Omitted Variable Bias in Difference-in-Differences Designs

Juejue Wang[1], Pedro H.C. Sant'Anna[2], Victor Chernozhukov[3], and Carlos Cinelli[1]

[1]Department of Statistics, University of Washington

[2]Department of Economics, Emory University

[3]Department of Economics, MIT

May 19, 2026

Abstract

We study the omitted-variable bias (OVB) problem in canonical difference-in-differences (DiD) designs when unobserved confounding induces departures from the parallel trends assumption. Our results provide a novel characterization of the OVB formula for the average treatment effect on the treated (ATT), which may be of independent interest. We show how the ATT bias is mainly governed by the strength of confounding in the treatment-selection mechanism and provide alternative ways of quantifying this strength, such as (i) changes in the average odds of treatment among the treated, (ii) confounding imbalance between treated and control units, or (iii) variation explained in treatment odds among the untreated. Building on these results, we offer sensitivity statistics for routine reporting, describing the minimum strength of confounding required to overturn the conclusions of a DiD study, as well as formal bounds on the strength of confounders based on comparisons to observed covariates or pre-trends. Finally, we provide flexible and efficient statistical inference methods for the bounds on ATT, which can leverage modern machine learning algorithms for estimation. We demonstrate the utility of our approach in an empirical example that estimates the effects of minimum wage on teen employment.

1 Introduction

Difference-in-differences (DiD) has become the most widely used research design for causal inference with observational data in economics and other empirical sciences (Goldsmith-Pinkham, 2024). In its canonical form, DiD identifies the average treatment effect on the treated (ATT) under a parallel

trends assumption (PTA), which states that, in the absence of treatment, the average outcomes of treated and control units would have evolved in parallel over time—possibly only after conditioning on a set of observed pre-treatment covariates. Because the PTA concerns counterfactual trajectories that are never observed for treated units after treatment, it is fundamentally untestable, and investigators marshal what evidence they can to defend its plausibility in the context of their investigations.

By far, the most common form of gathering such evidence is performing placebo tests using pre-treatment information. Researchers typically plot or formally test whether the outcomes of treated and control groups evolve in parallel before treatment, and read the absence of differential pre-trends as evidence for post-treatment parallel trends. Recent work has improved the statistical properties of these tests (Roth, 2022; Bilinski and Hatfield, 2018; Dette and Schumann, 2020), but a more fundamental limit remains. Pre-trends speak only to what happened before treatment, and parallel trends may well hold ex-ante and fail ex-post. Another common practice is to check whether treated and control groups are balanced on observed pretreatment covariates that are also thought to be determinants of changes in the untreated potential outcome (Baker et al., 2025). But again, while null findings on such balance tests are consistent with the parallel trends assumption, they are not dispositive—there could still be *unobserved* confounders that induce violations of parallel trends. Moreover, in many settings, observed covariates may be imbalanced, and pre-trends may differ across groups. Yet, researchers would still like to learn something about the ATT, even if there is evidence that the parallel trends assumption does not hold exactly.

To this end, sensitivity analyses have been proposed that use observed pre-treatment deviations from parallel trends to bound the plausible magnitude of post-treatment violations (e.g., Rambachan and Roth, 2023). This framework has been widely adopted, and is now recommended as a standard sensitivity analysis tool in the DiD literature (Roth et al., 2023; Baker et al., 2025). While a step forward from naively assuming that parallel trends hold exactly, this approach still suffers from shortcomings. By construction, pre-trend extrapolation anchors the sensitivity analysis to features of the pre-treatment data rather than to the mechanisms that drive violations of parallel trends. It is informative only insofar as pre-treatment dynamics are a reliable proxy for the post-treatment counterfactual—an assumption that it cannot itself adjudicate, as it offers no framework for reasoning about *why* parallel trends might fail in the first place. One immediate symptom of this deficiency is that the framework cannot be applied to the simplest, canonical two-period DiD design, where pre-

trends are not available. Importantly, applied work typically discusses violations of parallel trends as arising from *unobserved confounders* that *jointly* shape the *selection* of units *into treatment* and their untreated *outcome trends* (Callaway and Sant'Anna, 2021; Roth et al., 2023; Baker et al., 2025).[1] How can we translate these factors into formal statements about deviations from parallel trends and the ATT bias? How strong would unobserved confounding have to be to overturn a given conclusion in a DiD study?

In this paper, we develop a suite of sensitivity analysis tools for DiD that allows one to easily answer such questions. We first derive a novel characterization of the OVB formula for the ATT. While motivated by DiD, this result is important in its own right and may be of independent interest—we elaborate on this point in the related literature below. We show how the bias due to violations of parallel trends admits a formal decomposition into the product of a scaling factor estimable from the data, and three interpretable bias factors that researchers already informally invoke: (i) the strength of unobserved confounders in shaping selection into treatment, (ii) their strength in shaping the evolution of untreated potential outcomes, and (iii) the alignment between these two channels. Among these, we further discuss how selection into treatment is the most important factor for bounding this bias (since the other two factors are each upper bounded by one), and we offer a rich set of equivalent ways that practitioners can use to reason about it, such as how unmeasured confounding increases the average odds of treatment among the treated, widens the covariate imbalance between treated and control units, or explains variation in treatment odds among the untreated.

Next, we derive (extreme) robustness values for DiD. These are sensitivity statistics that characterize the minimum strength of unobserved confounding needed to overturn a given conclusion about the ATT (Cinelli and Hazlett, 2020, 2025a; Tsao et al., 2026; Chernozhukov et al., 2026). Routine reporting of these quantities, alongside point estimates and standard errors, provides a quick and simple way to communicate how robust DiD findings are to violations of the parallel trends assumption. We also offer formal bounds on the ATT based on plausibility judgments about how the strength of unobserved confounders compares against the explanatory power of observed covariates. These bounds connect naturally to balance checks already performed by applied researchers,

[1] For example, Roth et al. (2023) states that *"even if one conditions on observable covariates, there are often concerns that the necessary parallel trends assumption may be violated due to time-varying unobserved confounding factors."* And, as Baker et al. (2025) put it, parallel trends deviations may be due to *"shocks that jointly shape treatment decisions and untreated outcomes."*

who compare observed covariates across treated and control groups as informal evidence for parallel trends. Our framework gives these comparisons a direct role in sensitivity analysis, allowing the observed imbalance on measured covariates to discipline judgments about the unobserved imbalance one is willing to entertain. When multiple pre-treatment periods are available, the same logic applies to pre-treatment dynamics, allowing our framework to be used in tandem with current pre-trend extrapolation approaches.

Finally, we provide flexible and efficient methods for statistical inference using Debiased Machine Learning (DML), which allows the use of modern machine learning algorithms for estimation—though we note our approach can also be used with standard parametric regressions. We demonstrate the utility of our approach in an empirical application revisiting the analysis of the effect of minimum wage on teen employment in Callaway and Sant'Anna (2021). Open-source software for R implements the methods discussed in this paper.

Related literature. Our work contributes to the growing literature on sensitivity analysis for DiD, where the dominant approach is the pre-trend extrapolation framework of Rambachan and Roth (2023). Pre-trend extrapolation is best understood as a way of benchmarking the plausible magnitude of post-treatment violations against observed pre-treatment dynamics. What it does not do, however, is explain how such violations arise in the first place. Our framework fills this gap by decomposing the bias into interpretable factors tied to selection into treatment and untreated outcome evolution, and in doing so, it opens the door to a broader set of benchmarking strategies. Pre-trend extrapolation becomes one such strategy—a way of calibrating the magnitude of the bias factors against pre-treatment dynamics—but it is no longer the only one. Observed covariates, for instance, can also be used to benchmark the plausibility of unobserved confounding directly, further connecting sensitivity analysis to the balance tests that applied researchers already perform as a matter of course. We provide a formal way of doing both in this paper (see Section 4).

Our work is also related to the broader literature on sensitivity analysis for the ATT. In particular, our results are most closely related to Chernozhukov et al. (2026), who develop a general OVB framework for linear functionals of the conditional expectation of the outcome, which encompasses the ATT as a special case; see Bach et al. (2025) for a direct application of Chernozhukov et al. (2026) to DiD designs using this unconditional parameterization. Though we build on Chernozhukov et al. (2026), our OVB formula for the ATT differs from the one presented there because we exploit a structural feature of the ATT—namely, that the bias admits a parameterization conditional solely

on the untreated units. This parameterization has several desirable properties, such as requiring plausibility judgments only on the local parameters that directly drive the bias. We further extend Chernozhukov et al. (2026) by providing alternative ways of quantifying selection strength in terms of treatment odds as well as covariate imbalance, both for our preferred conditional parameterization and for the original unconditional parameterization. See Appendix F.1 for a detailed discussion.

Our results are also closely related to the variance-based sensitivity analysis proposed by Huang and Pimentel (2025) for weighting estimators, which bounds the bias of the ATT by restricting an R^2 measure of the balancing weights. Our approach differs from theirs in four main ways. First, we use a non-centered parameterization of selection strength, which naturally avoids a zero-denominator issue acknowledged in Huang and Pimentel (2025). Second, we further decompose the bias into three components—the strength of confounding in treatment selection, its strength in untreated outcome trends, and the alignment between the two. We show that Huang and Pimentel (2025)'s parameterization bundles the trend and alignment components together, so that their upper bound on the bias implies setting these two components to one. Third, their benchmarking analysis uses observed covariates in a way that is anti-conservative relative to ours. Fourth, we use debiased machine learning for statistical inference. In Appendix F.2 we provide a detailed discussion of all these points, while also extending and completing the derivations in Huang and Pimentel (2025) to obtain results analogous to ours in their centered parameterization.

Outline of the paper. The rest of the paper is organized as follows. Section 2 introduces the running example, in which we examine the effect of minimum wage increases on teen employment. Section 3 formalizes the omitted variable bias problem in the canonical two-period DiD design. We derive an exact bias decomposition for the ATT under violations of parallel trends and show how the magnitude of the resulting bias can be summarized in terms of interpretable measures of the strength of confounding in driving treatment selection and explaining the outcome evolution. Section 4 then develops tools for assessing the plausibility of such confounding strengths by benchmarking sensitivity parameters against observed covariates or pre-trends. Section 5 discusses estimation and inference for our proposed framework using debiased machine learning. We then return to the minimum-wage application in Section 6, where we illustrate the proposed methods for sensitivity analysis in practice. We offer some final remarks in Section 7. Proofs and additional results are deferred to the appendix.

Notation. We write $R^2_{Y\sim X}$ for the (possibly uncentered) R^2 from the orthogonal linear projection of a scalar random variable Y onto a random vector X. We write $\eta^2_{Y\sim X}$ for the corresponding non-parametric R^2, obtained by replacing the best linear predictor with the best predictor of Y given X, i.e., the conditional expectation function $E[Y \mid X]$. The partial R^2 of Y with U given X is defined as $R^2_{Y\sim U|X} := (R^2_{Y\sim U+X} - R^2_{Y\sim X})/(1 - R^2_{Y\sim X})$, with analogous definitions for the non-parametric version. We use the conventional notation $\mathrm{d}L/\,\mathrm{d}P$ to denote the Radon-Nikodym derivative of measure L with respect to P. For two probability distributions $P \ll Q$ we define the chi-squared divergence of P from Q as $\chi^2(P\|Q) = \int(\mathrm{d}P/\,\mathrm{d}Q - 1)^2\,\mathrm{d}Q$. We use $P_{V|W=w}$ to denote the probability distribution of vector V given $W = w$. We use E_n to denote the empirical expectation, i.e., $E_n[f(X_i)] := \frac{1}{n}\sum_{i=1}^n f(X_i)$.

2 Background and running example

In this section we introduce the running example used throughout the paper. This example serves three main purposes: (i) it illustrates the standard use of pre-trends and covariates as a tool for assessing the plausibility of parallel trends; (ii) it shows how our framework already refines this standard analysis through a richer characterization of *observed* biases; and (iii) it sets the stage for the formal sensitivity analysis to *unobserved* confounders that we develop in the rest of the paper.

2.1 The effect of minimum wage on teen employment

Whether increases in minimum wages reduce teen employment has been actively debated in labor economics for decades, with credible evidence on both sides (Neumark and Shirley, 2022; Dube and Lindner, 2024). Here we revisit the analysis of Callaway and Sant'Anna (2021), who study this question using a difference-in-differences design.

Between 2001 and 2007, the federal minimum wage remained at \$5.15, but several states raised their own minimums above this floor at different points in time. Callaway and Sant'Anna (2021) exploit this variation in treatment timing to estimate group-time average treatment effects for each cohort of states, defined by when they first raised their minimum wage above the federal level. To fix ideas, here we focus on the last cohort, the states that raised their minimum wage in 2007. The treated group consists of the 584 counties in those states, and the control group consists of the 1,377 counties in states that never raised their minimum wage above the federal floor before 2007.

An investigator interested in estimating the ATT might start the analysis by positing that, in

the absence of any policy change, teen employment in treated and control counties would have evolved in parallel over time. Under this parallel trends assumption, the ATT is identified by the usual canonical two-by-two difference-in-differences estimand. The results are shown in panel (a) of Figure 1 (see also Table 1)[2]. The treatment effect of interest is the estimate for 2007, and we find that increases in minimum wages resulted in a 2.77% reduction in teen employment for that group. This is a sizable, policy-relevant negative effect, but how credible is this estimate?

Beyond the effect of interest, Figure 1(a) also shows pre-treatment estimates ("pre-trends") from 2002 to 2006, which are now commonly used to assess the plausibility of the parallel trends assumption. If the untreated trends of both groups were indeed evolving in parallel over time, these pre-trends should be equal to zero, barring sampling variation. Yet, the 2006 pre-trend is -3.9%, which is not only large and statistically significant, but also of the same order of magnitude as the treatment effect estimate itself. This suggests that treated and control counties were already on diverging trends before any policy changes.

