B \tag{16}$$

with $B = \overline{\text{bias}}_{a_\lambda,C}$. Consequently, the `regulaTE` *CI is the shortest fixed-length CI based on linear estimators.*

The first part of the theorem establishes that $\hat{\beta}_\lambda$ lies on the bias–variance frontier for every $\lambda > 0$: at bias level $B = \overline{\text{bias}}_{a_\lambda,C}$, no linear estimator achieves smaller variance.

This follows from the modulus of continuity characterization of optimal inference (Donoho, 1994; Armstrong and Kolesár, 2018): the generalized ridge regression solves the modulus problem associated with the VCATE constraint. The second part follows because the family $\{\hat{\beta}_\lambda : \lambda > 0\}$ spans the entire Pareto frontier. Since the FLCI half-length $\chi_{a,C}$ depends on the weights a only through the variance $a'a$ and worst-case bias $\overline{\text{bias}}_{a,C}$, minimizing CI length over all linear estimators reduces to minimizing over the ridge family, which is what (15) achieves.

3.2 Implementation and validity under more general error distributions

We begin with an initial estimator $\hat{\theta}^{\text{init}} = (\hat{\beta}^{\text{init}}, \hat{\gamma}^{\text{init}\prime}, \hat{\delta}^{\text{init}\prime})'$ for $\theta = (\beta, \gamma', \delta')'$, obtained either from the long regression (when defined) or from a cross-validated generalized ridge regression that penalizes the weighted ℓ_2 norm $\mathbb{E}_n[(\widetilde{x}'_{i,-1}\delta)^2]$. Define the residuals from this initial estimator $\hat{\varepsilon}_{\text{init},i} = y_i - d_i\hat{\beta}^{\text{init}} - x_i'\hat{\gamma}^{\text{init}} - d_i\widetilde{x}'_{i,-1}\hat{\delta}^{\text{init}}$, and the corresponding initial variance estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}^2_{\text{init},i}$.

For each λ, compute $\hat{\beta}_\lambda$ as before and obtain λ^*_C via (15) by plugging in $\hat{\sigma}$. Let $W = [D,\, X,\, D \circ \tilde{X}_{-1}]$ denote the design matrix. Form the robust variance estimator

$$\hat{V}_{\lambda^*_C,\text{rob}} = \sum_{i=1}^n a^2_{\lambda^*_C,i}\hat{\varepsilon}^2_{\text{init},i}, \quad \text{where } a_{\lambda^*_C} = W\left(W'W + \lambda^*_C \begin{pmatrix} 0 & 0 \\ 0 & V_x \end{pmatrix}\right)^{-1} e_1.$$

The feasible CI is then $\hat{\beta}_{\lambda^*_C} \pm \text{cv}_\alpha\left(\overline{\text{bias}}_{a_{\lambda^*_C},C}/\hat{V}^{1/2}_{\lambda^*_C,\text{rob}}\right)\cdot\hat{V}^{1/2}_{\lambda^*_C,\text{rob}}$. Cluster-robust versions follow analogously.

Remark 1. Due to the heteroskedastic nature of the error terms, the exact optimality results stated in Section 3 no longer hold in general. Nonetheless, the procedure mirrors the common practice of using weights that are optimal under homoskedasticity (i.e., OLS weights) combined with robust standard errors.

The asymptotic validity of the feasible CI is formally stated in Appendix A.2, which follows from a central limit theorem (CLT) applied to $\hat{\beta}_{\lambda^*_C}$. The key requirement is that the

maximal Lindeberg weight associated with the estimator,

$$\text{Lind}(a_{\lambda_C^*}) := \max_{1\le i\le n} \frac{a_{\lambda_C^*,i}^2}{\sum_{j=1}^n a_{\lambda_C^*,j}^2}$$

shrinks sufficiently quickly relative to the error of the initial estimator used to form the residuals. While we provide formal conditions for asymptotic validity in Appendix A.2, the Lindeberg weights can be computed in any given application and serve as a diagnostic for the reliability of the normal approximation; see Noack and Rothe (2024) for an analogous discussion in the context of fuzzy regression discontinuity designs. The companion R package reports the maximal Lindeberg weight and issues a warning when it is large. Note that, due to limited overlap, the convergence rate may be slower than the parametric rate of $n^{-1/2}$.

3.3 Necessity of bounding VCATE

The previous sections demonstrate how to construct `regulaTE` CIs given a bound on VCATE. One might instead wish to construct a "wider" CI that remains valid under unrestricted treatment effect heterogeneity, or a CI that adapts to the true underlying VCATE, shrinking in length according to the true bound while still guaranteeing coverage over a broader class of heterogeneity. We show that neither approach is feasible, thereby establishing the necessity of imposing an *a priori* bound on VCATE.

Intuitively, absent any restriction on the parameter space, the worst-case bias of any linear estimator must be unbounded when overlap fails. The reason is that the data contains no information about treatment effects for strata in which only treated (or untreated) individuals are observed. Formally, recall from (7) that the worst-case bias is linear in the parameters. Thus, only unbiased estimators have bounded bias (in fact, zero) when the parameter space is unrestricted. As evident from the expression, unbiasedness requires $a'D = 1$, $a'X = 0$ and $a'(D \circ \tilde{X}_{-1}) = 0$. Suppose overlap fails for the binary jth covariate so that $X_{j+1} \le D$, where $\le$ is interpreted elementwise. Writing $\bar{x}_{j+1} := \mathbb{E}_n[x_{i,j+1}]$, the conditions for unbiasedness imply $a'D = 1$, $a'X_{j+1} = 0$ and $a'(X_{j+1} + D\bar{x}_{j+1}) = 0$, but it is clear that no weight vector a can satisfy these conditions.[8] Hence, no unbiased linear estimator exists, and the worst-case

[8]This is because due to lack of overlap $D \circ \tilde{X}_{-1} = X_{j+1} + D\bar{x}_{j+1}$.

bias of *any* linear estimator is unbounded. A formal statement and proof of this result can be found in Appendix A.3.

Moreover, it turns out that it is impossible to construct a CI that adapts to the true VCATE. By definition, an adaptive CI has length that automatically reflects the true magnitude of VCATE while maintaining coverage under a conservative *a priori* bound on VCATE. However, Corollary 2.1 of Armstrong et al. (2023) implies that any valid CI must have expected length that reflects the conservative *a priori* bound C^2, even when VCATE is much smaller than C^2. In other words, one cannot automate the choice of the VCATE bound C^2 when constructing the CI.

Therefore, while C is an important input to our method, it cannot be set to $C = \infty$, nor can it be selected in a data-driven way with the goal of adapting to the true VCATE. In Section 4, we discuss how to conduct sensitivity analysis with respect to C.

3.4 Calibrated simulations

We illustrate the theoretical results so far in realistic settings through simulations calibrated to the data generating process in Angrist (1998). Angrist (1998) links social security earnings records to administrative data on a sample of U.S. military applicants from 1979 to 1982 to estimate the effects of voluntary military service on veterans' earnings. The data include the following observable characteristics: year of application, year of birth, education at the time of application, race, and Armed Forces Qualification Test (AFQT) score group. The paper documents heterogeneity in treatment effects: military service modestly increased the civilian earnings of non-white veterans while reducing those of white veterans. Further heterogeneity is observed across background characteristics such as education and AFQT scores, which prompted Angrist (1998) to theoretically analyze the different estimands of the short regression and the long regression (referred to therein as the controlled contrast estimator). The estimates were found to be significantly different, both statistically and economically.

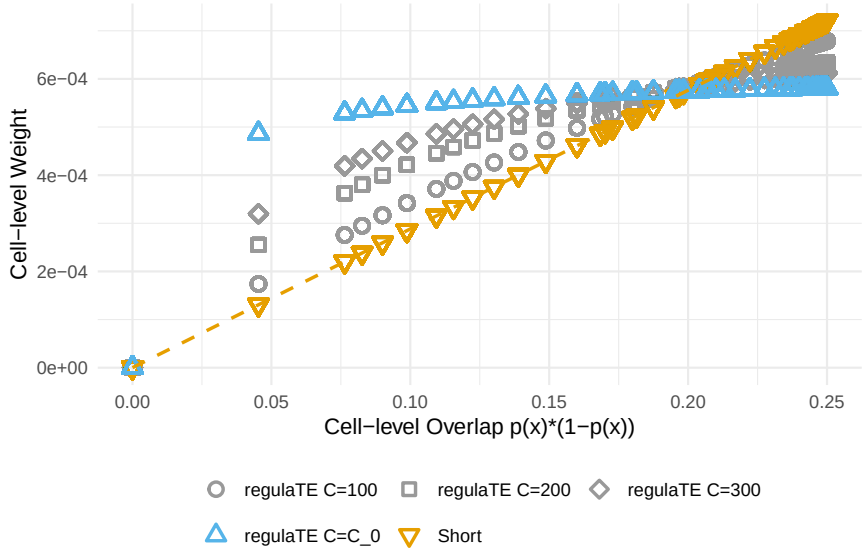
For simplicity, our simulation exercises focus on inference for the average treatment effect on earnings in 1988 among white applicants. Within this population, there are approximately 400 covariate cells. The public replication data from Angrist (1998) report cell-level summary

statistics—such as the mean and standard deviation of earnings, the share of veterans, and the cell frequency—constructed from administrative records covering roughly 100,000 individuals.

To calibrate the simulation to Angrist (1998) and to construct a micro-level dataset, we draw 2,000 individuals, which can be interpreted as a 2% subsample of the original population. Treatment status is assigned using the cell-level share of veterans as the true propensity score, which ranges from 4.6% to 81.2%. Given a fixed realization of treatment assignments, the relatively small sample size leads to a lack of overlap at some covariate values, equivalent to 12.4% of the sample size. As a result, the long regression is not well defined and the ATE is not point identified.

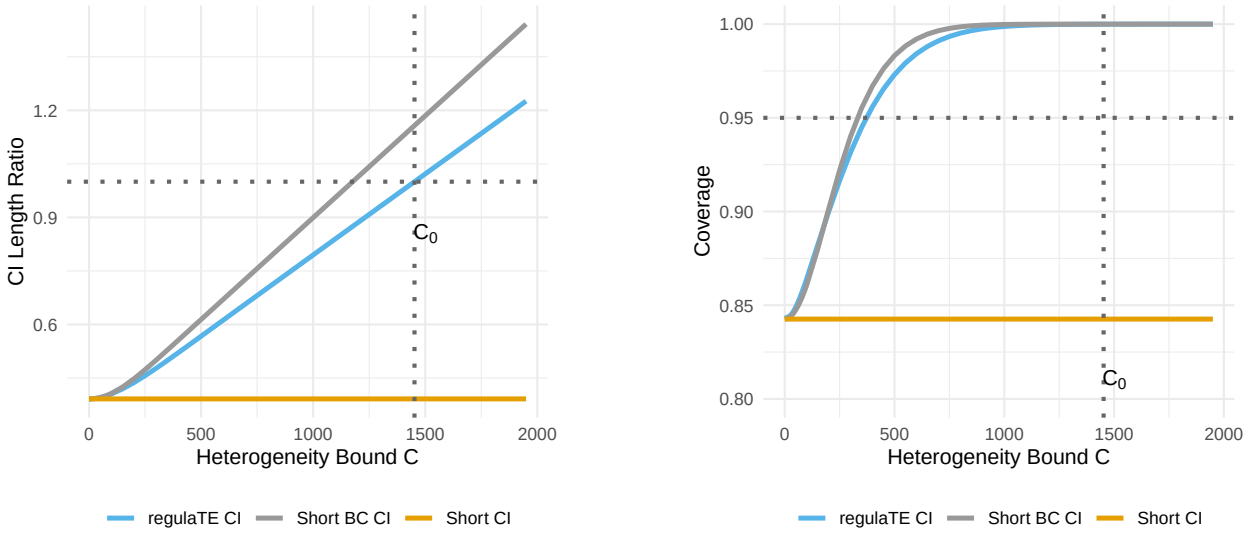
To preserve both heteroskedasticity and treatment effect heterogeneity from the original data of Angrist (1998) in the data-generating process, outcomes are generated by treating the cell-level summary statistics as the true means and standard deviations of earnings and assuming normally distributed earnings within each cell. We treat the cell-level standard deviations as known for the simulation. The true standard deviation of the CATEs is $C_0 = 1452.195$.

Figure 1: Weighted Average Interpretation of `regulaTE` Estimators under Lack of Overlap in DGP Calibrated to Angrist (1998)



Because all covariates are discrete in the data-generating process, the `regulaTE` estimator admits a transparent interpretation as a weighted average of the cell-level treatment effects $\tau(x)$, as characterized in (13). Figure 1 plots the `regulaTE` weights under several values of the heterogeneity bound C against the within-cell treatment variance $p(x)(1-p(x))$. When $C = 0$, the estimator coincides with the short regression, placing relatively high weight on well-overlapped cells and aggressively shrinking contributions from poorly overlapped ones where $p(x)(1-p(x))$ is small. As C increases, the estimator progressively shifts weight toward a more uniform allocation across cells, reflecting the increasing importance of worst-case bias relative to variance. With $C = C_0$, the weights become substantially more even, illustrating how `regulaTE` moves away from the short regression in response to the bias-variance trade-off induced by treatment effect heterogeneity.

Figure 2: Sensitivity of Coverage and CI Length under Lack of Overlap in DGP Calibrated to Angrist (1998)



Notes: "`regulaTE` CI" refers to the bias-aware fixed-length confidence interval based on the `regulaTE` estimate. "Short CI" refers to the CI based on the short regression estimate. "Short BC CI" refers to the bias-corrected short regression CI. Both "`regulaTE` CI" and "Short BC CI" are heteroskedasticity-robust with 95% coverage guarantees under each heterogeneity bound C on the horizontal axis. The ratio of the CI lengths is relative to the length of the `regulaTE` CI under $C = C_0$.

On the horizontal axis of Figure 2, we consider various heterogeneity bounds C and use them both to bias-correct the short-regression CI and to construct the heteroskedasticity-robust `regulaTE` CI. When no correction is applied, the short-regression CI, while quite

short, exhibits substantial undercoverage, reflecting bias induced by omitting treatment effect heterogeneity. As the bound C increases, both the bias-corrected short-regression CI and the `regulaTE` CI achieve coverage close to the nominal level. But across the range of C, the bias-corrected short-regression CI is longer than the `regulaTE` CI, as shown in the left panel. Correct coverage is attained for values of C that are strictly smaller than the true heterogeneity level C_0 because the validity guarantee is derived under worst-case heterogeneity, whereas the data-generating process considered here is less adversarial.

In Appendix B, we illustrate the behavior of `regulaTE` in settings with overlap and compare it with the long regression, which is now well-defined. We also compare with the adaptive estimator of Armstrong et al. (2025), which combines the short and long regression estimators for efficiency gain. `regulaTE` is constructed to adjust the CI based on the user-specified bound C for sensitivity analysis. Therefore, it connects the long and short regression CIs as C increases from zero. The adaptive estimator does not adjust to the user-specified bound C and remains close to the long regression in this DGP, while the bias-corrected short-regression CI is again overly long.

4 Sensitivity analysis and empirical illustrations

Researchers often use the short regression to estimate the ATE under the (frequently implicit) assumption that treatment effect heterogeneity is not too large. Our method provides a formal way to assess this assumption. In this case, rather than taking a definitive stand on C, one can begin with $C = 0$ (which corresponds to using the short regression) and gradually increase C until the results become insignificant.[9] We denote the smallest value of C that renders the estimated treatment effect insignificant as C^*. One can then evaluate whether the corresponding breakdown point C^* is plausible, which is feasible since VCATE is a highly interpretable quantity. This is analogous to the breakdown frontier analysis considered in, e.g., Kline and Santos (2013), Masten and Poirier (2020) and Li and Müller (2021). Our R package provides a plotting functionality that aids such sensitivity analysis.

[9]Note that we take C, rather than C^2, as our sensitivity parameter because it has the same units as the outcome.

4.1 Unconfoundedness

Aizer et al. (2016) evaluate the long-run effects of early 20th-century Mothers' Pension (MP) cash transfers on children's lifetime outcomes, using administrative data. Their study compares children of approved MP applicants to those whose approvals were subsequently reversed, accounting for observed characteristics, to estimate causal effects. To illustrate our method, we focus on child longevity, which is the main outcome of interest. The original study estimates the following short regression as in Aizer et al. (2016, Equation (1)):

$$\log(Age\ at\ Death)_{ifts} = \theta_0 + \theta_1 MP_f + \theta_2 \boldsymbol{X}_{if} + \theta_3 \boldsymbol{Z}_{st} + \boldsymbol{\theta}_c + \boldsymbol{\theta}_t + \varepsilon_{if},$$

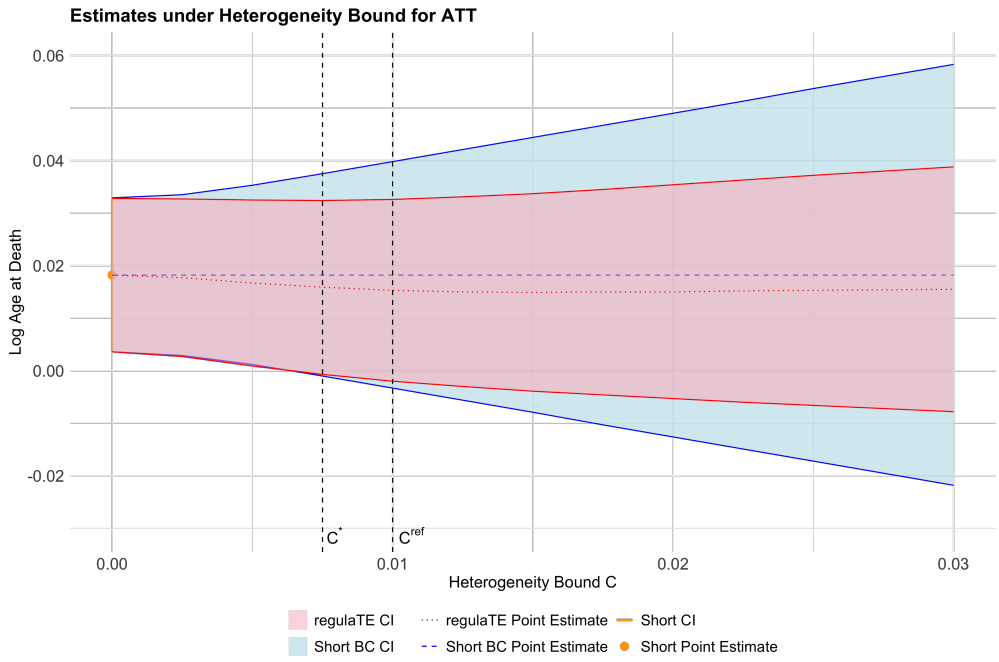
where the outcome is the natural logarithm of the age at death for a given individual i in family f born in year t living in a county c (state s), and the treatment MP_f indicates MP receipt. The authors control linearly for $\boldsymbol{X}$, a vector of relevant family characteristics (marital status, number of siblings, etc.) and child characteristics (year of birth and age at application), and $\boldsymbol{Z}_{st}$, a vector of county-level characteristics in 1910 and state characteristics in the year of application. The binary covariates are $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_t$, which are county and cohort fixed effects, respectively.