What might explain such deviations from parallel trends? A natural concern is that employment in counties with *different characteristics* may not evolve in parallel over time. To probe this concern, another common approach is to check whether the treated and control groups are *balanced* in terms of covariates that could plausibly affect employment trajectories. Table 2 reports standardized differences in means for several such characteristics. We find that treated counties are more populous, whiter, more educated and have lower poverty rates than control counties; they are also concentrated in the Midwest and West, with the South substantially underrepresented. Most differences exceed the 0.25 threshold, conventionally regarded as problematic (Imbens and Rubin, 2015; Baker et al., 2025). These are the kind of imbalances that could account for parallel trends violations, and they motivate adjusting for observed covariates before drawing any conclusions about the policy effect.

To alleviate such concerns, panel (b) of Figure 1 reports DiD estimates under a *conditional* parallel trends assumption, which requires only that counties with the *same observed characteristics* would have evolved in parallel over time. These estimates account for observed covariates non-parametrically, using random forests and debiased machine learning[3], as we will discuss in Section 5. The treatment effect estimate is now -3.66%, somewhat larger in magnitude than the unconditional estimate, and points to the same conclusions. However, the 2006 pre-trend estimate also remains

[2]We use uniform confidence bands constructed via multiplier bootstrap (Callaway and Sant'Anna, 2021) with 1,000 iterations. See Algorithm 2 in Appendix G for details.

[3]We use 10-fold cross-fitting. Details of the tuning parameters can be found in Appendix G.

essentially unchanged. In other words, though adjusting for observed covariates has not changed our treatment effect estimate, it has not resolved the pre-trend problem either. This persistent pre-trend deviation suggests confounding is acting through factors not captured by the observed data.

At this point, the standard recommendation in the DiD literature is to proceed with a sensitivity analysis of the type of Rambachan and Roth (2023). It proceeds as follows: take the 2006 deviation of -3.8%, assume the same deviation persists in 2007, and subtract it from the conditional estimate of -3.66%. Since 3.8% is larger than 3.66%, this is evidently enough bias to overturn our original conclusion. This approach is straightforward and an improvement over simply assuming parallel trends and ignoring the problem. But it is also somewhat unsatisfactory as it leaves several questions unanswered. Why have the observed covariates barely changed these estimates, despite their substantial imbalance? Should we be looking at standardized mean differences or something else? What kind of unobserved confounding would have to be at work to produce these biases? And is pre-trend extrapolation the only way to gauge the plausible magnitude of such confounders?

As a preview of how the results of this paper can help analysts answer these questions, we show that the difference between the conditional and unconditional estimates in fact admits the following exact decomposition:

$$\underbrace{(-0.0277)}_{\text{Unconditional Estimate}} - \underbrace{(-0.0366)}_{\text{Conditional Estimate}} = \underbrace{0.0089}_{\text{Bias}} = \underbrace{0.194}_{\text{Alignment}} \times \underbrace{0.117}_{\text{Trend}} \times \underbrace{2.548}_{\text{Selection}} \times \underbrace{0.154}_{\text{Scale}}.$$

Table 3 reproduces the above decomposition using all covariates, and also presents analogous decompositions obtained by adjusting for each observed covariate separately. Almost every applied DiD paper conditions on covariates, yet the standard narrative about why they matter rarely extends beyond covariate imbalance. Decompositions such as the one above provide an exact post-mortem of how covariates affect (or not) the results.

First, the "selection" component confirms that indeed there is a sizable distributional imbalance between treated and control groups, corresponding to a chi-squared divergence of $2.548^2 \approx 6.5$—note the relevant metric of distributional imbalance here is not the commonly used standardized mean difference, though in some cases the two may be close to each other, as we will discuss. Second, one of the reasons why this imbalance does not translate into a substantial difference in estimates in our running example is because these covariates explain only $0.117^2 \approx 1.4\%$ of the variation in trends among the control group (the "trend" component). Third, for bias to arise, covariates must affect both channels jointly—their effect on the outcome trend must correlate with their effect on

selection. This correlation, the "alignment" component, equals only 0.194 in our example, further attenuating the bias. As with the conditional estimate, all these components can also be estimated non-parametrically using debiased machine learning with the tools we provide in this paper. We argue that exercises such as the one above should be a routine part of DiD studies.

But more importantly, this same decomposition used to explain how observed covariates *did* change our estimates can also be used to contemplate how unobserved confounders *would* have changed our estimates. For example, it explains what unobserved confounders need to look like, in terms of their strength in predicting outcome trends and treatment selection, in order to induce any certain amount of bias—including biases derived from pre-trend extrapolation. Crucially, it also opens up different ways of gauging the plausible magnitude of confounding. For instance, if an investigator has grounds to claim that no plausible unobserved confounder is more strongly imbalanced than the observed covariate region—or, equivalently, in terms of how it changes the odds of belonging to the treated group—this is also sufficient to bound the bias. We develop these tools formally in the rest of the paper, beginning with the bias decomposition itself in the next section.

3 Omitted variable bias in difference-in-differences designs

In this section, we formally define and derive the omitted variable bias formula for the ATT in the canonical DiD design. We show how it can be expressed in terms of interpretable bias factors that capture the strength of omitted confounding in treatment selection, untreated outcome evolution, and the alignment between these two channels.

3.1 Problem setup

We begin by introducing the omitted variable bias problem in the canonical two-by-two DiD setup. Let $t \in \{1, 2\}$ index time, where $t = 1$ denotes the pre-treatment period and $t = 2$ denotes the post-treatment period. Let $D \in \{0, 1\}$ denote treatment-group assignment, with $D = 1$ corresponding to the treated group and $D = 0$ corresponding to the control group. For each period t, $Y_t(1)$ and $Y_t(0)$ denote the treated and untreated potential outcomes at time t, respectively. We assume the observed outcome at time t satisfies

$$Y_t = DY_t(1) + (1 - D)Y_t(0). \tag{1}$$

Equation (1) is often referred to as the consistency assumption, or the stable unit treatment value assumption (SUTVA).

The causal parameter of interest is the average treatment effect on the treated (ATT) at period 2, defined as

$$\text{ATT} := E[Y_2(1) \mid D = 1] - E[Y_2(0) \mid D = 1]. \tag{2}$$

Now let $\Delta Y := Y_2 - Y_1$ denote the outcome evolution, X denote a vector of observed covariates and U denote a vector of *unobserved* covariates. Suppose that, conditionally on both X and U, the outcomes of treated and control units would have evolved in parallel over time in the absence of treatment, i.e.,

$$E[\Delta Y(0) \mid D = 1, X, U] = E[\Delta Y(0) \mid D = 0, X, U]. \tag{3}$$

Under this conditional parallel trends assumption—along with the assumptions of no anticipation and other regularity conditions[4], such as weak overlap below—we obtain the (well-known) result that the ATT is identified by the following *difference-in-differences estimand*,

$$\theta := E[\ E[\Delta Y \mid D = 1, X, U] - E[\Delta Y \mid D = 0, X, U] \ \mid \ D = 1\ ]. \tag{4}$$

We refer to θ as the "long" DiD estimand because it depends both on the observed covariates X and the unobserved confounders U. The starting point of our analysis is that the investigator has determined that she is interested in the DiD estimand θ as in (4). We further impose the following regularity conditions.

Assumption 1 (Regularity Conditions)**.** *As regularity conditions for the DiD estimand, we assume the outcome evolution is square-integrable* $E[\Delta Y^2] < \infty$ *and the weak overlap condition* $E\left[P(D = 1)^{-1}P(D = 0|X, U)^{-1} \mid D = 1\right] < \infty$.

Remark 1. *We intentionally place limited emphasis on the specific identification assumptions that endow the DiD estimand* θ *with causal meaning. Different formulations of the DiD assumptions can lead to the same statistical estimand (4) above, after appropriately redefining assumptions and variables. See, for example, Callaway and Sant'Anna (2021) for identification results involving group-time average treatment effects with multiple time-periods, which invoke various different choices of control groups, parallel trends and no-anticipation assumptions. Another example is the identification of the ATT under a regular unconfoundedness assumption, instead of parallel trends—for that*

[4]See Appendix A.4 to recall the usual standard DiD assumptions.

case, one simply replaces ΔY with Y. Our results apply to any such estimand that can be expressed in the generic form (4) above.

In practice, however, the unobserved confounders U are not available to the researcher, so we cannot estimate the "long" DiD parameter θ. Instead, we are forced to estimate its "short" version, omitting U from the analysis,

$$\theta_s := E[\ E[\Delta Y \mid D=1, X] - E[\Delta Y \mid D=0, X]\ \mid\ D=1\ ]. \tag{5}$$

Note that the parallel trends assumption may not hold if we condition on X alone, and our estimates may thus suffer from omitted variable bias. In fact, under standard DiD assumptions, the omitted variable bias due to the omission of U exactly equals the average deviation from parallel trends, among the treated, from conditioning on X alone, i.e.,

$$\theta - \theta_s = -E\left[\ \underbrace{E[\Delta Y(0) \mid X, D=1] - E[\Delta Y(0) \mid X, D=0]}_{=:\delta(X)\text{ deviation from conditional parallel trends given } X.}\ \mid\ D=1\ \right].$$

Consequently, judgments about deviations from parallel trends can be translated into judgments about omitted confounding. Conversely, the presence of omitted confounders provides a natural way to explain deviations from parallel trends.

Our goal is to characterize the omitted variable bias—or, equivalently, the deviations from parallel trends—in terms of the strength of the omitted variable U, capturing the distinct channels through which U confounds our estimand.

3.2 The omitted variable bias formula

Denote by g_0 and g_{0s} the long and short regression functions of the outcome evolution for the untreated group,

$$g_0 := E[\Delta Y | D=0, X, U], \qquad g_{0s} := E[\Delta Y | D=0, X].$$

The long and short DiD estimands can thus be compactly written as

$$\theta = E[\Delta Y - g_0 \mid D=1], \qquad \theta_s = E[\Delta Y - g_{0s} \mid D=1].$$

Note that the average treated trend $E[\Delta Y | D=1]$ is the same in both the long and short parameters, and it cancels out from the bias. The OVB is therefore fully governed by errors in the extrapolation of the untreated trend from the control group to the treated group. That is,

$$\theta - \theta_s = -(\theta_0 - \theta_{0s}),$$

where,

$$\theta_0 := E[g_0|D=1], \qquad \text{and,} \qquad \theta_{0s} := E[g_{0s}|D=1]. \tag{6}$$

In order to better understand the nature of this extrapolation, consider the marginal odds, as well as the short and long conditional odds of treatment,

$$O := \frac{P(D=1)}{P(D=0)}, \qquad O_{XU} := \frac{P(D=1 \mid X,U)}{P(D=0 \mid X,U)}, \qquad O_X := \frac{P(D=1 \mid X)}{P(D=0 \mid X)}.$$

A key step for our approach is the following lemma, which rewrites θ_0 and θ_{0s} as inner products of the conditional regression functions and rebalancing weights given by the conditional to marginal odds-ratio, under the control population.

Lemma 1. *Under Assumption 1, the long and short parameters θ_0 and θ_{0s} in (6) can be written as*

$$\theta_0 = E\left[g_0\left(\frac{O_{XU}}{O}\right) \mid D=0\right], \quad \theta_{0s} = E\left[g_{0s}\left(\frac{O_X}{O}\right) \mid D=0\right].$$

Moreover, $g_{0s} = E[g_0 \mid D=0, X]$ and $O_X = E[O_{XU} \mid D=0, X]$.

The fact that $g_{0s} = E[g_0 \mid D=0, X]$ follows directly from the law of iterated expectations. The property $O_X = E[O_{XU} \mid D=0, X]$ can be verified by applying Bayes' rule. Using this lemma, we can derive the following characterization of the OVB for the ATT. It shows that deviations from parallel trends can be exactly characterized by the ability of omitted confounders to explain treatment selection—as parameterized by the odds of treatment—and trend variation, jointly.

Theorem 1 (OVB for the ATT)**.** *Consider the long and short estimands θ and θ_s in (4) and (5). Under the regularity conditions of Assumption 1, the OVB is given by the (negative of the) covariance of errors induced by the omission of U both in the regression function and in the odds-ratio, among the untreated units,*

$$\theta - \theta_s = -\operatorname{Cov}\left(g_0 - g_{0s},\ \frac{O_{XU}}{O} - \frac{O_X}{O} \;\middle|\; D=0\right).$$

Moreover, this bias can be further expressed in terms of R^2 measures:

$$\theta - \theta_s = -\underbrace{\rho_0}_{\textit{alignment}}\ \underbrace{C_{0\Delta Y}}_{\textit{trend}}\ \underbrace{C_{0D}}_{\textit{selection}}\ \underbrace{S_0}_{\textit{scale}},$$

where,

$$\rho_0 := \operatorname{Cor}\left(g_0 - g_{0s}, O_{XU} - O_X \mid D=0\right), \quad C^2_{0\Delta Y} := \eta^2_{\Delta Y \sim U|X,D=0}, \quad C^2_{0D} := \frac{1 - R^2_{O_{XU} \sim O_X|D=0}}{R^2_{O_{XU} \sim O_X|D=0}},$$

and

$$S_0^2 := \sigma_{0s}^2 \nu_{0s}^2, \qquad \sigma_{0s}^2 := E[\text{Var}(\Delta Y \mid X, D=0) \mid D=0], \qquad \nu_{0s}^2 := E\left[\left(\frac{O_X}{O}\right)^2 \middle| D=0\right].$$

Theorem 1 rewrites the average deviation from parallel trends in terms that more conveniently rely on scale-free partial R^2 measures of association characterizing the strength of unobserved confounding. The bias decomposes into the product of three non-identifiable bias factors, ρ_0, $C_{0\Delta Y}$, C_{0D}—which need to be restricted by plausibility judgments on the strength of confounders—and an identifiable scale factor, S_0, which is estimable from the observed data. Note that bias arises only when all bias factors are non-zero. We now examine each term below.