We assess the sensitivity of the estimate for ATT, which is the average impact of MP receipt among MP recipients, as it is the most relevant target parameter for program evaluation. Based on the short regression, the ATT is estimated to be 1.82% and is statistically significant at the 5% significance level. Note that the long regression is infeasible because in eleven counties—accounting for about 9% of the sample—all applicants received the MP. This lack of overlap renders the long regression undefined due to collinearity among the interaction terms and county fixed effects.

After reporting their ATT estimates based on the short regression (Aizer et al., 2016, Table 4 Panel A), the authors examine heterogeneous effects of the MP program across subgroups defined by family income, the child's age, and urban residence. Some subgroup estimates are statistically insignificant, and the magnitudes are broadly similar, ranging between 1% and 2% (Aizer et al., 2016, Table 5).

To formalize the robustness checks based on the subgroup analysis originally conducted

Figure 3: Sensitivity Analysis for Aizer et al. (2016)



Notes: "`regulaTE` Point Estimate" refers to the `regulaTE` estimate optimized under each heterogeneity bound C on the horizontal axis. "Short Point Estimate" and "Short BC Point Estimate" both refer to the short regression estimate. "`regulaTE` CI" and "Short BC CI" refer to the bias-aware fixed length confidence interval based on the corresponding estimates, with 95% coverage guarantees valid under each heterogeneity bound C.

by the authors, Figure 3 conducts the sensitivity analysis based on the `regulaTE` CI. We also present the bias-corrected short regression CI for comparison. Standard errors are clustered at the county-level, as in the original analysis. As C increases, the `regulaTE` CI for ATT expands to account for the possibility of greater worst-case bias. Note that the bias-corrected short regression CI is substantially wider, underscoring the efficiency gains from `regulaTE`. The breakdown point is around 0.75% ($C^* = 0.0075$). To put this breakdown frontier value into perspective, note that given an average age at death of 72.44 years, even a 2% increase represents a substantively meaningful effect. This consideration motivates a benchmark on the standard deviation of heterogeneous percentage effects of $C^{\text{ref}} = 1\%$, corresponding to bounding the effects to be between 0% and 2% and applying Popoviciu's inequality on variances. The breakdown point is only marginally smaller than C^{ref}, suggesting the statistically significantly positive impact of MP receipt on child longevity among receiving families is robust to economically meaningful treatment effect heterogeneity to some extent.

4.2 Staggered adoption

The recent literature on staggered adoption designs also emphasizes the bias in common specifications from omitting treatment effect heterogeneity; see Roth et al. (2023) for a review. Consider a staggered adoption setting where we observe a sample of i.i.d. draws $\left\{\{Y_{it}, d_{it}\}_{t=0}^{T}\right\}_{i=1}^{n}$ over $\mathcal{T} = T + 1$ total time periods. Define $e_i := \min\{t : d_{it} = 1\}$ as the period in which unit i first receives treatment. Let $\mathcal{E}$ denote the set of all such first treatment periods, and define a cohort as the set of units treated for the first time in the same period: $\{i : e_i = e\}$.

Under standard assumptions of parallel trends and no anticipation, the DGP for the observed outcome can be written as

$$Y_{it} = \alpha_e + \lambda_t + \sum_{\ell \geq 0,\, e \in \mathcal{E}} \tau_{e,\ell} \cdot d_{it} \cdot \mathbf{1}\{e_i = e, t - e_i = \ell\} + \varepsilon_{it}, \tag{17}$$

where α_e and λ_t are cohort and time fixed effects respectively. The framework developed earlier for unconfoundedness settings in (1) assumes $\tau(\cdot)$ is a linear function of the confounders x_i. Here $\tau(\cdot)$ is a linear function of group and relative time indicators, rather than

the confounders. Specifically, $\tau_{e,\ell} = \mathbb{E}[Y_{i,e+\ell}(1) - Y_{i,e+\ell}(0) \mid e_i = e]$ is the conditional average treatment effect (CATE) for cohort e at relative time ℓ, which can vary across cohorts and over time in an unrestricted fashion. Therefore, the original framework continues to hold for the following "short" and "long" regressions.

Rather than the unrestricted (17), researchers often estimate a simpler specification also known as the "static" two-way fixed effects (TWFE) regression

$$Y_{it} = \alpha_e + \lambda_t + \beta_{\text{short}} \cdot d_{it} + \varepsilon_{it} \tag{18}$$

which omits the interactions between d_{it} and cohort and relative-time indicators. Analogous to the short regression in the unconfoundedness setting, as argued in de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021), the "static" specification (18) implicitly assumes constant effects across cohorts and over time (i.e., $C = 0$). When the effects indeed vary, the "static" specification (18) can be severely biased for a class of reasonable estimands due to negative weighting of $\tau_{e,\ell}$ for large ℓ.

One reasonable estimand in this setting is the average effect over treated cohorts (ATT), defined as

$$\text{ATT} = \frac{\sum_{t\in\mathcal{T}} \sum_{e\in\mathcal{E}} \mathbb{P}_n\{d_{it} = 1\} \cdot \mathbb{P}_n\{e_i = e \mid d_{it} = 1\} \cdot \tau_{e,t-e}}{\sum_{t\in\mathcal{T}} \mathbb{P}_n\{d_{it} = 1\}},$$

where $\mathbb{P}_n$ denotes empirical frequencies. That is, the ATT is a weighted average of the cohort-by-event-time effects $\tau_{e,t-e}$, where the weights reflect the empirical distribution of treated observations across cohort-time cells. To recover this estimand, we can reparameterize (17) as the "long" regression

$$Y_{it} = \alpha_e + \lambda_t + d_{it}\beta^{\text{ATT}}_{\text{long}} + (d_{it} \cdot \tilde{x}_{it})' \, \delta + \varepsilon_{it}$$

where $\tilde{x}_{it}$ collects all but one of the re-centered cohort and relative time indicators:

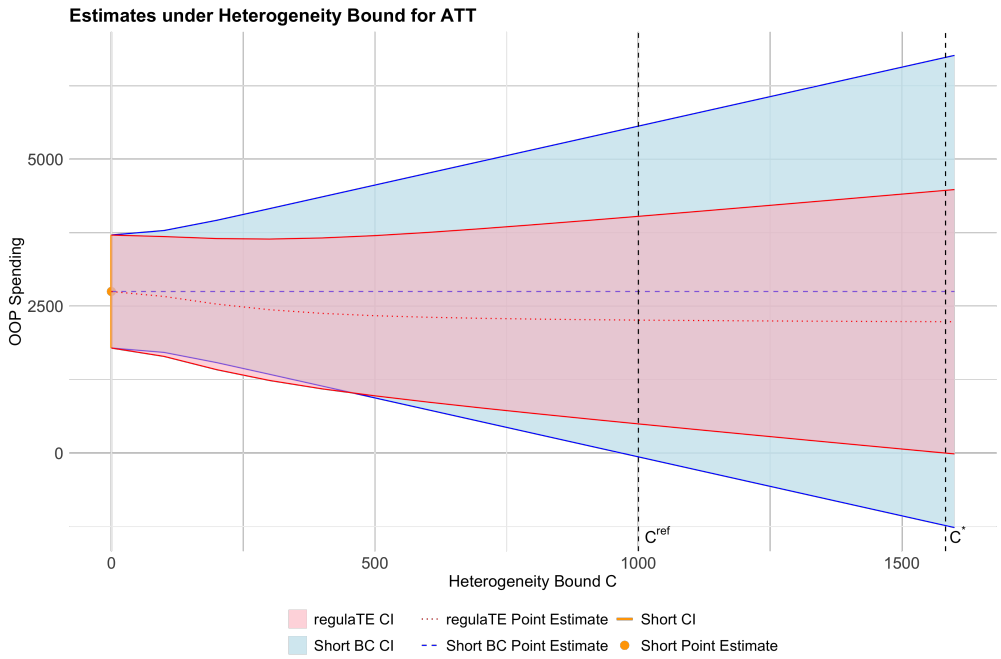
$$\tilde{x}_{it} = \begin{pmatrix} \vdots \\ \mathbf{1}\{e_i = e,\, t - e_i = \ell\} - \dfrac{\sum_{t\in\mathcal{T}} \mathbb{P}_n\{d_{it} = 1\}\mathbb{P}_n\{e_i = e \mid d_{it} = 1\}\mathbf{1}\{t - e = \ell\}}{\sum_{t\in\mathcal{T}} \mathbb{P}_n\{d_{it} = 1\}} \\ \vdots \end{pmatrix}.$$

This long regression coincides with the extended TWFE specification from Wooldridge (2025). However, this long regression can be noisy or even infeasible in practice. Estimation of (17) corresponds to aggregating cohort- and time-specific DID estimators $\hat{\tau}_{e,t-e}$ for each $\tau_{e,\ell}$, which can be noisy when few units are treated at a given time. Moreover, if all units are treated after time t, the required counterfactuals are not identified, similar to a lack of overlap in cross-sectional settings, and the ATT is no longer identified without further assumptions on treatment effect heterogeneity. To address this, we impose a bound $C > 0$ on the variance of treatment effect heterogeneity, restricting the deviation of $\tau_{e,t-e}$ from the ATT. Under this assumption, we can construct valid confidence intervals using our `regulaTE` procedure, thereby extending the sensitivity analysis to the staggered adoption setting.