The "trend" component $C_{0\Delta Y}^2$ measures the strength of confounding in explaining trend variation. Formally, this is measured by $\eta_{\Delta Y \sim U|X,D=0}^2$, which is the *non-parametric* partial R^2 of ΔY with U, given X, among the untreated. This quantifies how much residual variance of the trend in the control group is explained by the omitted confounders, after taking into account what is already explained by X. The less confounders can explain trend variation in the control group, the less the potential of such confounders to induce deviations from parallel trends. Note this component can always be left unconstrained by the investigator, as it is naturally upper bounded by 1.

The "selection" component C_{0D}^2 measures the strength of confounding in explaining selection into treatment. Formally, this is driven by $1 - R_{O_{XU} \sim O_X|D=0}^2$, i.e., the share of variation in the long treatment odds of the control group which cannot be explained by observed covariates alone. While this latter quantity is also an R^2 which can be upper bounded by 1, note the selection term C_{0D}^2 is given by the ratio, $(1 - R_{O_{XU} \sim O_X|D=0}^2)/R_{O_{XU} \sim O_X|D=0}^2$, which can be arbitrarily large. Therefore, the treatment selection component must always be restricted. To aid plausibility judgments regarding C_{0D}^2, we provide alternative characterizations of treatment selection in the next section below.

Together, the two sensitivity parameters, $\eta_{\Delta Y \sim U|X,D=0}^2$ and $1 - R_{O_{XU} \sim O_X|D=0}^2$, characterize the strength of confounding in explaining trends and treatment odds. However, for bias to arise, it is not sufficient for confounders to create errors in the trend and selection components; these errors must be systematically aligned. This is captured by the "alignment" component ρ_0, defined as the correlation between these two errors in the control arm. Similarly to the trend component, in the worst case, the alignment component can also be left unconstrained by the investigator, as its magnitude is upper bounded by 1.

Finally, the identifiable scale factor S_0 translates these three scale-free measures of the strength of confounding back into the original units of the ATT. The scale factor is characterized by the strength

of *observed covariates* in explaining variation in outcome trend and selection into treatment. The first component, σ^2_{0s}, measures the left-over variation of the trend in the control group after taking into account the part explained by X. Note the more variation X explains of the outcome trend, the less room there is for unobserved confounding to create bias. Moving to ν^2_{0s}, it shows the opposite relationship holds with respect to the strength of covariates in explaining treatment assignment. This term captures the norm of rebalancing weights, and it measures how well observed covariates separate treated from control units. The stronger this separation, the less identifying variation is being used to estimate the ATT, and the more leverage unobserved confounding has to drive the bias. Note ν^2_{0s} is also directly interpretable as the imbalance of observed covariates between treated and control units, as measured by the chi-squared divergence, as we show next.

3.3 Making sense of treatment-selection strength

As we have seen, the OVB formula for the ATT is particularly sensitive to selection strength, both unobserved (via C^2_{0D}) and observed (via ν^2_{0s}). It is therefore useful to develop alternative ways of understanding and reasoning about these quantities. In this section, we provide such characterizations in terms of increases in average treatment odds and covariate imbalance, with the goal of helping researchers specify plausible upper bounds on the unobserved selection strength C^2_{0D} using the scale most familiar to them in their own applications.

For example, in genetics, psychology, economics and other quantitative social sciences, researchers routinely think in terms of variance explained and related R^2 measures (Cohen, 2013; Imbens, 2003; Oster, 2019; Cinelli and Hazlett, 2020, 2025a; Cinelli et al., 2022; Chernozhukov et al., 2026). In medicine and epidemiology, odds and odds ratios are commonly used to characterize selection into treatment (Rothman et al., 2008; Ding and VanderWeele, 2016). Covariate imbalance—often summarized by the standardized mean difference, itself a common effect size, (Cohen, 2013)—is a widely used balance diagnostic in observational studies and randomized trials (Rubin, 2001; Stuart, 2010). Each of these communities has developed intuition about what counts as "large" or "small" on its own scale. Our goal is to formally connect these measures, allowing researchers to approach the OVB problem from these different perspectives. The following lemma provides the basic identities linking treatment odds, changes of measure, and covariate imbalance.

Lemma 2. *Under the weak overlap condition of Assumption 1, the following identities hold:*

1. ***Odds-ratio as change of measure:***

$$\frac{O_{XU}}{O} = \frac{\mathrm{d}P_{X,U|D=1}}{\mathrm{d}P_{X,U|D=0}}.$$

2. ***Control-to-treated odds identity:***

$$E\left[O_{XU}^2 \mid X, D=0\right] = O_X \times E[O_{XU} \mid X, D=1].$$

3. ***χ^2-Divergence as an average odds-ratio:***

$$\chi^2(P_{X,U|D=1} \| P_{X,U|D=0}) = E\left[\frac{O_{XU}}{O}\middle| D=1\right] - 1.$$

Applying these identities yields equivalent expressions for the selection-related components C_{0D}^2 and ν_{0s}^2 in terms of increase in average treatment odds and χ^2-divergence.

Corollary 1 (Alternative characterizations of C_{0D}^2 and ν_{0s}^2)**.** *Under the weak overlap condition of Assumption 1, the components C_{0D}^2 and ν_{0s}^2 of Theorem 1 admit the following additional interpretations:*

1. *Increase in average treatment odds among the treated:*

$$C_{0D}^2 = \frac{E[O_{XU} \mid D=1] - E[O_X \mid D=1]}{E[O_X \mid D=1]}, \qquad \nu_{0s}^2 = \frac{E[O_X|D=1]}{O}.$$

2. *Covariate imbalance between treated and control units:*

$$C_{0D}^2 = \frac{\chi^2(P_{X,U|D=1} \| P_{X,U|D=0}) - \chi^2(P_{X|D=1} \| P_{X|D=0})}{\chi^2(P_{X|D=1} \| P_{X|D=0}) + 1}, \qquad \nu_{0s}^2 = \chi^2(P_{X|D=1} \| P_{X|D=0}) + 1.$$

The odds characterization gives a direct way to translate substantive restrictions on the propensity score into bounds on C_{0D}^2. For example, if one believes that the inclusion of U at most doubles the average odds of treatment, among treated units, then Corollary 1 implies $C_{0D}^2 \leq 1$. This relationship also clarifies the connection between the selection component C_{0D}^2 of the OVB formula and sensitivity models based on worst-case bounds on odds ratios. The marginal sensitivity model of Tan (2006), for instance, assumes

$$\Lambda^{-1} \leq \frac{O_{XU}}{O_X} \leq \Lambda$$

almost surely, for some $\Lambda > 1$. Such a restriction implies

$$C_{0D}^2 \leq \frac{(\Lambda - 1)^2}{\Lambda}.$$

Thus, for example, if one believes U cannot double (or halve) the odds of treatment not only on average, but uniformly, this implies the tighter restriction $C_{0D}^2 \leq 1/2$. This also makes it clear that worst-case restrictions on odds ratios are sufficient, but not necessary, for bounding the bias.

Beyond treatment odds, Corollary 1 also provides a covariate-balance interpretation of C_{0D}^2. It shows that C_{0D}^2 captures the additional imbalance introduced by U, relative to the imbalance already captured by X. Checking covariate balance is standard practice in applied research, often through standardized mean differences (Imbens and Rubin, 2015; Baker et al., 2025). As the OVB formula shows, however, the imbalance measure directly relevant for quantifying the bias of the ATT is the χ^2-divergence, which summarizes the full distributional difference between treated and control groups. Thinking about standardized mean differences, however, can still provide a useful calibration exercise. For example, if U is binary and independent of X within each treatment group[5], then C_{0D}^2 reduces to the square of the standardized mean difference, computed using the control-group standard deviation. In this case, a standardized mean difference of 0.5 would correspond, for instance, to $C_{0D}^2 = 0.25$, or equivalently, a $1 - R^2_{O_{XU} \sim O_X | D=0} = 0.2$. In the appendix we discuss other similar ways of calibrating intuition about the χ^2-divergence.

4 Benchmarking the strength of unobserved confounding

The main difficulty in sensitivity analysis is specifying plausible strengths for unobserved confounding. While in some cases researchers may not be able to make absolute judgments about the magnitude of an omitted confounder, they may still have grounds to make judgments of its relative importance. For example, a researcher may be willing to argue that any unobserved confounder is no stronger than some observed covariate, or that post-treatment confounding is no stronger than the confounding revealed by pre-treatment periods. In this section, we develop two benchmarking procedures that formalize these comparisons.

4.1 Comparing unobserved confounders against observed covariates

Our first analysis compares the gains in explanatory power due to unobserved confounders with the observed gains in explanatory power from key observed covariates. Let X_j denote the benchmark covariates against which we wish to calibrate the strength of U, and let X_{-j} denote the remaining

[5]Note this is unlikely to happen, because D is a collider for U and X, but this is still an interesting thought exercise to calibrate judgments about χ^2.

covariates, so that $X = (X_j, X_{-j})$. We define the following relative strength parameters:

$$k_{0\Delta Y,j} := \frac{\eta^2_{\Delta Y \sim U, X_j | X_{-j}, D=0} - \eta^2_{\Delta Y \sim X_j | X_{-j}, D=0}}{\eta^2_{\Delta Y \sim X_j | X_{-j}, D=0}}, \quad k_{0D,j} := \frac{R^2_{O_X \sim O_{X_{-j}} | D=0} - R^2_{O_{XU} \sim O_{X_{-j}} | D=0}}{1 - R^2_{O_X \sim O_{X_{-j}} | D=0}}.$$

Note the parameter $k_{0\Delta Y,j}$ measures how much additional variation in the untreated outcome evolution is explained by U, relative to the observed gains in variation explained by X_j itself. For example, $k_{0\Delta Y,j} = 1$ implies that further adding U to the trend regression would lead to similar additive gains in explanatory power to those observed by adding X_j. The parameter $k_{0D,j}$ uses the same logic to measure the relative gain in explanatory power with the treatment odds. As discussed in Section 3.3, this parameter also admits alternative interpretations in terms of the relative increase in the *average treatment odds* or the additional *imbalance between treated and control groups* attributable to U, relative to that attributable to X_j.

The usefulness of these measures is that they allow us to re-express the absolute strength of U in the OVB formula of Theorem 1 in terms of its relative strength as compared to that of X_j:

$$\eta^2_{\Delta Y \sim U | X, D=0} = k_{0\Delta Y,j} \left(\frac{\eta^2_{\Delta Y \sim X_j | X_{-j}, D=0}}{1 - \eta^2_{\Delta Y \sim X_j | X_{-j}, D=0}} \right), \quad 1 - R^2_{O_{XU} \sim O_X | D=0} = k_{0D,j} \left(\frac{1 - R^2_{O_X \sim O_{X_{-j}} | D=0}}{R^2_{O_X \sim O_{X_{-j}} | D=0}} \right).$$

Therefore, plausibility judgments on the *relative importance* of U as compared to X_j, both in explaining outcome evolution, and treatment odds, can be leveraged to bound the bias. Note that we can use different subgroups of covariates to bound $\eta^2_{\Delta Y \sim U | X, D=0}$ and $1 - R^2_{O_{XU} \sim O_X | D=0}$. Thus, if one group of covariates is known to be especially important for explaining treatment selection, while another group is known to be especially important for explaining outcome evolution, the researcher can incorporate this information directly into the sensitivity analysis.

Finally, one may also benchmark plausible values for the correlation of errors ρ_0. Notice that ρ_0 is not a measure of explanatory power of the confounders. Rather, it measures how systematically U shifts the odds of treatment and outcome evolution in the same direction. This is partly connected to the functional form of confounding. For example, if U enters the treatment and trend equations with different functional forms, this will generally attenuate the bias. To calibrate this judgment empirically, we may use as a reference the observed Pearson correlation between the outcome and treatment-odds errors induced by X_j, as measured by

$$\rho_{0,j} := \mathrm{Cor}\big(g_{0s} - g_{0s,-j},\ O_X - O_{X_{-j}} \,\big|\, D = 0\big)\,.$$

An advantage of benchmarking the bias factors ρ_0, $C_{0\Delta Y}$ and C_{0D} separately is that it allows researchers to entertain richer confounding scenarios. For example, one may posit that an unobserved

confounder is comparable to X_j in terms of explaining treatment selection and outcome evolution, while remaining agnostic about alignment. Further details can be found in Appendix C.1.

4.2 Comparing post-treatment bias against pre-treatment bias

We now connect the OVB framework to pre-trend extrapolation approaches such as Rambachan and Roth (2023). Existing methods typically bound the post-treatment bias by extrapolating deviations from parallel trends observed before treatment. Our decomposition shows that these extrapolation assumptions can instead be interpreted as restrictions on the underlying confounding mechanism itself, and also suggests natural alternative ways of both interpreting and augmenting them. We consider the simplest case of one additional pre-treatment period, and leave extensions to multiple time periods to future work.

Quantifying pre-treatment Bias

Consider three time periods $t \in \{0, 1, 2\}$, with treatment occurring after $t = 1$. To estimate the (placebo) effect in the pre-treatment period, we take $t = 0$ as the reference and estimate the effect at $t = 1$. Let θ^{pre} and θ_s^{pre} denote the corresponding long and short parameters. Since no treatment has yet occurred, $\theta^{\text{pre}} = 0$, and the *pre-treatment bias* from omitting confounders reduces to θ_s^{pre}. As discussed in Section 3, this bias characterizes the (average) deviation from parallel trends before treatment.

Let $\Delta Y^{\text{pre}} := Y_1 - Y_0$ denote the observed outcome evolution prior to treatment, and let $(X^{\text{pre}}, U^{\text{pre}})$ denote the corresponding set of covariates that would make parallel trends hold in the pre-treatment period. Define

$$g_0^{\text{pre}} := E[\Delta Y^{\text{pre}} \mid X^{\text{pre}}, U^{\text{pre}}, D = 0], \qquad g_{0s}^{\text{pre}} := E[\Delta Y^{\text{pre}} \mid X^{\text{pre}}, D = 0],$$

and,

$$O_{XU}^{\text{pre}} := \frac{P(D = 1|X^{\text{pre}}, U^{\text{pre}})}{P(D = 0|X^{\text{pre}}, U^{\text{pre}})}, \qquad O_X^{\text{pre}} := \frac{P(D = 1|X^{\text{pre}})}{P(D = 0|X^{\text{pre}})}.$$

By our bias characterization, the pre-treatment bias is governed by three quantities associated with the omitted confounders U^{pre}:

$$\rho_0^{\text{pre}} := \text{Cor}(g_0^{\text{pre}} - g_{0s}^{\text{pre}},\, O_{XU}^{\text{pre}} - O_X^{\text{pre}} \mid D = 0),\; C_{0\Delta Y}^{2,\text{pre}} := \eta^2_{\Delta Y^{\text{pre}} \sim U^{\text{pre}} | X^{\text{pre}}, D=0},\; C_{0D}^{2,\text{pre}} := \frac{1 - R^2_{O_{XU}^{\text{pre}} \sim O_X^{\text{pre}} | D=0}}{R^2_{O_{XU}^{\text{pre}} \sim O_X^{\text{pre}} | D=0}},$$

together with the pre-treatment scaling factor $S_0^{2,\text{pre}} := \sigma_{0s}^{2,\text{pre}} \nu_{0s}^{2,\text{pre}}$. Note we allow $(X^{\text{pre}}, U^{\text{pre}})$ to be different from (X, U), but in many cases, such as in our empirical application, they will be the same.

When $(X, U) = (X^{\text{pre}}, U^{\text{pre}})$, note the treatment selection component of the bias is immediately transportable across periods, since $C_{0D}^{2,\text{pre}} = C_{0D}^2$. In this case we also have $\nu_{0s}^{2,\text{pre}} = \nu_{0s}^2$.

Note that θ_s^{pre} does not fully identify the pre-treatment bias factors, but it constrains their product. That is, the triple of pre-treatment bias factors $(\rho_0^{\text{pre}}, C_{0\Delta Y}^{2,\text{pre}}, C_{0D}^{2,\text{pre}})$ lies in the set of triples consistent with the (nonzero) pre-treatment bias estimand θ_s^{pre}:

$$|\rho_0^{\text{pre}}| C_{0\Delta Y}^{\text{pre}} C_{0D}^{\text{pre}} S_0^{\text{pre}} = |\theta_s^{\text{pre}}|.$$

Therefore, two natural strategies emerge for extrapolating the pre-treatment bias to the post-treatment period in this setting. One is to directly transport the bias magnitude, as is currently done in pre-trend extrapolation approaches; another is to transport only (some of) the bias factors, while re-estimating the scaling factor S_0 from post-treatment data.

Extrapolating the bias magnitude

As discussed in Rambachan and Roth (2023), researchers may be willing to assume that the bias in the post-treatment period is no larger than k times the bias observed in the pre-treatment period. This corresponds to the assumption that $|\theta - \theta_s| \leq k|\theta_s^{\text{pre}}|$, and yields the following bounds

$$\theta_{\pm} = \theta_s \pm k|\theta_s^{\text{pre}}|,$$

where both θ_s and θ_s^{pre} are estimable from the data. Mapping this to unobserved confounders, this approach is equivalent to imposing a constraint on the full product of bias factors and the scale factor, i.e.,

$$|\rho_0|\, C_{0\Delta Y}\, C_{0D}\, S_0 \leq k|\rho_0^{\text{pre}}| C_{0\Delta Y}^{\text{pre}} C_{0D}^{\text{pre}}\, S_0^{\text{pre}} = k|\theta_s^{\text{pre}}|.$$

This connection allows us to also think about how different values of k could arise. Indeed, we can write

$$|\theta - \theta_s| = \left(\frac{|\rho_0|}{|\rho_0^{\text{pre}}|}\right)\left(\frac{C_{0\Delta Y}}{C_{0\Delta Y}^{\text{pre}}}\right)\left(\frac{C_{0D}}{C_{0D}^{\text{pre}}}\right)\left(\frac{S_0}{S_0^{\text{pre}}}\right)|\theta_s^{\text{pre}}|.$$

Defining

$$k_\rho := \frac{|\rho_0|}{|\rho_0^{\text{pre}}|}, \quad k_{\Delta Y} := \frac{C_{0\Delta Y}}{C_{0\Delta Y}^{\text{pre}}}, \quad k_D := \frac{C_{0D}}{C_{0D}^{\text{pre}}}, \quad k_S := \frac{S_0}{S_0^{\text{pre}}},$$

we have

$$|\theta - \theta_s| = \underbrace{k_\rho k_{\Delta Y} k_D k_S}_{k} |\theta_s^{\text{pre}}| = k|\theta_s^{\text{pre}}|.$$

Thus, the scalar multiplier k can be understood as collecting the relative changes in the components of the bias when moving from the pre-treatment period to the post-treatment period (further note that k_S is estimable from the data).

For example, suppose that the same observed and latent variables enter the pre-treatment and post-treatment comparisons. In this case, as discussed above, $C_{0D} = C_{0D}^{\text{pre}}$ and $\nu_{0s} = \nu_{0s}^{\text{pre}}$. Suppose further that the alignment term is stable, so that $|\rho_0| = |\rho_0^{\text{pre}}|$. This gives

$$|\theta - \theta_s| = \underbrace{\left(\frac{C_{0\Delta Y}}{C_{0\Delta Y}^{\text{pre}}}\right)}_{k_{\Delta Y}} \underbrace{\left(\frac{\sigma_{0s}}{\sigma_{0s}^{\text{pre}}}\right)}_{k_\sigma} |\theta_s^{\text{pre}}| = \underbrace{k_{\Delta Y} k_\sigma}_{k} |\theta_s^{\text{pre}}| = k|\theta_s^{\text{pre}}|.$$

In other words, the correct multiplier to transport the pre-trend bias to the post-treatment period should account for the change in how much variation the latent variables explain of the untreated outcome evolution, as well as the observed change in the residual variance of the outcome trend.

Extrapolating the bias factors

The previous considerations suggest another natural approach for extrapolating the bias. Rather than transporting the full magnitude of the pre-treatment bias directly, we can transport only the bias factors associated with the unobserved confounding mechanism, while allowing the observed scale component to change across periods. This yields the following bounds

$$\theta_\pm = \theta_s \pm k \left(\frac{S_0}{S_0^{\text{pre}}}\right) |\theta_s^{\text{pre}}|,$$

where θ_s, θ_s^{pre}, S_0^{pre}, and S_0 are estimable from the data, and $k = k_\rho k_{\Delta Y} k_D$.

This approach differs from the first in how it handles the scale factor. The first approach transports the full bias magnitude directly. The second approach, instead, estimates S_0^2 from the post-treatment data, and combines it with the transported bias factors from the pre-treatment period. Which restriction is preferable depends on the assumptions one is willing to make about the nature of confounding across periods.

Regardless of the choice of pre-trend extrapolation, the OVB decomposition of Section 3, together with benchmarking against observed covariates in Section 4.1, allows researchers to clearly translate these assumptions into statements about the underlying strength of confounding. For example, one can ask what kind of unobserved confounder would be required to generate the extrapolated post-treatment bias, in terms of its strength in treatment selection, outcome trends, and the correlation between the two, and whether such strength of confounding is comparable in magnitude to that of observed covariates.

5 Estimation and inference

The previous sections characterize the OVB formula and the associated sensitivity parameters at the population level. We now discuss estimation and inference in finite samples. In what follows, we assume we have an i.i.d. sample from the observed data distribution. Appendix E reports extensive simulation exercises to verify the properties of our estimators and confidence intervals.

5.1 Inference under direct restrictions on confounding

Given plausibility judgments on the magnitude of sensitivity parameters $|\rho_0|$, $C_{0\Delta Y}$, C_{0D}, we have the following bounds on θ:

$$\theta_\pm = \theta_s \pm |\rho_0|\, C_{0\Delta Y}\, C_{0D}\, S_0, \quad \text{with} \quad S_0^2 = \sigma_{0s}^2\, \nu_{0s}^2,$$

where θ_s and S_0^2 are estimable from the observed data.

We estimate these components using debiased machine learning (DML), which combines Neyman orthogonal scores with cross-fitting to enable the use of machine learning methods for nuisance estimation (Chernozhukov et al., 2018, 2026). Specifically, let $Z := (\Delta Y, D, X)$, $p := P(D = 1)$ and $\pi := P(D = 1 \mid X)$. Estimation proceeds via Algorithm 1, with orthogonal scores

$$\psi_{\theta_s}(Z; g_{0s}, \pi, p) = \left(\frac{D}{p} - \frac{(1-D)}{(1-p)}\frac{O_X}{O}\right)(\Delta Y - g_{0s}) - \frac{D\theta_s}{p}, \tag{7}$$

$$\psi_{\sigma_{0s}^2}(Z; g_{0s}, p) = \frac{1-D}{1-p}\left((\Delta Y - g_{0s})^2 - \sigma_{0s}^2\right), \tag{8}$$

$$\psi_{\nu_{0s}^2}(Z; \pi, p) = 2\frac{D}{p}\left(\frac{O_X}{O} - \nu_{0s}^2\right) - \frac{1-D}{1-p}\left(\left(\frac{O_X}{O}\right)^2 - \nu_{0s}^2\right). \tag{9}$$

We denote the resulting estimates as

$$\widehat{\theta}_s := \mathrm{DML}(\psi_{\theta_s}), \qquad \widehat{\sigma}_{0s}^2 := \mathrm{DML}(\psi_{\sigma_{0s}^2}) \qquad \widehat{\nu}_{0s}^2 := \mathrm{DML}(\psi_{\nu_{0s}^2}).$$

Let $\beta \in \{\theta_s, \sigma_{0s}^2, \nu_{0s}^2\}$ denote a generic target parameter and $\eta = (g_{0s}, \pi, p)$ the vector of nuisance parameters. The scores (7)-(9) admit the representation $\psi_\beta(Z;\eta) = \psi_\beta^a(Z;\eta)\,\beta + \psi_\beta^b(Z;\eta)$. An application of the results in Chernozhukov et al. (2018) for linear score functions yields the following lemma, which establishes the asymptotic linearity and normality of the DML estimators, and characterizes their influence functions.

Lemma 3 (Asymptotic linearity of the DML estimators)**.** *Suppose that Assumptions 3.1 and 3.2 from Chernozhukov et al. (2018) hold for each of the scores ψ_β above, and for the estimators of the*

Algorithm 1 DML for estimable components: DML(ψ_β)

Input: The Neyman orthogonal score $\psi_\beta(Z;\eta)$ where β is the target to estimate, and η is a vector of the nuisance parameters. A sample $\{Z_i\}_{i=1}^n$ and an appropriate number of folds L.

Sample Splitting: Randomly partition the sample into L folds of approximately equal size. Denote each fold by I_l and its complement by I_l^c, for $l = 1, \ldots, L$.

for $l = 1, \ldots, L$ **do**

Estimate the nuisance parameters using observations in I_l^c. Denote the estimates by $\widehat{\eta}_l$.

end for

Estimate: Construct the estimator $\widehat{\beta}$ as the solution of $0 = \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\psi_{\widehat{\beta}}(Z_i;\widehat{\eta}_l)$.

Return: The estimate $\widehat{\beta}$ and the estimated scores $\psi_{\widehat{\beta}}(Z_i;\widehat{\eta}_l)$ for each $i \in I_l$ and each l.

nuisance parameters $\widehat{\eta}_l$. Then, the DML estimators $\widehat{\beta}$ are asymptotically linear and Gaussian with the following properties

$$\sqrt{n}(\widehat{\beta}-\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underbrace{\frac{\psi_\beta(Z_i;\eta)}{-E[\psi_\beta^a(Z;\eta)]}}_{=:\varphi_\beta^0(Z_i)} + o_p(1) \xrightarrow{d} N(0,\sigma_{\varphi_\beta}^2), \text{ with } \sigma_{\varphi_\beta}^2 := E[(\varphi_\beta^0(Z))^2],$$

where $\varphi_\beta^0(\cdot)$ is the influence function. Following Theorem 3.2 in Chernozhukov et al. (2018), we denote the estimate of the influence function as $\widehat{\varphi}_{\widehat{\beta}}(\cdot) := -\psi_{\widehat{\beta}}(\cdot;\widehat{\eta}_l)/E_n[\psi^a(Z;\widehat{\eta}_l)]$ and use $\widehat{\sigma}_{\varphi_\beta}^2 := E_n\left[(\widehat{\varphi}_{\widehat{\beta}}(Z))^2\right]$ as the estimator for $\sigma_{\varphi_\beta}^2$.

Remark 2. *Note that the DML estimator for ν_{0s}^2 can also be used to estimate the chi-squared divergence of observed covariates, since $\nu_{0s}^2 - 1 = \chi^2(P_{X|D=1}\|P_{X|D=0})$—see Corollary 1.*

The plug-in estimator for the bounds thus takes the following form,

$$\widehat{\theta}_\pm := \widehat{\theta}_s \pm |\rho_0|\, C_{0\Delta Y}\, C_{0D}\, \sqrt{\widehat{\sigma}_{0s}^2\, \widehat{\nu}_{0s}^2}.$$

Since $\widehat{\theta}_\pm$ is a smooth function of asymptotically linear estimators, we can obtain confidence intervals for the bounds by applying the delta method.

Theorem 2 (Confidence intervals for the bounds)**.** *Under the Assumptions of Lemma 3, the estimator $\widehat{\theta}_\pm$ is asymptotically linear and Gaussian with the following properties*

$$\sqrt{n}(\widehat{\theta}_\pm - \theta_\pm) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{\theta_\pm}^0(Z_i) + o_p(1) \xrightarrow{d} N(0,\sigma_{\varphi_{\theta_\pm}}^2), \text{ with } \sigma_{\varphi_{\theta_\pm}}^2 = E[(\varphi_{\theta_\pm}^0(Z))^2],$$

where $\varphi^0_{\theta_\pm}(Z) = \varphi^0_{\theta_s}(Z) \pm |\rho_0| C_{0\Delta Y} C_{0D} \times \underbrace{\frac{1}{2S_0}\left(\sigma^2_{0s}\varphi^0_{\nu^2_{0s}}(Z) + \nu^2_{0s}\varphi^0_{\sigma^2_{0s}}(Z)\right)}_{=:\varphi^0_{S_0}(Z)}$.