To illustrate, we revisit the empirical application in Sun and Abraham (2021), which builds on Dobkin et al. (2018) to estimate the average effect of an unexpected hospitalization on out-of-pocket medical spending among adults who experienced at least one such hospitalization between waves 8 and 11 of the Health and Retirement Study (HRS). Each wave corresponds to approximately two calendar years. Using a balanced panel from wave 7 through wave 11 ($T = 5$), all individuals are treated in the final period ($t = 5$), so the ATT from waves 8 to 11 is not point identified and the long regression (17) is not well defined. Nevertheless, the "static" specification delivers a statistically significant and positive estimate, as shown in Figure 4.

To evaluate the robustness of this estimate to treatment effect heterogeneity, Figure 4 reports sensitivity analysis based on the `regulaTE` CIs. Again, we present the bias-corrected short regression CIs as well for comparison. Standard errors are clustered at the individual level, as in the original analysis of Dobkin et al. (2018). The CIs based on `regulaTE` remain significant up to a breakdown value of $C^* = \$1{,}583$. To interpret this breakdown value, consider the original analysis in Dobkin et al. (2018), which focuses on heterogeneity over time. They report an increase in out-of-pocket spending of roughly \$3,000 in the first year after hospitalization and \$1,000 by the third year. Under the assumption that heterogeneity operates primarily over time, treating these magnitudes as rough bounds and applying Popoviciu's inequality on variances, we get a reference treatment effect heterogeneity value

Figure 4: Sensitivity Analysis for Dobkin et al. (2018)



Notes: "`regulaTE` Point Estimate" refers to the `regulaTE` estimate optimized under each heterogeneity bound C on the horizontal axis. "Short Point Estimate" and "Short BC Point Estimate" both refer to the short/static regression (18) estimate. "`regulaTE` CI" and "Short BC CI" refer to the bias-aware fixed length confidence interval based on the corresponding estimates, with 95% coverage guarantees valid under each heterogeneity bound C.

of $C^{\text{ref}} = \$1{,}000$. Therefore, using sensitivity analysis based on the `regulaTE` CIs, we find the statistically significant average increase in out-of-pocket spending due to unexpected hospitalization remains robust to substantial treatment effect heterogeneity. In contrast, the bias-corrected short regression CIs widen substantially, and a sensitivity analysis based on them would suggest a lack of robustness to treatment effect heterogeneity, potentially overstating the fragility of the original results.

5 Conclusion

Many specifications commonly used in empirical research implicitly restrict treatment effects to be constant, often favoring precision at the expense of robustness. While such a restriction yields narrower confidence intervals, the resulting estimates may be substantially biased in the presence of heterogeneity. This paper develops a sensitivity analysis based on the

proposed `regulaTE` CIs by varying the bound on treatment effect heterogeneity.

There are several directions for future work. One natural extension is to consider alternative forms of restrictions on treatment effect heterogeneity. While variance bounds naturally capture the dispersion in heterogeneous effects, one could instead impose an absolute bound on the deviation of individual treatment effects from the average effect if such prior knowledge is available. Another promising direction is to generalize the framework to accommodate multiple treatments or layered sources of heterogeneity, as in Goldsmith-Pinkham et al. (2024) and Sun and Abraham (2021).

References

Aizer, A., S. Eli, J. Ferrie, and A. Lleras-Muney (2016). The long-run impact of cash transfers to poor families. *American Economic Review 106*(4), 935–71.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica 66*(2), 249–288.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Armstrong, T. B., P. Kline, and L. Sun (2025). Adapting to misspecification. *Forthcoming at Econometrica*.

Armstrong, T. B. and M. Kolesár (2018). Optimal inference in a class of regression models. *Econometrica 86*(2), 655–683.

Armstrong, T. B. and M. Kolesár (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica 89*(3), 1141–1177.

Armstrong, T. B., M. Kolesár, and S. Kwon (2023). Bias-aware inference in regularized regression models. *arXiv:2012.14823 [econ.EM]*.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased

inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Bobonis, G. J., L. R. Cámara Fuertes, and R. Schwabe (2016). Monitoring corruptible politicians. *American Economic Review 106*(8), 2371–2405.

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

de Chaisemartin, C. (2024). Trading-off bias and variance in stratified experiments and in staggered adoption designs, under a boundedness condition on the magnitude of the treatment effect. *arXiv:2105.08766 [econ.EM]*.

de Chaisemartin, C. and A. Deeb (2024). Estimating treatment-effect heterogeneity across sites, in multi-site randomized experiments with few units per site. *arXiv:2405.17254 [econ.EM]*.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Dobkin, C., A. Finkelstein, R. Kluender, and M. J. Notowidigdo (2018). The economic consequences of hospital admissions. *American Economic Review 108*(2), 308–52.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics 22*(1), 238–270.

Favara, G. and J. Imbs (2015). Credit supply and the price of housing. *American economic review 105*(3), 958–992.

Gibbons, C. E., J. C. S. Serrato, and M. B. Urbancic (2019). Broken or fixed effects? *Journal of Econometric Methods 8*(1), 1–12.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2024). Contamination bias in linear regressions. *American Economic Review 114*(12), 4015–4051.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Heiler, P. and E. Kazak (2021). Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics 222*(2), 1083–1108.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature 47*(1), 5–86.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Kline, P. and A. Santos (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the us wage structure. *Quantitative Economics 4*(2), 231–267.

Lechner, M. (2008). A note on the common support problem in applied evaluation studies. *Annales d'Économie et de Statistique*, 217–235.

Lee, S. and M. Weidner (2021). Bounding treatment effects by pooling limited information across observations. *arXiv preprint arXiv:2111.05243*.

Levy, J., M. van der Laan, A. Hubbard, and R. Pirracchio (2021). A fundamental measure of treatment effect heterogeneity. *Journal of Causal Inference 9*(1), 83–108.

Li, C. and U. K. Müller (2021). Linear regression with many controls of limited explanatory power. *Quantitative Economics 12*(2), 405–442.

Li, K.-C. (1982). Minimaxity of the method of regularization of stochastic processes. *The Annals of Statistics 10*(3), 937–942.

Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics 25*(6), 2547–2554.

Ma, X. and J. Wang (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association 115*(532), 1851–1860.

Manski, C. F. and J. V. Pepper (2018). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics 100*(2), 232–244.

Martinez-Bravo, M. (2014). The role of local officials in new democracies: Evidence from indonesia. *American Economic Review 104*(4), 1244–1287.

Masten, M. A. and A. Poirier (2020). Inference on breakdown frontiers. *Quantitative Economics 11*(1), 41–111.

Michalopoulos, S. and E. Papaioannou (2016). The long-run effects of the scramble for africa. *American Economic Review 106*(7), 1802–1848.

Noack, C. and C. Rothe (2024). Bias-aware inference in fuzzy regression discontinuity designs. *Econometrica 92*(3), 687–711.

Poirier, A. and T. Słoczyński (2024). Quantifying the internal validity of weighted estimands. *arXiv:2404.14603 [econ.EM]*.

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica 85*(2), 645–660.

Sanchez-Becerra, A. (2023). Robust inference for the treatment effect variance in experiments using machine learning. *arXiv:2306.03363 [econ.EM]*.

Sasaki, Y. and T. Ura (2022). Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory 38*(1), 66–112.

Słoczyński, T. (2022). Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights. *The Review of Economics and Statistics 104*(3), 501–509.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics 225*(2), 175–199.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2 ed.). Cambridge, MA, USA: MIT Press.

Wooldridge, J. M. (2025). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Empirical Economics 69*, 2545–2587.

Xu, G. (2018). The costs of patronage: Evidence from the british empire. *American Economic Review 108*(11), 3170–3198.