Let Φ *denote the CDF of the standard normal and* α *denote the significance level. The confidence interval*

$$[l, u] = \left[\widehat{\theta}_- - \Phi^{-1}(1-\alpha)\sqrt{\frac{E[(\varphi^0_{\theta_-}(Z))^2]}{n}},\ \widehat{\theta}_+ + \Phi^{-1}(1-\alpha)\sqrt{\frac{E[(\varphi^0_{\theta_+}(Z))^2]}{n}}\right],$$

satisfies $P(\theta_- \in [l, \infty)) \to 1-\alpha$ *and* $P(\theta_+ \in (-\infty, u]) \to 1-\alpha$. *These results continue to hold if we replace* $E[(\varphi^0_{\theta_\pm}(Z))^2]$ *with* $E_n[(\widehat{\varphi}_{\widehat{\theta}_\pm}(Z))^2]$.

5.2 Inference under benchmarking restrictions on confounding

In the benchmarking analysis introduced in Section 4, additional components characterizing the strength of observed confounding are estimable from the data. We now discuss how to account for the uncertainty in estimating these components. We begin with benchmarking against observed covariates.

Given plausibility judgments on the relative strength parameters $k_{0\Delta Y,j}, k_{0D,j}$, the plug-in estimator of the bias bound now takes the following form:

$$\widehat{\theta}_\pm = \widehat{\theta}_s \pm |\widehat{\rho}_{0,j}| \times \sqrt{k_{0\Delta Y,j}\widehat{G}_{0\Delta Y,j} \times \frac{k_{0D,j}\widehat{G}_{0D,j}}{1 - k_{0D,j}\widehat{G}_{0D,j}}} \times \sqrt{\widehat{\sigma}^2_{0s}\,\widehat{\nu}^2_{0s}}$$

where,

$$\widehat{G}_{0\Delta Y,j} := \frac{\widehat{\sigma}^2_{0s,-j} - \widehat{\sigma}^2_{0s}}{\widehat{\sigma}^2_{0s}}, \quad \widehat{G}_{0D,j} := \frac{\widehat{\nu}^2_{0s} - \widehat{\nu}^2_{0s,-j}}{\widehat{\nu}^2_{0s,-j}}, \text{ and } \quad \widehat{\rho}_{0,j} = \frac{-(\widehat{\theta}_s - \widehat{\theta}_{s,-j})}{\sqrt{(\widehat{\sigma}^2_{0s,-j} - \widehat{\sigma}^2_{0s}) \times (\widehat{\nu}^2_{0s} - \widehat{\nu}^2_{0s,-j})}}.$$

Here $\theta_{s,-j}$ denotes the ATT estimand conditional on X_{-j} alone, and $\sigma^2_{0s,-j}$ and $\nu^2_{0s,-j}$ are the scaling factors in our OVB decomposition of $\theta_s - \theta_{s,-j}$. The orthogonal scores for these parameters have the same form as (7)-(9), with the only difference that X_{-j} is used for estimation in place of X.

We now move to benchmarking against pre-trend bias. For fixed k, the first approach yields the following plug-in estimator of the bias bound

$$\widehat{\theta}_\pm = \widehat{\theta}_s \pm k|\widehat{\theta}^{\text{pre}}_s|,$$

whereas the second approach gives

$$\widehat{\theta}_\pm = \widehat{\theta}_s \pm k\left(\frac{\widehat{S}_0}{\widehat{S}^{\text{pre}}_0}\right)|\widehat{\theta}^{\text{pre}}_s|, \qquad \text{with} \quad \widehat{S}^2_0 = \widehat{\sigma}^2_{0s}\,\widehat{\nu}^2_{0s}, \qquad \text{and} \quad \widehat{S}^{2,\text{pre}}_0 = \widehat{\sigma}^{2,\text{pre}}_{0s}\,\widehat{\nu}^{2,\text{pre}}_{0s}.$$

The orthogonal scores for the pre-treatment parameters again have the same form as (7)-(9), replacing the post-treatment data with pre-treatment data. In all these cases, $\widehat{\theta}_{\pm}$ is again a smooth function of asymptotically linear estimators whose influence functions were characterized above. Thus, inference for the bounds can be performed by applying the delta method. We provide the corresponding influence functions and asymptotic results in Appendix C.

5.3 Sensitivity statistics for routine reporting

The previous results allow users to perform detailed sensitivity analyses under a wide range of confounding scenarios. We now introduce two sensitivity statistics that answer the reverse question: what is the minimum strength of confounding required to invalidate the conclusions of a DiD study? These statistics require no assumptions about the strength of unobserved confounding; rather, they communicate what one must be prepared to believe in order to rule out confounding that would be problematic. They serve as quick summaries of the overall robustness of the ATT estimates against systematic biases (Cinelli and Hazlett, 2020, 2025a; Tsao et al., 2026; Chernozhukov et al., 2026).

Let $\text{CI}^{\max}_{1-\alpha,R^2_{\Delta Y},R^2_D}(\theta)$ denote the widest $(1-\alpha)$ confidence interval for the target parameter, constructed via Theorem 2, such that

$$|\rho| \leq 1, \qquad \eta^2_{\Delta Y \sim U|X,D=0} \leq R^2_{\Delta Y}, \qquad 1 - R^2_{O_{XU} \sim O_X|D=0} \leq R^2_D.$$

The first sensitivity statistic we consider answers the following question: what is the bare minimum selection strength that confounders need to have in order to overturn the results of a study? In other words, if we leave both the alignment and the trend components of the bias entirely unrestricted, what is the minimal strength of confounders with treatment selection alone, such that it would lead one to not reject the null hypothesis of interest, $H_0 : \theta = \theta^*$? This yields the *extreme robustness value* (XRV), defined as

$$\text{XRV}_{\theta^*,\alpha}(\theta) := \inf \left\{ \text{XRV} : \theta^* \in \text{CI}^{\max}_{1-\alpha,1,\text{XRV}}(\theta) \right\}$$

Any confounding scenario with selection strength below $\text{XRV}_{\theta^*,\alpha}(\theta)$ is logically incapable of overturning the original conclusions, *regardless* of how much variation such confounders explain of the untreated trend.

Leaving the association of the confounder with the trend completely unrestricted may be too conservative a scenario. Thus, the second sensitivity statistic we propose considers the minimum

strength of the confounder in explaining both outcome evolution and treatment selection jointly. The *robustness value* is defined as

$$\mathrm{RV}_{\theta^*,\alpha}(\theta) := \inf\left\{\mathrm{RV} : \theta^* \in \mathrm{CI}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\theta)\right\}.$$

Note that any unobserved confounding with trend and selection strength below $\mathrm{RV}_{\theta^*,\alpha}(\theta)$ is not capable of overturning the results of the study. More details on these sensitivity statistics are provided in Appendix D.

6 Applying the OVB framework to the sensitivity of DiD

We now return to the minimum wage example of Section 2, and show how to deploy the results of Sections 3 to 5 to answer the following questions: (i) How strong would unobserved confounders have to be to overturn the estimated negative effect in 2007? (ii) How does this required strength compare to that of the observed covariates? and, (iii) What does the 2006 pre-trend imply about the nature of unobserved confounding when read through the OVB decomposition?

6.1 Minimal sensitivity reporting

Table 4 shows our proposal for minimal sensitivity reporting of DiD estimates. The first columns report usual estimates of the treatment effect in 2007 under the assumption that parallel trends hold conditionally on the observed covariates alone. In addition to these estimates, we propose researchers report the extreme robustness value and the robustness value (Cinelli and Hazlett, 2020, 2025a; Tsao et al., 2026; Chernozhukov et al., 2026). As discussed in Section 5.3, these statistics quickly convey how robust the estimated effect is to the presence of omitted variables.

The extreme robustness value ($\mathrm{XRV}_{\theta^*=0,\ \alpha=0.05}$) of 0.2% means that confounders that explain at most 0.2% of the variation in treatment odds cannot explain away the estimated effect, at the 5% significance level, even if such confounders were to explain all leftover variation of the outcome evolution. Alternatively, the same number can be interpreted as confounders that induce at most a $\approx 0.2\%$ increase in the average odds of treatment among the treated. When the XRV is large, this may provide sufficient evidence for a causal effect in and of itself, by virtue of ruling out confounders that explain a large fraction of treatment assignment. This turns out not to be the case in our application, as it seems difficult to rule out confounders that change the odds of treatment by 0.2%.

We thus move to examining the minimum *joint* strength that confounders need to have, both in terms of explaining treatment selection and outcome evolution, in order to explain away the results.

The robustness value of $\text{RV}_{\theta^*=0,\ \alpha=0.05} = 4.38\%$ indicates that unobserved confounders explaining less than 4.38% of the residual variation in both treatment assignment and the outcome evolution of untreated units are not sufficiently strong to render the negative effect insignificant, in the sense of shifting the upper confidence bound to zero, at the 5% significance level. As before, the alternative characterizations of selection, discussed in Section 3.3, offer two additional interpretations of this value—it corresponds to a 4.58% relative increase in the observed average treatment odds among treated units, or, equivalently, to a 4.58% increase in covariate imbalance. For a single independent binary confounder, this corresponds to a standardized mean difference of about 0.2—which is not implausible in light of observed imbalances, as we have seen in Section 2.

6.2 Sensitivity contour plots and benchmarking

Overall, neither the XRV nor the RV allowed us to quickly rule out confounding of magnitudes that would be problematic. We thus turn to examining the full range of estimates we could have obtained under different confounding scenarios, using sensitivity contour plots (Imbens, 2003; Cinelli and Hazlett, 2020, 2025a; Chernozhukov et al., 2026). Since the estimated effect is negative, we focus on the upper confidence bound, which corresponds to the direction of bias relevant for overturning the original conclusion. The results are shown in Figure 3(a), where we consider the most conservative case, with alignment fixed at $|\rho_0| = 1$. The horizontal axis measures confounding in treatment selection, whereas the vertical axis measures confounding in the untreated outcome trend, both on the same R^2 scale. Each contour line corresponds to confounding scenarios that yield the same upper confidence bound for θ. The black triangle in the lower-left corner corresponds to the upper bound of the confidence interval reported in Table 4. The red dashed line represents the threshold scenarios that render the effect insignificant. Any confounding scenario below that line is not capable of overturning the original conclusion. The virtue of the contour plot is that the investigator can examine robustness to postulated confounding of *any* hypothetical strength, by simply reading off the value at the corresponding coordinates.

To better aid judgments about whether it is possible to rule out confounding strengths revealed to be problematic, the red diamonds show the scenarios implied by unobserved confounders comparable

in strength to observed covariates, as discussed in Section 4.1.[6] The results show that an adversarial latent confounder comparable to *region* could be sufficiently strong to completely overturn the results. Confounders comparable to median income (*lmedinc*) or race (*white*), though less extreme than the regional benchmark, could also bring the upper limit of the confidence interval just above the zero threshold. By contrast, a latent confounder comparable to poverty rate (*pov*) would not be sufficient to overturn the conclusion of a negative effect, though it would substantially reduce its magnitude. In addition, the blue dashed line represents the scenario implied by extrapolating the 2006 pre-trend deviation, and it shows what kinds of omitted variables reproduce that amount of bias. We can see, for instance, that such adversarial confounding would be substantially stronger than race, poverty or income, though weaker than regional confounding.

All the previous results are conservative, as they set the alignment to one. A less extreme scenario, again motivated by benchmarking, is to set the alignment to 0.3, which is still greater than the observed alignment of any of the benchmark covariates. Figure 3(b) shows that, under this specification, the original negative effect remains robust to unobserved confounding as strong as any of the observed benchmarks. Moreover, the 2006 pre-trend deviation itself can no longer be reproduced by confounding of such magnitudes. Taken together, these results show that the minimum wage estimate is not immune to omitted confounding of plausible strengths, but also that overturning the original analysis requires confounding of nontrivial magnitude. More importantly, it clarifies exactly what one needs to believe in order to sustain the estimated effect. For example, can we rule out biases as strong as the pre-trend observed for 2006? A critic arguing that the estimate is not credible must articulate what plausible omitted variables remain that are not only comparable in magnitude to regional confounding, but also sufficiently adversarial. Conversely, someone defending the estimate must not only argue against confounding of that magnitude, but also explain why the forces generating the 2006 deviation would no longer be operating in 2007. While researchers may not have much confidence in answering these questions, the sensitivity analysis shifts the discussion from a generic concern about whether parallel trends might fail, to a more disciplined and concrete discussion about the mechanisms required to overturn the conclusion.

[6]Note that here we do not propagate uncertainty due to estimated values in benchmarking, though this can be easily done using the results we discuss in Section 5.

7 Conclusion

In this paper we provide an omitted-variable bias framework for sensitivity analysis of difference-in-differences designs when unobserved confounding induces violations of parallel trends. We show that the bias in the average treatment effect on the treated admits an exact decomposition into an estimable scale factor and three bias factors, measuring confounding in treatment selection, confounding in untreated outcome evolution, and the alignment between these two channels. We also provide alternative characterizations of the treatment-selection component in terms of treatment odds and covariate imbalance. Building on these results, we derive sensitivity statistics for routine reporting, develop benchmarking procedures based on observed covariates and pre-treatment biases, and provide inference methods for the resulting bounds using debiased machine learning.

Possible extensions of our framework include accommodating multiple pre-treatment periods as well as staggered treatment adoption. In these cases, the OVB analysis we showed here can be performed period-by-period and aggregated across cohorts and time. Likewise, it is possible to perform richer benchmarking exercises in these settings, including calibration against the full sequence of pre-trends and pre-treatment covariates. Finally, extending the analysis to DiD with continuous and multi-valued treatments is also an interesting direction for future work.

References

Bach, P., Klaassen, S., Kueck, J., Mattes, M., and Spindler, M. (2025). Sensitivity analysis for treatment effects in difference-in-differences models using riesz representation. *arXiv preprint arXiv:2510.09064*.

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., and Sant'Anna, P. H. (2025). Difference-in-differences designs: A practitioner's guide. *arXiv preprint arXiv:2503.13323*.

Bilinski, A. and Hatfield, L. A. (2018). Seeking evidence of absence: Reconsidering tests of model assumptions. *arXiv preprint arXiv:1805.03273*, 8.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. (2026). Long story short: Omitted variable bias in causal machine learning. *The Review of Economics and Statistics*.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67.

Cinelli, C. and Hazlett, C. (2025a). An omitted variable bias framework for sensitivity analysis of instrumental variables. *Biometrika*, 112(2):asaf004.

Cinelli, C. and Hazlett, C. (2025b). The risks of informal benchmarking. R vignette.

Cinelli, C., LaPierre, N., Hill, B. L., Sankararaman, S., and Eskin, E. (2022). Robust mendelian randomization in the presence of residual population stratification, batch effects and horizontal pleiotropy. *Nature communications*, 13(1):1093.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Dette, H. and Schumann, M. (2020). Difference-in-differences estimation under non-parallel trends. Technical Report 22/2020, SFB 823 Discussion Paper.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27(3):368–377.

Dube, A. and Lindner, A. (2024). Minimum wages in the 21st century. *Handbook of Labor Economics*, 5:261–383.

Goldsmith-Pinkham, P. (2024). Tracking the credibility revolution across fields. *arXiv preprint arXiv:2405.20604*.

Huang, M. and Pimentel, S. D. (2025). Variance-based sensitivity analysis for weighting estimators results in more informative bounds. *Biometrika*, 112(1):asae040.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

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

Neumark, D. and Shirley, P. (2022). Myth or measurement: What does the new minimum wage research say about minimum wages and job loss in the united states? *Industrial Relations: A Journal of Economy and Society*, 61(4):384–417.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.

Rényi, A. (1959). New version of the probabilistic generalization of the large sieve. *Acta Math. Hung*, 10(1-2):217–226.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322.

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

Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

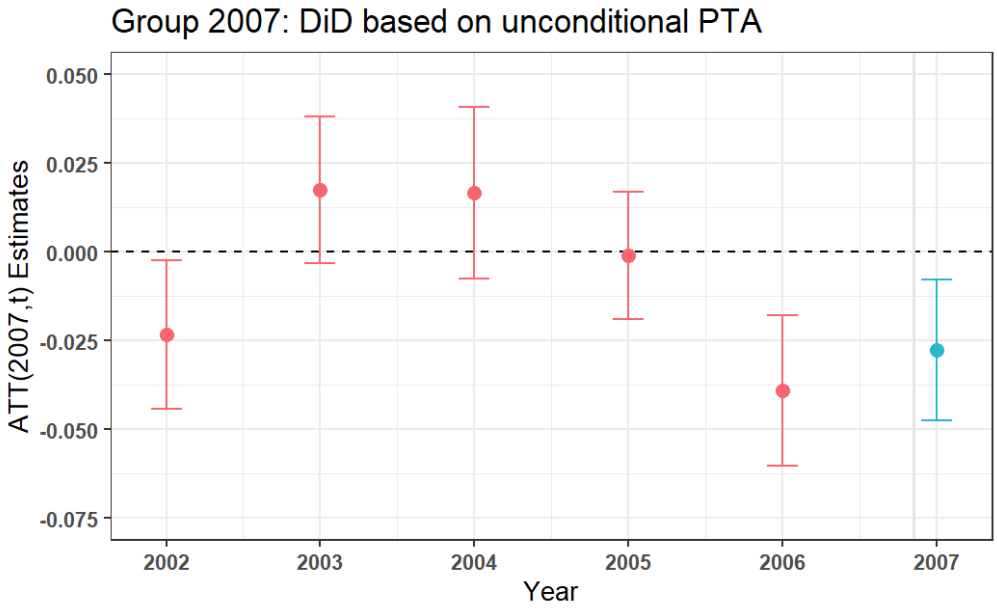
Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

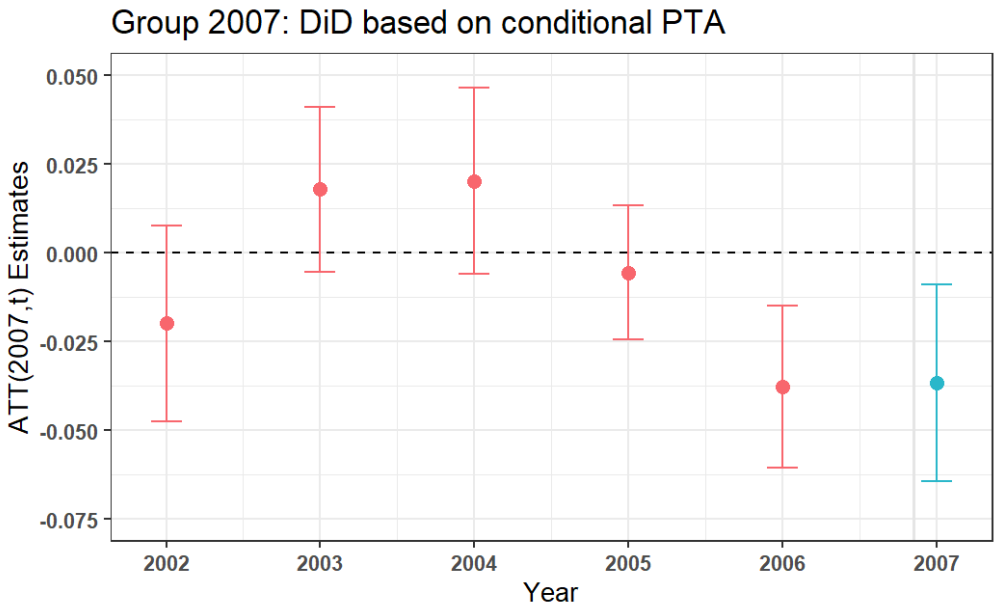
Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Tsao, D., Perry, R., and Cinelli, C. (2026). On the minimum strength of (unobserved) covariates to overturn an insignificant result. *Statistical Science*.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes: with applications to statistics*, pages 16–28. Springer.



(a) Unconditional Parallel Trends, never-treated as the control



(b) Conditional Parallel Trends, never-treated as the control

Figure 1: The effect of the minimum wage on teen employment estimated under the unconditional parallel trends assumption (Panel (a)) and the conditional parallel trends assumption (Panel (b)) using the **never-treated** as the comparison group. Estimates are obtained via DML, using the canonical DiD design with the preceding year as the base period. Red lines give point estimates and uniform 95% confidence bands for pre-treatment periods. Blue lines provide point estimates and uniform 95% confidence bands for the treatment effect of increasing the minimum wage.

	Year	$\hat{\theta}_s$	Std. Error	p-value	Confidence Bounds
Pre	2002	-0.0233	0.0078	0.0030	[-0.0442; -0.0023]
	2003	0.0175	0.0079	0.0260	[-0.0033; 0.0383]
	2004	0.0166	0.0089	0.0631	[-0.0076; 0.0409]
	2005	-0.0009	0.0072	0.8971	[-0.0189; 0.0170]
	2006	-0.0390	0.0077	0.0000	[-0.0603; -0.0178]
Post	2007	-0.0277	0.0073	0.0002	[-0.0476; -0.0078]

(a) Unconditional Parallel Trends, never-treated as the control

	Year	$\hat{\theta}_s$	Std. Error	p-value	Confidence Bounds
Pre	2002	-0.0207	0.0104	0.0473	[-0.0483; 0.0069]
	2003	0.0179	0.0086	0.0373	[-0.0053; 0.0411]
	2004	0.0203	0.0099	0.0394	[-0.0060; 0.0466]
	2005	-0.0056	0.0076	0.4612	[-0.0244; 0.0132]
	2006	-0.0377	0.0086	0.0000	[-0.0605; -0.0150]
Post	2007	-0.0366	0.0105	0.0005	[-0.0643; -0.0088]

(b) Conditional Parallel Trends, never-treated as the control

Table 1: Detailed numerical results corresponding to Figure 1.

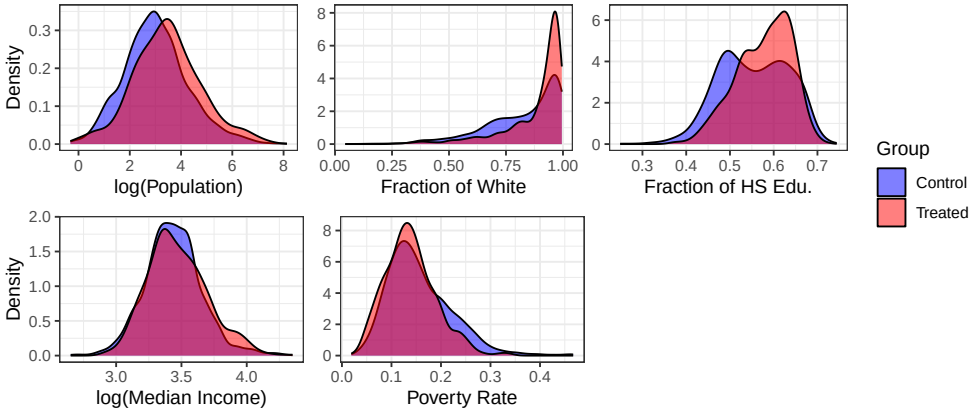


Figure 2: Distributions of county characteristics by treatment status. Never-treated as the control group.

	Untreated		Treated		Std. Mean Diff.	Var. Ratio Dev.
	μ_0	σ_0^2	μ_1	σ_1^2	$\frac{\mu_1-\mu_0}{\sqrt{(\sigma_1^2+\sigma_0^2)/2}}$	$\frac{\sigma_1^2}{\sigma_0^2}-1$
Midwest	0.336	0.223	0.483	0.250	0.303	0.120
South	0.593	0.241	0.301	0.212	-0.613	-0.128
West	0.072	0.067	0.216	0.169	0.419	1.536
Population (1000s)	53.425	23,438	84.142	32,850	0.183	0.402
log(Population)	3.018	1.584	3.467	1.781	0.346	0.125
Median Inc. (1000s)	31.889	54.565	33.080	66.654	0.153	0.222
log(Median Inc.)	3.438	0.048	3.471	0.053	0.150	0.111
White	0.826	0.026	0.885	0.018	0.397	-0.331
HS Graduates	0.553	0.006	0.577	0.004	0.336	-0.382
Poverty Rate	0.157	0.004	0.138	0.003	-0.324	-0.378

Table 2: This table reports covariate balance between 584 counties (29.8%) in states that increased the minimum wage above the federal level in 2007 (treated) and $1,377$ never-treated counties (controls). From left to right, the columns report group means, variances, standardized mean differences, and deviations of the variance ratio from one.

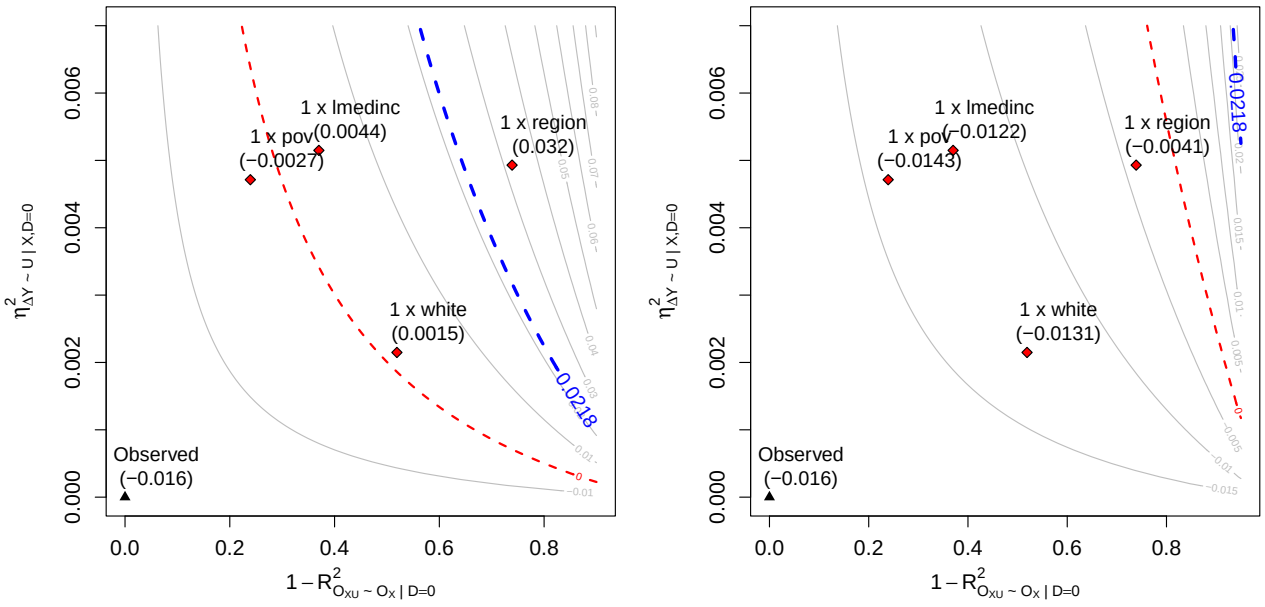
	Bias (from X_j)	Alignment ($\rho_{0,j}$)	Trend ($C_{0\Delta Y,j}$)	Imbalance ($C_{0D,j}$)
	$\theta_{s,j} - \theta_{s,\emptyset}$	$\text{Cor}\left(g_{0s,j}, O_{X_j} \mid D=0\right)$	$\sqrt{\eta^2_{\Delta Y \sim X_j \mid D=0}}$	$\sqrt{\chi^2(f_{X_j\mid 1}\|\|f_{X_j\mid 0})}$
Region (Overall)	-0.0010	0.291	0.036	0.589
	(0.0015)	(0.425)	(0.023)	(0.036)
Region (Midwest)	0.0016	-1.000	0.032	0.263
	(0.0006)	(0.498)	(0.021)	(0.030)
Region (South)	0.0013	-1.000	0.015	0.504
	(0.0012)	(1.921)	(0.030)	(0.029)
Region (West)	-0.0025	1.000	0.033	0.432
	(0.0011)	(0.611)	(0.028)	(0.039)
Population (1000s)	-0.0016	0.765	0.045	0.295
	(0.0007)	(0.373)	(0.010)	(0.032)
log(Population)	-0.0017	1.000	0.023	0.286
	(0.0007)	(1.834)	(0.027)	(0.032)
Median Inc. (1000s)	-0.0002	0.051	0.110	0.179
	(0.0006)	(0.186)	(0.018)	(0.034)
log(Median Inc.)	-0.0002	0.058	0.110	0.162
	(0.0006)	(0.210)	(0.019)	(0.037)
White	0.0011	-0.307	0.060	0.406
	(0.0010)	(0.261)	(0.026)	(0.029)
HS degree	-0.0026	0.607	0.072	0.389
	(0.0010)	(0.246)	(0.016)	(0.026)
Poverty rate	0.0001	-0.029	0.076	0.251
	(0.0008)	(0.283)	(0.026)	(0.026)
X_{all}	-0.0089	0.194	0.117	2.548
	(0.0079)	(0.166)	(0.045)	(0.170)

Note: $S_{0,\emptyset} = \sqrt{\sigma^2_{0s,\emptyset}\nu^2_{0s,\emptyset}} = 0.154$ with an SE of 0.006.
$\widehat{\sigma}^2_{0s,\emptyset} = 0.024$ with an SE of 0.002, and $\widehat{\nu}^2_{0s,\emptyset} = 1$ because there are no covariates.