A Details and proofs

A.1 Details for Section 3.1

Lemma A.1. *The short regression estimator is $\hat{\beta}_{short} = a_s'Y$, where $a_s = (D'H_XD)^{-1}H_XD$. Under model (6) and the assumption that VCATE is bounded by C^2, its worst-case bias is given by $\overline{\text{bias}}_{a_s,C} = C \cdot \sqrt{a_s'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'a_s}$.*

Proof for Lemma A.1. The expression for a_s holds by FWL. The worst case bias is

$$\begin{aligned}\overline{\text{bias}}_{a_s,C} &:= \sup_{(\beta,\delta',\gamma')'\in\Theta_C} \mathbb{E}[a_s'Y] - \beta \\ &= \sup_{(\beta,\delta',\gamma')'\in\Theta_C} a_s'D\beta + a_s'(D \circ \tilde{X}_{-1})\delta + a_s'X\gamma - \beta \\ &= \sup_{\delta'V_x\delta\leq C^2} a_s'(D \circ \tilde{X}_{-1})\delta,\end{aligned} \tag{A.1}$$

where the last equality holds because $a_s'D = 1$ and $a_s'X = 0$. We optimize a linear objective function subject to a quadratic constraint of $\delta'V_x\delta \leq C^2$. Assuming the constraint binds, KKT condition gives the solution:

$$\delta = V_x^{-1}(D \circ \tilde{X}_{-1})'a_s\sqrt{\frac{C^2}{a_s'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'a_s}}$$

and plugging in the solution gives the expression for the worst-case bias. □

Lemma A.2 (Two-step representation of generalized ridge regression)**.** *Consider a generalized ridge regression estimator*

$$\min_{\beta,\gamma}\|Y - D\beta - W\gamma\|^2 + \gamma'A\gamma$$

where A is positive semidefinite and $(W'W + A)$ is invertible. The solution β^* is equivalent to

$$\beta^* = \frac{(D - W\tilde{\gamma})'Y}{(D - W\tilde{\gamma})'D}$$

where $\tilde{\gamma}$ solves

$$\min_{\gamma}\|D - W\gamma\|^2 + \gamma'A\gamma.$$

Proof for Lemma A.2. Let (β^*, γ^*) denote the solution to the generalized ridge regression. The first-order condition for γ gives $\gamma^* = (W'W + A)^{-1} W' (Y - D\beta^*)$. Plugging this into the first-order condition for β gives

$$D' \left(Y - W (W'W + A)^{-1} W'Y \right) = D' \left(D - W (W'W + A)^{-1} W'D \right) \beta^*.$$

Note the LHS can be written as

$$D'Y - D'W (W'W + A)^{-1} W'Y = \left(D - W (W'W + A)^{-1} W'D \right)' Y.$$

At the same time, we note $(W'W + A)^{-1} W'D$ solves

$$\min_{\gamma} \|D - W\gamma\|^2 + \gamma' A \gamma.$$

□

Proof of Theorem 1. Let $(\pi_{1,\lambda}, \pi_{2,\lambda})$ be the ridge coefficients solving

$$\min_{\pi_1, \pi_2} \|D - X\pi_1 - (D \circ \tilde{X}_{-1})\pi_2\|^2/n + \lambda \, \pi_2' V_x \pi_2.$$

Following the derivation of (10), the weights of the generalized ridge regression coefficient estimator are based on

$$\tilde{D}_\lambda := D - X\pi_{1,\lambda} - (D \circ \tilde{X}_{-1})\pi_{2,\lambda}, \qquad a_\lambda := \frac{\tilde{D}_\lambda}{\tilde{D}_\lambda' D}.$$

The first-order condition with respect to $\pi_{1,\lambda}$ and $\pi_{2,\lambda}$ are

$$X'\tilde{D}_\lambda = 0, \ (D \circ \tilde{X}_{-1})'\tilde{D}_\lambda/n = \lambda V_x \pi_{2,\lambda}. \tag{A.2}$$

Using the definition of a_λ, we have $X'a_\lambda = 0$ and $D'a_\lambda = 1$. Therefore, the derivation for

the worst-case bias in (A.1) still applies and the term within the squared root is equal to

$$a_\lambda'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'a_\lambda = \frac{\tilde{D}_\lambda'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'\tilde{D}_\lambda}{(\tilde{D}_\lambda' D)^2}.$$

Applying (A.2),

$$\tilde{D}_\lambda'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'\tilde{D}_\lambda = n^2\lambda^2\, \pi_{2,\lambda}' V_x \pi_{2,\lambda},$$

and notice that $\tilde{D}_\lambda' D > 0$ because

$$\begin{aligned}\tilde{D}_\lambda' D &= \tilde{D}_\lambda'\left(\tilde{D}_\lambda + X\pi_{1,\lambda} + (D \circ \tilde{X}_{-1})\pi_{2,\lambda}\right) = \tilde{D}_\lambda'\tilde{D}_\lambda + \tilde{D}_\lambda'(D \circ \tilde{X}_{-1})\pi_{2,\lambda} \\ &= \tilde{D}_\lambda'\tilde{D}_\lambda + n\lambda\, \pi_{2,\lambda}' V_x \pi_{2,\lambda} > 0.\end{aligned}$$

Therefore

$$\sqrt{a_\lambda'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'a_\lambda} = \frac{n\lambda\sqrt{\pi_{2,\lambda}' V_x \pi_{2,\lambda}}}{\tilde{D}_\lambda' D}. \tag{A.3}$$

Consider the formula

$$\frac{n\lambda\sqrt{\pi_{2,\lambda}' V_x \pi_{2,\lambda}}}{\tilde{D}_\lambda' D} = \frac{n\lambda\pi_{2,\lambda}' V_x \pi_{2,\lambda}}{\tilde{D}_\lambda' D\sqrt{\pi_{2,\lambda}' V_x \pi_{2,\lambda}}}$$

Next, using (A.2) in the reverse direction yields

$$n\lambda\pi_{2,\lambda}' V_x \pi_{2,\lambda} = \tilde{D}_\lambda'(D \circ \tilde{X}_{-1})\pi_{2,\lambda} = \tilde{D}_\lambda'(X\pi_{1,\lambda} + (D \circ \tilde{X}_{-1})\pi_{2,\lambda}).$$

Substituting this identity into (A.3) and using the definition of a_λ gives

$$\sqrt{a_\lambda'(D \circ \tilde{X}_{-1})V_x^{-1}(D \circ \tilde{X}_{-1})'a_\lambda} = \frac{a_\lambda'(X\pi_{1,\lambda} + (D \circ \tilde{X}_{-1})\pi_{2,\lambda})}{\sqrt{\pi_{2,\lambda}' V_x \pi_{2,\lambda}}},$$

which implies the worst-case bias is

$$\overline{\text{bias}}_{a_\lambda, C} = C\frac{a_\lambda'(X\pi_{1,\lambda} + (D \circ \tilde{X}_{-1})\pi_{2,\lambda})}{\sqrt{\pi_{2,\lambda}' V_x \pi_{2,\lambda}}}. \tag{A.4}$$

We now map our setting to the framework of Armstrong et al. (2023), Theorem 2.1. In their notation, the estimand is the linear functional $L(\theta) = \beta$, the observed signal is $f(\theta) = D\beta + X\gamma + (D \circ \tilde{X}_{-1})\delta$, the nuisance parameters are $(\gamma', \delta')'$, and the penalty is $\text{Pen}((\gamma', \delta')') = \sqrt{\delta' V_x \delta}$, which is a seminorm. The constraint set Θ_C corresponds to $\{\theta : \text{Pen}((\gamma', \delta')') \leq C\}$ with β unrestricted. Under this identification, Theorem 2.1 of Armstrong et al. (2023) implies that the weights a_λ, which arise from the penalized regression (9) with penalty parameter λ, solve the bias-constrained variance-minimization problem (16) with $B = \overline{\text{bias}}_{a_\lambda, C}$ as in (A.4). Finally, since the constraint in (16) must bind, $\overline{\text{bias}}_{a_\lambda, C}$ is indeed the worst-case bias of $\hat{\beta}_\lambda$. □

A.1.1 Specialized results under discrete covariates

Lemma A.3. *Suppose x_i is generated from discrete covariate values. Specifically, let J denote a matrix that replaces the first row of an identity matrix I with a row of ones. Let $\bar{x}_i$ denote the vector where each entry is an indicator for observation i's covariate taking the corresponding value. Concretely, let $c(i) \in \{1, 2, ..., k+1\}$ denote the covariate value of observation i, with 1 as the reference category. Then $x_i = J\bar{x}_i$ for*

$$J = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \bar{x}_i = \begin{pmatrix} \mathbf{1}\{c(i) = 1\} \\ \mathbf{1}\{c(i) = 2\} \\ \vdots \\ \mathbf{1}\{c(i) = k+1\} \end{pmatrix},$$

which removes the indicator for the reference value, and replaces it with an intercept. Then the solution to the generalized ridge regression is equivalent to

$$\beta_\lambda = \frac{\mathbb{E}_n\left[(d_i - p(x_i))\left(1 - \widetilde{x}'_{i,-1}\pi_{2,\lambda}\right) Y_i\right]}{\mathbb{E}_n\left[(d_i - p(x_i))\left(1 - \widetilde{x}'_{i,-1}\pi_{2,\lambda}\right) d_i\right]}$$

where $p(x_j) = \mathbb{E}_n\left[d_i \mathbf{1}\{c(i) = c(j)\}\right] / \mathbb{E}_n\left[\mathbf{1}\{c(i) = c(j)\}\right]$ and $\pi_{2,\lambda}$ is the solution to the following objective function

$$\min_{\pi_2} \; n\mathbb{E}_n\left[\left((d_i - p(x_i))\left(1 - \widetilde{x}'_{i,-1}\pi_2\right)\right)^2\right] + \lambda \mathbb{E}_n\left[(\widetilde{x}'_{i,-1}\pi_2)^2\right]$$

Proof for Lemma A.3. By Lemma A.2 we have

$$\beta_\lambda = \frac{\mathbb{E}_n\left[\left(d_i - x_i'\pi_{1,\lambda} - d_i\widetilde{x}_{i,-1}'\pi_{2,\lambda}\right) Y_i\right]}{\mathbb{E}_n\left[\left(d_i - x_i'\pi_{1,\lambda} - d_i\widetilde{x}_{i,-1}'\pi_{2,\lambda}\right) d_i\right]}$$

where $\pi_{1,\lambda}$ and $\pi_{2,\lambda}$ solve the following regularized propensity score regression:

$$\min_{\pi_1,\pi_2} \; n\mathbb{E}_n\left[\left(d_i - x_i'\pi_1 - d_i\widetilde{x}_{i,-1}'\pi_2\right)^2\right] + \lambda\mathbb{E}_n\left[(\widetilde{x}_{i,-1}'\pi_2)^2\right].$$

Since π_1 is not regularized, and x_i is based on discrete covariates, for any value of π_2, we can concentrate out π_1. Because $(1-\widetilde{x}_{i,-1}'\pi_2)$ is constant within each cell, the OLS projection of $d_i(1-\widetilde{x}_{i,-1}'\pi_2)$ onto x_i equals $p(x_i)(1-\widetilde{x}_{i,-1}'\pi_2)$, giving: $\min_{\pi_1}\ \mathbb{E}_n\left[\left(d_i - x_i'\pi_1 - d_i\widetilde{x}_{i,-1}'\pi_2\right)^2\right] = \min_{\pi_1}\ \mathbb{E}_n\left[\left(d_i\left(1-\widetilde{x}_{i,-1}'\pi_2\right) - x_i'\pi_1\right)^2\right] = \mathbb{E}_n\left[\left(\left(d_i - p(x_i)\right)\left(1-\widetilde{x}_{i,-1}'\pi_2\right)\right)^2\right]$.

□

The following lemma shows that the estimand weights in (13) admit a closed form when the covariates are generated by saturating discrete variables.

Lemma A.4 (Closed-form weights under discrete covariates)**.** *Suppose x_i is generated by saturating discrete covariates as in Lemma A.3, with $k+1$ cells, each containing at least one treated and one untreated unit. Then for every $\lambda > 0$, the estimand in (13) equals (14). In particular, the weights are non-negative and sum to one.*

Proof. Write $\sigma^2(x_i) = p(x_i)(1-p(x_i))$ for the conditional variance of treatment at covariate value x_i. By Lemma A.3, concentrating out π_1 from the penalized propensity score objective (11) under discrete covariates reduces it to a function of π_2 alone, and the population-level objective whose minimizer determines the estimand weights is

$$\mathbb{E}_n\left[\sigma^2(x_i)(1-\widetilde{x}_{i,-1}'\pi)^2 + \lambda(\widetilde{x}_{i,-1}'\pi)^2\right],$$

using $\pi' V_x \pi = \mathbb{E}_n[(\widetilde{x}_{i,-1}'\pi)^2]$. Since all observations in the same cell share the same covariate value, $\widetilde{x}_{i,-1}'\pi$ takes a common value within each cell. Moreover, $\mathbb{E}_n[\widetilde{x}_{i,-1}] = 0$ implies $\mathbb{E}_n[\widetilde{x}_{i,-1}'\pi] = 0$ for any π. Conversely, every assignment of $k+1$ cell-level values satisfying this weighted-mean-zero property is attainable. To see this, let $(r_1, \ldots, r_{k+1})$ satisfy $\mathbb{E}_n[r_{c(i)}] = 0$,

where $c(i) \in \{1, \ldots, k+1\}$ denotes the cell of observation i. Setting $\pi_{t-1} = r_t - r_1$ for $t = 2, \ldots, k+1$, we have $\widetilde{x}'_{i,-1}\pi = r_{c(i)}$ for all i, because the t-th component of $\widetilde{x}_{i,-1}$ equals $\mathbf{1}\{c(i) = t+1\} - \mathbb{E}_n[\mathbf{1}\{c(i) = t+1\}]$.

Since the map $\pi \mapsto (\widetilde{x}'_{i,-1}\pi)_{i=1}^n$ is determined by the $k+1$ cell-level values and the correspondence above is a bijection between $\pi \in \mathbb{R}^k$ and the k-dimensional set of weighted-mean-zero cell-level vectors, we may minimize the objective directly over these $k+1$ values subject to the single constraint $\mathbb{E}_n[\widetilde{x}'_{i,-1}\pi] = 0$. Since each coefficient $\sigma^2(x_i) + \lambda$ is strictly positive, the objective is strictly convex in these cell-level values. Hence, there exists a Lagrangian multiplier η_λ such that, for every observation i,

$$-\sigma^2(x_i)(1 - \widetilde{x}'_{i,-1}\pi_{2,\lambda}) + \lambda\, \widetilde{x}'_{i,-1}\pi_{2,\lambda} + \eta_\lambda = 0,$$

which gives

$$\widetilde{x}'_{i,-1}\pi_{2,\lambda} = \frac{\sigma^2(x_i) - \eta_\lambda}{\sigma^2(x_i) + \lambda}.$$

Imposing $\mathbb{E}_n[\widetilde{x}'_{i,-1}\pi_{2,\lambda}] = 0$ and solving for η_λ:

$$\eta_\lambda = \frac{\mathbb{E}_n\big[\sigma^2(x_i)/(\sigma^2(x_i) + \lambda)\big]}{\mathbb{E}_n\big[1/(\sigma^2(x_i) + \lambda)\big]}.$$

This is a weighted average of $\sigma^2(x_1), \ldots, \sigma^2(x_n)$ with positive weights proportional to $1/(\sigma^2(x_i) + \lambda)$, so $\eta_\lambda \geq 0$. It follows that

$$1 - \widetilde{x}'_{i,-1}\pi_{2,\lambda} = \frac{\lambda + \eta_\lambda}{\sigma^2(x_i) + \lambda} > 0 \qquad (\lambda > 0),$$

since $\lambda + \eta_\lambda > 0$ and $\sigma^2(x_i) + \lambda > 0$. Substituting into (13), the weight on observation i's CATE is proportional to $\sigma^2(x_i)(\lambda + \eta_\lambda)/(\sigma^2(x_i) + \lambda)$. The common factor $(\lambda + \eta_\lambda)$ cancels between numerator and denominator, yielding (14). □

Lemma A.5 (Convergence as $\lambda \to 0$)**.** *Assume that $S(\theta)$ is quadratic and continuous in $\theta \in \mathbb{R}^{p_\theta}$, and is strongly convex with unique minimizer θ^*. Moreover, assume $P(\theta, \psi) \geq 0$ is*

quadratic and continuous in θ and $\psi \in \mathbb{R}^{p_\psi}$. Let the penalized objective be

$$F_\lambda(\theta, \psi) \;=\; S(\theta) \;+\; \lambda\, P(\theta, \psi), \qquad \lambda \geq 0.$$

Let $(\hat{\theta}_\lambda, \hat{\psi}_\lambda)$ be the minimizer of F_λ for $\lambda > 0$. Then $\hat{\theta}_\lambda \to \theta^*$ as $\lambda \to 0$.

Proof. Let $\varepsilon > 0$ be arbitrary. Since $S(\theta)$ is quadratic and strongly convex, there exists a constant $c > 0$ such that

$$S(\theta) \;\geq\; S(\theta^*) + c\|\theta - \theta^*\|^2 \quad \text{whenever} \quad \|\theta - \theta^*\|^2 \leq \epsilon.$$

By continuity of P on compact domain, there exists $M < \infty$ such that

$$\sup_{\|\theta - \theta^*\| \leq \varepsilon} P(\theta, 0) \;\leq\; M.$$

Now set $\delta := c\,\varepsilon^2/2$. For λ small enough so that $\lambda M \leq \delta$, we have

$$\inf_{(\theta,\psi) \in \mathbb{R}^{p_\theta + p_\psi}} F_\lambda(\theta, \psi) \;\leq\; \inf_{\|\theta - \theta^*\| \leq \varepsilon} F_\lambda(\theta, 0) \;\leq\; \inf_{\|\theta - \theta^*\| \leq \varepsilon} S(\theta) + \lambda M \;\leq\; S(\theta^*) + \delta.$$

Suppose $\|\hat{\theta}_\lambda - \theta^*\| > \varepsilon$. Then we reach a contradiction as

$$F_\lambda(\hat{\theta}_\lambda, \hat{\psi}_\lambda) \;\geq\; S(\hat{\theta}_\lambda) \;\geq\; S(\theta^*) + c\varepsilon^2 \;=\; S(\theta^*) + 2\delta \;>\; \inf_{(\theta,\psi) \in \mathbb{R}^{p_\theta + p_\psi}} F_\lambda(\theta, \psi).$$

Therefore, for all sufficiently small λ, every global minimizer $(\hat{\theta}_\lambda, \hat{\psi}_\lambda)$ of F_λ must satisfy $\|\hat{\theta}_\lambda - \theta^*\| \leq \varepsilon$. Since $\varepsilon > 0$ was arbitrary, $\hat{\theta}_\lambda \to \theta^*$.

□

Lemma A.6 (Behavior under no overlap)**.** *Suppose x_i is generated by saturating discrete covariates as in Lemma A.3, so that observations can be grouped into $j = 1, \ldots, k+1$ mutually exclusive cells. Let $\mathcal{J}_1$ denote the set of cells with all-treated observations ($d_i = 1$ for all i in the cell) and $\mathcal{J}_0$ denote the set of cells with all-untreated observations ($d_i = 0$ for all i in the cell). Assume overlap holds for all other cells.*

Then the generalized ridge regression estimator satisfies

$$\lim_{\lambda \to 0} \hat{\beta}_\lambda = \hat{\beta}_{trim}$$

where $\hat{\beta}_{trim}$ is the long regression estimator computed on the subsample of groups with overlap, i.e., the sample with groups $\{1, \ldots, k+1\} \setminus (\mathcal{J}_0 \cup \mathcal{J}_1)$.