Table 3: Decomposition of the observed confounding strength. For each covariate, the first row reports the estimates and the second row reports the corresponding standard errors. The decomposition follows from $\theta_{s,j} - \theta_{s,\emptyset} = -\rho_{0,j} C_{0\Delta Y,j} C_{0D,j} S_{0,\emptyset}$.

Results Under Conditional Parallel Trends			Robustness Value	
Short Estimate ($\widehat{\theta}_{s,2007}$)	**Std. Error**	**Confidence Interval**	$\text{RV}_{\theta^*=0,\ \alpha=0.05}$	$\text{XRV}_{\theta^*=0,\ \alpha=0.05}$
-0.0366	0.0105	[-0.0572; -0.0160]	4.38%	0.20%

Table 4: Minimal Sensitivity Reporting: Increase in Minimum Wage on Teen Employment.
We use random forests for both outcome evolution and propensity scores, with tuning parameters chosen by cross-validation (CV).



(a) Upper limit conf. bound, $|\rho_0| = 1$.

(b) Upper limit conf. bound, $|\rho_0| = 0.3$.

Figure 3: Sensitivity contour plots for the minimum wage example at a significance level of $\alpha = 0.05$. The blue dashed lines denote the upper confidence bound of $\widehat{\theta}_{\pm} = \widehat{\theta}_s \pm |\widehat{\theta}_s^{\text{pre}}|$ where $\widehat{\theta}_s^{\text{pre}}$ is treated as fixed when constructing the confidence bound for $\widehat{\theta}_{\pm}$.

A Preliminaries

A.1 Notation

Let Y denote the outcome, D the treatment indicator, X the observed covariates, and U the unobserved covariates. To simplify notation, we denote $p = P(D = 1)$, $\pi = P(D = 1 \mid X)$, $O = p/(1-p)$, and $O_X = \pi/(1-\pi)$. We recall the following standard definitions and results, which will be used in subsequent derivations.

Linear Projection and Residuals. Let $\widehat{Y^X} := X^\top \beta$ denote the linear projection of Y onto X, where $\beta := \arg\min_b E[(Y - X^\top b)^2] = (E[XX^\top])^{-1}E[XY]$. We define the residual as $Y^{\perp X} := Y - \widehat{Y^X}$, which represents the component of Y orthogonal to the linear span of X.

Radon-Nikodym Derivative. Let $(\mathscr{X}, \mathscr{A})$ be a measurable space on which two σ-finite measures, μ and ν, are defined. If ν is absolutely continuous with respect to μ, denoted $\nu \ll \mu$, then by the Radon-Nikodym theorem, there exists an $\mathscr{A}$-measurable function $f : \mathscr{X} \to [0, \infty)$ such that for any measurable set $A \in \mathscr{A}$, $\nu(A) = \int_A f \, \mathrm{d}\mu$. The function f, denoted $f = \mathrm{d}\nu/\,\mathrm{d}\mu$, is called the Radon-Nikodym derivative of ν with respect to μ.

Neyman χ^2-divergence. Let $P_{X|D=0}$ and $P_{X|D=1}$ denote the distributions of X in the control and treated groups, respectively. Suppose $P_{X|D=1} \ll P_{X|D=0}$. We define the Neyman χ^2-divergence of $P_{X|D=1}$ from $P_{X|D=0}$ as $\chi^2(P_{X|D=1} \| P_{X|D=0}) = \int \left(\frac{\mathrm{d}P_{X|D=1}}{\mathrm{d}P_{X|D=0}} - 1\right)^2 \mathrm{d}P_{X|D=0}$.

Coefficient of Variation (CV). Let Q_X denote the distribution of X. The coefficient of variation, CV_X, is a standardized measure of dispersion of Q_X, defined as $\mathrm{CV}_X := \frac{\sigma_X}{\mu_X}$, where μ_X and σ_X denote the mean and standard deviation of X, respectively.

Riesz-Frechet Representation Theorem. Let H be a Hilbert space over $\mathbb{R}$ with an inner product $\langle \cdot, \cdot \rangle$, that is complete w.r.t. the norm $\|\cdot\|$ generated by this inner product. Let T be a continuous linear functional on H. Then there exists a unique $w \in H$ such that for every $g \in H$, we have $T(g) = \langle g, w \rangle$.

A.2 Definition and properties of R^2

We use $R^2_{Y\sim X}$ to denote the (possibly uncentered) R^2 of the orthogonal linear projection of a random variable Y on a random vector X. That is, writing the orthogonal decomposition $Y = \widehat{YX} + Y^{\perp X}$,

$$R^2_{Y\sim X} := \frac{E[(\widehat{YX})^2]}{E[Y^2]} = 1 - \frac{E[(Y^{\perp X})^2]}{E[Y^2]} = \frac{\left(E[(\widehat{YX})Y]\right)^2}{E[Y^2]E[(\widehat{YX})^2]}.$$

If X is univariate we also have the equality,

$$R^2_{Y\sim X} = \frac{E[XY]^2}{E[Y^2]E[X^2]}.$$

We define the *partial* R^2 of Y with U given X as:

$$R^2_{Y\sim U|X} := \frac{R^2_{Y\sim U+X} - R^2_{Y\sim X}}{1 - R^2_{Y\sim X}}.$$

It is easy to show that this equals the regular R^2 of $Y^{\perp X}$ with $U^{\perp X}$. Note that $(Y^{\perp X})^{\perp (U^{\perp X})} = Y^{\perp U,X}$, so,

$$R^2_{Y\sim U|X} = 1 - \frac{E[(Y^{\perp U,X})^2]}{E[(Y^{\perp X})^2]}.$$

We can use these definitions to generalize the usual (centered) R^2. That is, if we want to recover the centered R^2, we simply partial out the constant. Formally,

$$R^2_{Y\sim X|1} = \frac{\mathrm{Var}(\widehat{YX1})}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(Y^{\perp X1})}{\mathrm{Var}(Y)} = \mathrm{Cor}^2(\widehat{YX1}, Y),$$

where, for univariate X, $R^2_{Y\sim X|1} = \mathrm{Cor}^2(Y, X)$. Note that some other works use $R^2_{Y\sim X}$ to denote the centered version (see, e.g., Cinelli and Hazlett, 2020, 2025a).

Notice that if X is a vector including the constant, then

$$R^2_{Y\sim U|X} = 1 - \frac{E[(Y^{\perp U,X})^2]}{E[(Y^{\perp X})^2]} = 1 - \frac{\mathrm{Var}(Y^{\perp U,X})}{\mathrm{Var}(Y^{\perp X})},$$

which is centered.

We define the nonparametric analogue of the (uncentered) R^2 as:

$$\eta^2_{Y\sim X} := \frac{E[(E[Y \mid X])^2]}{E[Y^2]},$$

which is equivalent to the linear (uncentered) R^2 from the projection of Y on $E[Y \mid X]$. We define the nonparametric *partial* R^2 of Y with U given X as:

$$\eta^2_{Y\sim U|X} := \frac{\eta^2_{Y\sim (X,U)} - \eta^2_{Y\sim X}}{1 - \eta^2_{Y\sim X}},$$

which quantifies how much of the residual variation in Y is explained by U, after accounting for X.

Note that the nonparametric partial R^2 is automatically centered, since $E[Y - E[Y \mid X]] = E[E[Y|X,U] - E[Y|X]] = 0$ by the tower property of conditional expectations.

Proposition 1 (Nonparametric and linear R^2)**.**

(i) $\eta^2_{Y\sim X} = 1 - \frac{E\left[(Y-E[Y|X])^2\right]}{E[Y^2]} = R^2_{Y\sim E[Y|X]}$.

(ii) $\eta^2_{Y\sim U|X} = 1 - \frac{E\left[(Y-E[Y|X,U])^2\right]}{E[(Y-E[Y|X])^2]} = R^2_{Y-E[Y|X]\sim E[Y|X,U]-E[Y|X]}$.

Proof.

$$\begin{aligned} R^2_{Y\sim E[Y|X]} &:= 1 - \frac{E\left[(Y - E[Y \mid X])^2\right]}{E[Y^2]} \\ &= \frac{E[Y^2] - \left(E[Y^2] + E[(E[Y \mid X])^2] - 2E[Y E[Y \mid X]]\right)}{E[Y^2]} \\ &\overset{\text{LTE}}{=} \frac{E[(E[Y \mid X])^2]}{E[Y^2]} = \eta^2_{Y\sim X}. \end{aligned}$$

Additionally,

$$\begin{aligned} R^2_{Y-E[Y|X]\sim E[Y|X,U]-E[Y|X]} &:= 1 - \frac{E\left[(Y - E[Y \mid X,U])^2\right]}{E[(Y - E[Y \mid X])^2]} \\ &= \frac{E[(Y - E[Y \mid X])^2] - E\left[(Y - E[Y \mid X,U])^2\right]}{E[(Y - E[Y \mid X])^2]} \\ &\overset{\text{LTE}}{=} \frac{E[(E[Y \mid X,U])^2] - E[(E[Y \mid X])^2]}{E[Y^2] - E[(E[Y \mid X])^2]} \\ &= \frac{\eta^2_{Y\sim(X,U)} - \eta^2_{Y\sim X}}{1 - \eta^2_{Y\sim X}} = \eta^2_{Y\sim U|X}. \end{aligned}$$

□

A.3 Important relationships

Proposition 2 (Relationship between selection odds)**.** *Under the weak overlap condition of Assumption 1, the following identities hold:*

(1) ***Odds-ratio as change of measure.***

$$\frac{O_{XU}}{O_X} = \frac{\mathrm{d}P_{U|X,D=1}}{\mathrm{d}P_{U|X,D=0}}.$$

(2) ***Short odds as the projection of long odds.***

$$O_X = E[O_{XU} \mid X, D = 0].$$

(3) ***Control-to-treated expectation identity for odds.***

$$E[O^2_{XU} \mid X, D = 0] = O_X \times E[O_{XU} \mid X, D = 1].$$

(4) ***From conditional to marginal odds identity.***

$$E[O_{XU}] = p \times (E[O_{XU} \mid D = 1] + 1) = (1 - p)E[O^2_{XU} \mid D = 0] + p.$$

Proof. We establish each of the four properties in turn.

(1) By Bayes' rule, the following holds:

$$\begin{aligned}\frac{O_{XU}}{O_X} &= \frac{P(D=1 \mid X,U)P(U \mid X)}{P(D=1 \mid X)} \times \frac{P(D=0 \mid X)}{P(D=0 \mid X,U)P(U \mid X)} \\ &= \frac{P(D=1,U \mid X)}{P(D=1 \mid X)} \times \frac{P(D=0 \mid X)}{P(D=0,U \mid X)} \\ &= \frac{P(U \mid X,D=1)}{P(U \mid X,D=0)} \overset{\text{def.}}{=} \frac{\mathrm{d}P_{U|X,D=1}}{\mathrm{d}P_{U|X,D=0}}.\end{aligned}$$

(2) We will use (1) to show the equality:

$$E\left[\frac{O_{XU}}{O_X} \mid X, D=0\right] \overset{(1)}{=} \int \frac{\mathrm{d}P_{U|X,D=1}(u)}{\mathrm{d}P_{U|X,D=0}(u)} \times \mathrm{d}P_{U|X,D=0}(u) = \int \mathrm{d}P_{U|X,D=1}(u) = 1,$$

$$\implies O_X = E[O_{XU} \mid X, D=0].$$

(3) We will use (1) to show the equality:

$$\begin{aligned}E\left[\left(\frac{O_{XU}}{O_X}\right)^2 \mid X, D=0\right] &\overset{(1)}{=} \int \left(\frac{\mathrm{d}P_{U|X,D=1}(u)}{\mathrm{d}P_{U|X,D=0}(u)}\right)^2 \times \mathrm{d}P_{U|X,D=0}(u) \\ &= \int \frac{\mathrm{d}P_{U|X,D=1}(u)}{\mathrm{d}P_{U|X,D=0}(u)} \times \mathrm{d}P_{U|X,D=1}(u) \\ &\overset{(1)}{=} E\left[\frac{O_{XU}}{O_X} \mid X, D=1\right],\end{aligned}$$

$$\implies E[O_{XU}^2 \mid X, D=0] = O_X \times E[O_{XU} \mid X, D=1].$$

(4) By the tower property, the following holds:

$$\begin{aligned}E[O_{XU}] &\overset{\text{LTE}}{=} E\left[E[O_{XU} \mid D]\right] \\ &= E[O_{XU} \mid D=1] \times p + E[O_{XU} \mid D=0] \times (1-p) \\ &\overset{(2)}{=} E[O_{XU} \mid D=1] \times p + O \times (1-p) \\ &= p \times (E[O_{XU} \mid D=1] + 1) \\ &\overset{(3)}{=} p + p \times \frac{E[O_{XU}^2 \mid D=0]}{O} \\ &= p + (1-p)E[O_{XU}^2 \mid D=0].\end{aligned}$$