Proof. With saturated discrete covariates, the generalized ridge problem can be reparameterized as in the proof for Lemma A.3,

$$\min_{\bar{\gamma}, \bar{\tau}} \mathbb{E}_n\big[(y_i - \bar{\gamma}'\bar{x}_i - \bar{\tau}'\bar{x}_i d_i)^2\big] + \lambda\, \mathbb{E}_n\big[(\bar{\tau}'\bar{x}_i - \mathbb{E}_n[\bar{\tau}'\bar{x}_i])^2\big],$$

where $\bar{x}_i$ is the vector of cell indicators as in Lemma A.3. With this reparameterization, the ridge regression coefficient estimator that we are interested in is $\hat{\beta}_\lambda = \mathbb{E}_n[\hat{\bar{\tau}}_\lambda' \bar{x}_i]$ where $\hat{\bar{\tau}}_\lambda$ minimizes the above ridge problem when the penalty is λ. Write $f_j = \mathbb{E}_n[\mathbf{1}\{c(i) = j\}]$ for the cell share.

For any cell $j \in \mathcal{J}_0$ (all-untreated), $d_i = 0$ for all i in the cell, so $\bar{x}_{ij} d_i = 0$. Hence $\bar{\tau}_j$ does not enter the least-squares term.

For any cell $j \in \mathcal{J}_1$ (all-treated), $d_i = 1$ for all i in the cell, so $\bar{x}_{ij} d_i = \bar{x}_{ij}$. Writing $\tilde{\bar{\gamma}}_j = \bar{\gamma}_j + \bar{\tau}_j$, $\bar{\tau}_j$ again does not enter the squared-error term separately.

Denote $\mathcal{N} = \mathcal{J}_0 \cup \mathcal{J}_1$ to be the set of cells without overlap. As $\lambda \to 0$, by Lemma A.5 the limit of the estimators $\lim_{\lambda \to 0} \hat{\bar{\tau}}_{k,\lambda}$ for $k \notin \mathcal{N}$ are the OLS estimators that minimize

$$\min_{\bar{\gamma}, \bar{\tau}_{k \notin \mathcal{N}}} \mathbb{E}_n \left[\left(y_i - \bar{\gamma}'\bar{x}_i - \sum_{k \notin \mathcal{N}} \bar{\tau}_k \bar{x}_{ik} d_i \right)^2 \right]. \tag{A.5}$$

The OLS problem (A.5) gives the OLS coefficient estimators

$$\begin{aligned}
\hat{\bar{\tau}}_k &= \frac{\mathbb{E}_n[y_i \bar{x}_{ik} d_i]}{\mathbb{E}_n[\bar{x}_{ik} d_i]} - \frac{\mathbb{E}_n[y_i \bar{x}_{ik}(1 - d_i)]}{\mathbb{E}_n[\bar{x}_{ik}(1 - d_i)]} \text{ for } k \notin \mathcal{N} \\
\hat{\bar{\gamma}}_k &= \frac{\mathbb{E}_n[y_i \bar{x}_{ik}(1 - d_i)]}{\mathbb{E}_n[\bar{x}_{ik}(1 - d_i)]} \text{ for } k \notin \mathcal{N} \\
\hat{\bar{\gamma}}_j &= \frac{\mathbb{E}_n[y_i \bar{x}_{ij}]}{\mathbb{E}_n[\bar{x}_{ij}]} \text{ for } j \in \mathcal{N}
\end{aligned}$$

For any $j \in \mathcal{N}$, the coefficient $\bar{\tau}_j$ does not enter the squared-error term, so the ridge objective depends on $\bar{\tau}_j$ only through the penalty $\mathbb{E}_n[(\bar{\tau}'\bar{x}_i - \mathbb{E}_n[\bar{\tau}'\bar{x}_i])^2]$. Writing $\sigma_{jk} := \mathbb{E}_n[\bar{x}_{ij}\bar{x}_{ik}] - \mathbb{E}_n[\bar{x}_{ij}]\mathbb{E}_n[\bar{x}_{ik}]$ for the sample covariance between cell indicators j and k, this penalty can be expanded as $\sum_{j,k} \bar{\tau}_j \bar{\tau}_k \sigma_{jk}$. Taking the derivative with respect to $\bar{\tau}_j$ for $j \in \mathcal{N}$, and substituting the limit values $\hat{\bar{\tau}}_k$ for $k \notin \mathcal{N}$, gives the linear system:

$$\bar{\tau}_j \, \sigma_{jj} + \sum_{k \in \mathcal{N},\, k \neq j} \bar{\tau}_k \, \sigma_{jk} = -\sum_{k \notin \mathcal{N}} \bar{\tau}_k \, \sigma_{jk}, \qquad j \in \mathcal{N}.$$

For categorical indicators, $\sigma_{jj} = f_j(1 - f_j)$ and $\sigma_{jk} = -f_j f_k$ for $j \neq k$. Plugging the limit ridge coefficient estimators $\lim_{\lambda \to 0} \widehat{\bar{\tau}}_{k,\lambda} = \hat{\bar{\tau}}_k$ into the above and solving the first-order condition for $\bar{\tau}_{j,\lambda}$ gives this system a closed-form solution:

$$\lim_{\lambda \to 0} \hat{\bar{\tau}}_{j,\lambda} = \frac{\sum_{k \notin \mathcal{N}} f_k \hat{\bar{\tau}}_k}{1 - \sum_{k \in \mathcal{N}} f_k}, \quad \forall j \in \mathcal{N}.$$

Finally, by linearity of expectation:

$$\begin{aligned} \lim_{\lambda \to 0} \hat{\beta}_\lambda &= \lim_{\lambda \to 0} \mathbb{E}_n[\hat{\bar{\tau}}_\lambda' \bar{x}_i] = \mathbb{E}_n[(\lim_{\lambda \to 0} \hat{\bar{\tau}}_\lambda)' \bar{x}_i] \\ &= \sum_{k \notin \mathcal{N}} f_k \hat{\bar{\tau}}_k + \sum_{j \in \mathcal{N}} f_j \cdot \lim_{\lambda \to 0} \hat{\bar{\tau}}_{j,\lambda} = \frac{\sum_{k \notin \mathcal{N}} f_k \hat{\bar{\tau}}_k}{\sum_{k \notin \mathcal{N}} f_k} = \hat{\beta}_{\text{trim}}, \end{aligned}$$

which is the long regression coefficient in the sample restricted to cells with overlap. □

A.2 Details for Section 3.2

We adapt the results in Appendix B.2 of Armstrong et al. (2023) to our setting. We allow the distribution Q of ε to be unknown and possibly non-Gaussian, and only maintain the assumption that ε_i is independent across i. The extension to clustered errors is similar and omitted here for brevity.

The class of possible distributions for Q is denoted by $\mathcal{Q}_n$. We use $P_{\theta,Q}$ and $E_{\theta,Q}$ to denote probability and expectation when Y is drawn according to $Q \in \mathcal{Q}_n$ and $\theta \in \Theta$, and we use the notation P_Q and E_Q for expressions that depend on Q only and not on θ.

Consider the estimator introduced in Section 3.2. Let $V_Q = \text{Var}_Q(\hat{\beta}_{\lambda_C^*}) = \sum_{i=1}^n a_{\lambda_C^*,i}^2 E_Q \varepsilon_i^2$,

the variance of the estimator.

Theorem A.1. *Suppose that, for some $\eta > 0$, $\eta \leq E_Q \varepsilon_i^2$ and $E_Q|\varepsilon_i|^{2+\eta} \leq 1/\eta$ for all i and all $Q \in \mathcal{Q}_n$. Suppose also that, for some sequence c_n with $c_n = \mathcal{O}(\sqrt{n})$, we have*

(i) $\max\{\sqrt{n}c_n, 1\} \cdot \text{Lind}(a_{\lambda_C^*}) \to 0$*; and*

(ii) $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_{\theta,Q}(\|W(\hat{\theta}^{init} - \theta)\|_2 \leq c_n) \to 1$.

Then, for any $\delta > 0$, $\inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q\left(|(\hat{V}_{\lambda_C^*,rob} - V_Q)/V_Q| < \delta\right) \to 1$. Furthermore,

$$\liminf_n \inf_{\theta \in \Theta, Q \in \mathcal{Q}_n} P_Q\left(\beta \in \left\{\hat{\beta}_{\lambda_C^*} \pm cv_\alpha(\overline{\text{bias}}_{a_{\lambda_C^*},C}/\sqrt{\hat{V}_{\lambda_C^*,rob}}) \cdot \sqrt{\hat{V}_{\lambda_C^*,rob}}\right\}\right) \geq 1 - \alpha. \quad \text{(A.6)}$$

The convergence rate c_n depends on whether $\hat{\theta}^{\text{init}}$ is estimated via the long regression or the cross-validated generalized ridge regression.

Proof sketch. The argument adapts Appendix B.2 of Armstrong et al. (2023). The estimator $\hat{\beta}_{\lambda_C^*} = a'_{\lambda_C^*} Y$ is a linear estimator with non-random weights applied to independent errors. Under condition (i), the Lindeberg condition holds for the triangular array $\{a_{\lambda_C^*,i}\varepsilon_i\}_{i=1}^n$, yielding $(\hat{\beta}_{\lambda_C^*} - \mathbb{E}[\hat{\beta}_{\lambda_C^*}])/V_Q^{1/2} \xrightarrow{d} N(0,1)$. For the variance estimator, writing $\hat{\varepsilon}_{\text{init},i} = \varepsilon_i + W_i'(\theta - \hat{\theta}^{\text{init}})$, condition (ii) ensures the cross-terms are asymptotically negligible, so that $|\hat{V}_{\lambda_C^*,\text{rob}} - V_Q|/V_Q \xrightarrow{p} 0$. The coverage result (A.6) then follows from the bias bound $|b| \leq \overline{\text{bias}}_{a_{\lambda_C^*},C}/V_Q^{1/2}$ and Slutsky's theorem. □

A.3 Details for Section 3.3

We now formally prove an impossibility result we discussed in Section 3.3: any CI $\mathcal{C}$ that has (uniformly) valid coverage under this unrestricted parameter space must be trivial in the sense that the worst-case expected length is unbounded, $\sup_{(\beta,\gamma',\delta') \in \mathbb{R}^{2+2k}} \mathbb{E}_{(\beta,\gamma',\delta')}\mu(\mathcal{C}) = \infty$. Here, $\mu(A)$ denotes the Lebesgue measure of a measurable set A. This follows from applying the worst-case CI length bounds derived by Low (1997) to the specific class of linear functions we consider.

Proposition A.1. *Let $\mathcal{C}$ be a CI for β with nominal coverage probability $1-\alpha$ (i.e., $\inf_{(\beta,\gamma',\delta')\in\mathbb{R}^{2+2k}} P(\beta \in \mathcal{C}) \geq 1-\alpha$). If there is no overlap, then $\sup_{(\beta,\gamma',\delta')\in\mathbb{R}^{2+2k}} \mathbb{E}_{(\beta,\gamma',\delta')}\mu(\mathcal{C}) = \infty$.*

Remark 2. The proposition applies to *any* CI $\mathcal{C}$ without restricting it to be fixed length CIs based on affine estimators. Hence, the result establishes that some restriction on the parameter space is necessary to obtain informative CIs when overlap fails. Note that when the focus is on fixed-length CIs, which includes many existing CIs, as is the case for this paper, the result implies that $\mu(\mathcal{C}) = \infty$. That is, there does not exist any non-trivial CI that has correct coverage if one were to restrict attention to CIs that take the usual form of "linear estimator $\pm$ critical value"

Proof. By Low (1997), the worst case length is lower bounded by $(1-\alpha-\varepsilon/4)\omega(\varepsilon)$, where $\omega(\varepsilon)$ is the *modulus of continuity* (see, e.g., Donoho, 1994 or Armstrong and Kolesár, 2018) defined as

$$\omega(\varepsilon) := \sup_{(\beta,\gamma',\delta')'\in\mathbb{R}^{2+2k}} 2\beta \quad \text{s.t.} \quad \|D\beta + X\gamma + (D \circ \tilde{X}_{-1})\delta\|^2 \leq \frac{\varepsilon^2}{4}.$$

It suffices to show that $\omega(\varepsilon) = \infty$. Suppose to the contrary that $\omega(\varepsilon) < \infty$. Then, there exists $\theta^* = (\beta^*, \gamma^{*\prime}, \delta^{*\prime})'$ such that θ^* satisfies the constraint and $\omega(\varepsilon) = 2\beta^*$. Note that it must be the case that $\beta^* > 0$. Since there is no overlap, it must be the case that there exists some j such that, without loss of generality, the binary $X_{j+1} \leq D$ where X_{j+1} denotes the $(j+1)$th column of X and $\leq$ is interpreted elementwise. That is, all individuals i with $X_{i,j+1} = 1$ are treated.

Now, consider the jth column of $D \circ \tilde{X}_{-1}$ which is by definition $D \circ (X_{j+1} - \bar{x}_{j+1}\mathbf{1})$. We have $D \circ (\bar{x}_{j+1}\mathbf{1}) = D\bar{x}_{j+1}$, and, due to lack of overlap, $D \circ X_{j+1} = X_{j+1}$. For any constant $c > 0$, consider $\theta_c^* = (\beta^* + c\bar{x}_{j+1}, \gamma^* - ce_{j+1}, \delta^* + ce_j)$. Note that θ_c^* satisfies the constraint

because

$$
\begin{aligned}
&D(\beta^* + c\bar{x}_{j+1}) + X(\gamma^* - ce_{j+1}) + (D \circ \tilde{X}_{-1})(\delta^* + ce_j) \\
=&D\beta^* + X\gamma^* + (D \circ \tilde{X}_{-1})\delta^* + (c\bar{x}_{j+1})D - cX_{j+1} + c(D \circ (X_{j+1} - \bar{x}_{j+1}\mathbf{1})) \\
=&D\beta^* + X\gamma^* + (D \circ \tilde{X}_{-1})\delta^* + (c\bar{x}_{j+1})D - cX_{j+1} + c(X_{j+1} - \bar{x}_{j+1}D) \\
=&D\beta^* + X\gamma^* + (D \circ \tilde{X}_{-1})\delta^*.
\end{aligned}
$$

Since $\beta^* + c\bar{x}_{j+1} > \beta^*$, this is a contradiction. Hence, it must be the case that $\omega(\varepsilon) = \infty$. □

B Additional simulation results under overlap

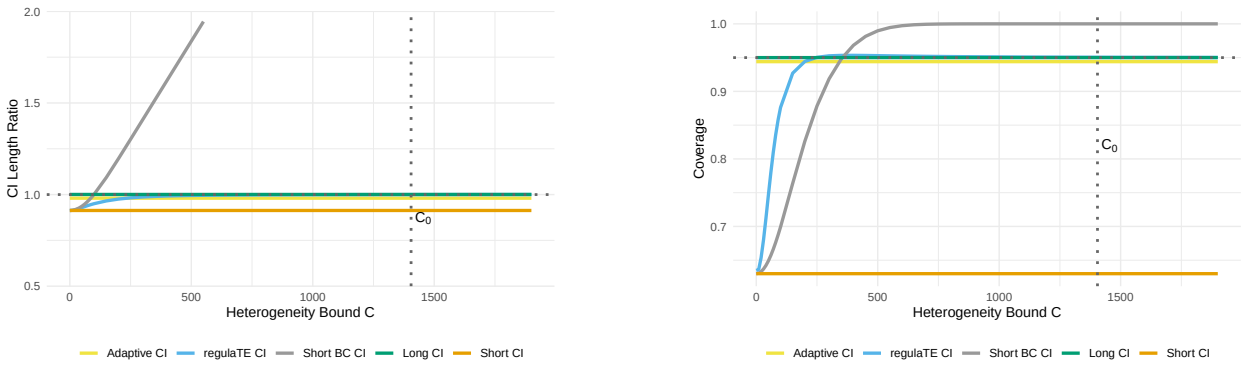
In this section, we use calibrated simulation to illustrate the behavior of `regulaTE` CI in settings with overlap where the long regression is well defined and the average treatment effect is point identified. We follow the simulation design described in Section 3.4, but increase the sample size to 10,000 individuals. For the realized treatment assignment used in this design, there is no lack of overlap across covariate values. In this setting, we can also compare the performance of `regulaTE` CI with the CI based on the adaptive estimator proposed by Armstrong et al. (2025), which is a weighted average between the short and long regression estimators where the weights depend on their realized discrepancy and the relative efficiency of the short and long regression. Therefore, such comparison is not feasible in Section 3.4 for settings without overlap.

Figure 5 shows that the `regulaTE` CI rapidly converges to the long regression CI once C exceeds approximately 250, at which point coverage is also correct at 95%. In contrast, the bias-corrected short-regression confidence interval becomes overly conservative as C increases. To preserve the readability of the ratio plot, values exceeding 2 for the bias-corrected interval are truncated.

The performance of the CI based on the adaptive estimator is rather different because it does not incorporate a user-specified bound C. We implement the adaptive procedure using the recommended soft-thresholding estimator in Armstrong et al. (2025). We use the critical value that guarantees at least 95% worst-case coverage when the bias for the short regression

is assumed to be within one standard error of the difference between the short and the long regression estimators. Coverage is evaluated using 100,000 Monte Carlo replications. As shown in Figure 5, the resulting CI has constant coverage at 94% that does not vary with C, and the CI is very close to the long regression for this setting.

Figure 5: Sensitivity of Coverage and CI Length under Overlap in DGP Calibrated to Angrist (1998)



Notes: "`regulaTE` CI" refers to the bias-aware fixed-length confidence interval based on the `regulaTE` estimate. "Short CI" refers to the CI based on the short regression estimate. "Short BC CI" refers to the bias-corrected short regression CI. Both "`regulaTE` CI" and "Short BC CI" are heteroskedasticity-robust with 95% coverage guarantees under each heterogeneity bound C on the horizontal axis. The ratio of the CI lengths is relative to the length of the `regulaTE` CI under $C = C_0$. "Long CI" refers to the CI based on the long regression estimate. "Adaptive CI" refers to the CI based on the adaptive estimate.