□

Proposition 3 (Odds and χ^2-divergence)**.** *Under the weak overlap condition of Assumption 1, the following identities hold:*

(1) ***χ^2-divergence as squared coefficient of variation.***

$$\chi^2(P_{X,U|D=1} \| P_{X,U|D=0}) = \mathrm{Var}\left(\frac{O_{XU}}{O} \mid D=0\right) = CV^2_{O_{XU}|0}.$$

(2) ***χ^2-divergence as average conditional expected odds.***

$$\chi^2(P_{X,U|D=1}\|P_{X,U|D=0}) = E\left[\frac{O_{XU}}{O} \mid D=1\right] - 1,$$
$$\chi^2(P_{U|X,D=1}\|P_{U|X,D=0}) = E\left[\frac{O_{XU}}{O_X} \mid X, D=1\right] - 1.$$

(3) ***Conditional χ^2-weighted decomposition of odds increase.***

$$E[O_{XU} \mid D=1] - E[O_X \mid D=1] = E\left[\chi^2(P_{U|X,D=1}\|P_{U|X,D=0}) \times O_X \mid D=1\right].$$

(4) ***Marginal-conditional decomposition of χ^2-divergence.***

$$\chi^2(P_{X,U|D=1}\|P_{X,U|D=0}) = \chi^2(P_{X|D=1}\|P_{X|D=0}) + E_{P_{X|D=0}}\left[\left(\frac{\mathrm{d}P_{X|D=1}}{\mathrm{d}P_{X|D=0}}\right)^2 \chi^2(P_{U|X,D=1}\|P_{U|X,D=0})\right].$$

Proof. We establish each of the four properties in turn, using Proposition 2. Throughout, we let $Z := (X, U)$.

(1)

$$\begin{aligned}
\chi^2(P_{Z|1}\|P_{Z|0}) &:= \int \left(\frac{\mathrm{d}P_{Z|1}(z)}{\mathrm{d}P_{Z|0}(z)} - 1\right)^2 \mathrm{d}P_{Z|0}(z) \\
&= \int \left(\frac{\mathrm{d}P_{Z|1}(z)}{\mathrm{d}P_{Z|0}(z)}\right)^2 \times \mathrm{d}P_{Z|0}(z) + \int \mathrm{d}P_{Z|0}(z) - 2\int \frac{P_{Z|1}(z)}{P_{Z|0}(z)}\,\mathrm{d}P_{Z|0}(z) \\
&= \int \left(\frac{\mathrm{d}P_{Z|1}(z)}{\mathrm{d}P_{Z|0}(z)}\right)^2 \mathrm{d}P_{Z|0}(z) + 1 - 2 \\
&= E\left[\left(\frac{O_Z}{O}\right)^2 \mid D=0\right] - 1 \text{ by Proposition 2 (1)}, \\
&= E\left[\left(\frac{O_Z}{O}\right)^2 \mid D=0\right] - \left(E\left[\frac{O_Z}{O} \mid D=0\right]\right)^2 \text{ by Proposition 2 (2)}, \\
&= \mathrm{Var}\left(\frac{O_Z}{O} \mid D=0\right) = \frac{\mathrm{Var}(O_Z \mid D=0)}{O^2} \\
&= \frac{\mathrm{Var}(O_Z \mid D=0)}{E^2[O_Z \mid D=0]} \text{by Proposition 2 (2)}, \\
&\overset{\text{def.}}{=} \mathrm{CV}^2_{O_{XU}|0}.
\end{aligned}$$

Following similar steps, we can show

$$\chi^2(P_{U|X,D=1}\|P_{U|X,D=0}) = \mathrm{Var}\left(\frac{O_{XU}}{O_X} \mid X, D=0\right) = \mathrm{CV}^2_{O_{XU}|X,0}.$$

(2)

$$\begin{aligned}
\chi^2(P_{Z|1}\|P_{Z|0}) &\overset{(1)}{=} \mathrm{Var}\left(\frac{O_Z}{O} \mid D=0\right) = \frac{\mathrm{Var}(O_Z \mid D=0)}{O^2} \\
&= \frac{E[O_Z^2 \mid D=0] - O^2}{O^2} \text{ by Proposition 2 (2)},
\end{aligned}$$

$$= E\left[\frac{O_Z}{O} \mid D=1\right] - 1 \text{ by Proposition 2 (3).}$$

Following similar steps, we can show

$$\chi^2(P_{U|X,D=1} \| P_{U|X,D=0}) = E\left[\frac{O_{XU}}{O_X} \mid X, D=1\right] - 1.$$

(3)

$$\begin{aligned}
&E[O_{XU} \mid D=1] - E[O_X \mid D=1] = E\left[\left(\frac{O_{XU}}{O_X} - 1\right) \times O_X \mid D=1\right] \\
&\overset{\text{LTE}}{=} E\left[\left(E\left[\frac{O_{XU}}{O_X} \mid X, D=1\right] - 1\right) \times O_X \mid D=1\right] \\
&\overset{(2)}{=} E\left[\chi^2(P_{U|X,D=1} \| P_{U|X,D=0}) \times O_X \mid D=1\right].
\end{aligned}$$

(4)

$$\begin{aligned}
&\chi^2(P_{X,U|1} \| P_{X,U|0}) - \chi^2(P_{X|1} \| P_{X|0}) \overset{(2)}{=} E\left[\frac{O_{XU}}{O} \mid D=1\right] - E\left[\frac{O_X}{O} \mid D=1\right] \\
&\overset{(3)}{=} \frac{1}{O} E\left[\chi^2(P_{U|X,D=1} \| P_{U|X,D=0}) \times O_X \mid D=1\right] \\
&= E\left[\chi^2(P_{U|X,D=1} \| P_{U|X,D=0}) \times \left(\frac{\mathrm{d}P_{X|1}}{\mathrm{d}P_{X|0}}\right) \mid D=1\right] \text{ by Proposition 2 (1),} \\
&= \int \chi^2(P_{U|X,D=1} \| P_{U|X,D=0}) \times \left(\frac{\mathrm{d}P_{X|1}}{\mathrm{d}P_{X|0}}\right) \mathrm{d}P_{X|1} \\
&= \int \chi^2(P_{U|X,D=1} \| P_{U|X,D=0}) \times \left(\frac{\mathrm{d}P_{X|1}}{\mathrm{d}P_{X|0}}\right)^2 \mathrm{d}P_{X|0} \\
&= E_{P_{X|0}}\left[\left(\frac{\mathrm{d}P_{X|1}}{\mathrm{d}P_{X|0}}\right)^2 \chi^2(P_{U|X,D=1} \| P_{U|X,D=0})\right].
\end{aligned}$$

□

A.4 Standard DiD assumptions and identification of ATT

This section presents the standard identifying assumptions for the ATT in the canonical DiD design and shows that, under these assumptions, our target estimand admits a causal interpretation.

Assumption A.1 (Consistency)**.** *Observed outcomes are generated as* $Y := Y(D)$.

Assumption A.2 (No anticipation)**.** *Treatment has no effect on pre-treatment outcomes.*

$$E[Y_1(0) \mid D=1, X, U] = E[Y_1(1) \mid D=1, X, U] \quad a.s.$$

Assumption A.3 (Conditional parallel trends assumption)**.** *Let* $\Delta Y(0) := Y_2(0) - Y_1(0)$ *denote the untreated potential outcome evolution, and assume that*

$$E[\Delta Y(0) \mid D=1, X, U] = E[\Delta Y(0) \mid D=0, X, U] \quad a.s.$$

Assumption A.4 (Strong overlap)**.** *There exists* $\epsilon > 0$ *such that*

$$p := P(D = 1) \geq \epsilon, \text{ and } \pi_{XU} := P(D = 1 \mid X, U) \leq 1 - \epsilon, \quad a.s.$$

Remark 3. *Assumption A.4 is stronger than what is needed to derive the OVB formula of Theorem 1, in which we impose only the weak overlap condition of Assumption 1.*

Proposition 4 (DiD identifies the ATT)**.** *Let* $\Delta Y := Y_2 - Y_1$ *denote the observed outcome evolution, and let* $g_0(X, U) := E[\Delta Y \mid D = 0, X, U]$*. Under Assumptions A.2-A.4, the ATT is identified by*

$$\theta := E\left[\Delta Y - g_0(X, U) \mid D = 1\right].$$

Proof. By no anticipation,

$$\begin{aligned}
\text{ATT} &= E\left[Y_2(1) - Y_2(0) | D = 1\right] \\
&= E[Y_2(1) - (Y_1(1) - Y_1(1)) - Y_2(0) + (Y_1(0) - Y_1(0)) \mid D = 1] \\
&= E[\Delta Y(1) \mid D = 1] - E[\Delta Y(0) \mid D = 1] + E[Y_1(1) - Y_1(0) \mid D = 1] \\
&= E[\Delta Y(1) \mid D = 1] - E[\Delta Y(0) \mid D = 1]
\end{aligned}$$

Then, we have

$$\begin{aligned}
\text{ATT} &= E[\Delta Y(1) \mid D = 1] - E[\Delta Y(0) \mid D = 1] \\
&= E[\Delta Y \mid D = 1] - E[\Delta Y(0) \mid D = 1] \text{ by consistency,} \\
&= E[\Delta Y \mid D = 1] - E\left[E[\Delta Y(0) \mid X, U, D = 1] \mid D = 1\right] \text{ by LTE,} \\
&= E[\Delta Y \mid D = 1] - E\left[E[\Delta Y(0) \mid X, U, D = 0] \mid D = 1\right] \text{ by conditional PTA,} \\
&= E[\Delta Y - g_0(X, U) \mid D = 1] \\
&= \theta.
\end{aligned}$$

□

B Deferred proofs

For simplicity, let $\alpha_0 := O_{XU}/O$, and $\alpha_{0s} := O_X/O$.

Proof of Lemma 1.

We begin by showing that $\theta_{0s} = E[g_{0s}\alpha_{0s} \mid D = 0]$.

$$\begin{aligned}
\theta_{0s} := E\left[g_{0s}(X) \mid D = 1\right] &= E\left[\frac{D}{p} \times g_{0s}(X)\right] = E\left[\frac{P(D = 1 \mid X)}{p} \times g_{0s}(X)\right] \text{ by LTE,} \\
&= E\left[\left(\frac{P(D = 1 \mid X)}{P(D = 0 \mid X)} \times \frac{1 - p}{p}\right) \times \frac{P(D = 0 \mid X)}{1 - p} \times g_{0s}(X)\right]
\end{aligned}$$

$$
\begin{aligned}
&= E\left[\alpha_{0s}(X) g_{0s}(X) \times \frac{P(D=0 \mid X)}{1-p}\right] \\
&= E\left[\alpha_{0s}(X) g_{0s}(X) \times \frac{1-D}{1-p}\right] \text{ by LTE,} \\
&= E[g_{0s}\alpha_{0s} \mid D=0].
\end{aligned}
$$

The same argument shows that $\theta_0 = E[g_0\alpha_0 \mid D=0]$. Next, by Proposition 2 (2), we have that $\alpha_{0s} = E[\alpha_0 \mid X, D=0]$ and that $E[\alpha_0 \mid D=0] = E[\alpha_{0s} \mid D=0] = 1$. □

Proof of Theorem 1.

We first show that $\theta - \theta_s = -\operatorname{Cov}(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D=0)$.

$$
\begin{aligned}
\theta - \theta_s &= -(\theta_0 - \theta_{0s}) \\
&= E[g_{0s}\alpha_{0s} \mid D=0] - E[g_0\alpha_0 \mid D=0] \\
&= E[g_{0s}\alpha_{0s} \mid D=0] - [E[(g_{0s} + g_0 - g_{0s})(\alpha_{0s} + \alpha_0 - \alpha_{0s}) \mid D=0]] \\
&= -\left[E[g_{0s}(\alpha_0 - \alpha_{0s}) \mid D=0] + E[\alpha_{0s}(g_0 - g_{0s}) \mid D=0] + E[(g_0 - g_{0s})(\alpha_0 - \alpha_{0s}) \mid D=0]\right] \\
&= -E[(g_0 - g_{0s})(\alpha_0 - \alpha_{0s}) \mid D=0] \\
&= -\operatorname{Cov}(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D=0).
\end{aligned}
$$

The second-to-last equality uses that given $D=0$, g_{0s} is orthogonal to $(\alpha_0 - \alpha_{0s})$ and α_{0s} is orthogonal to $(g_0 - g_{0s})$. The last equality uses that $E[\alpha_0 \mid D=0] = E[\alpha_{0s} \mid D=0] = 1$.

Next, we decompose this covariance in the following way:

$$
\begin{aligned}
|\theta - \theta_s|^2 &= \operatorname{Cov}^2(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D=0) \\
&= \underbrace{\frac{\operatorname{Cov}^2(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D=0)}{\operatorname{Var}(g_0 - g_{0s} \mid D=0)\operatorname{Var}(\alpha_0 - \alpha_{0s} \mid D=0)}}_{=:\rho_0^2} \times \operatorname{Var}(g_0 - g_{0s} \mid D=0)\operatorname{Var}(\alpha_0 - \alpha_{0s} \mid D=0)
\end{aligned}
$$

Since $E[g_0 - g_{0s} \mid D=0] = E[\alpha_0 - \alpha_{0s} \mid D=0] = 0$,

$$
\begin{aligned}
&= \rho_0^2 E[(g_0 - g_{0s})^2 \mid D=0] E[(\alpha_0 - \alpha_{0s})^2 \mid D=0] \\
&= \rho_0^2 \times \underbrace{\frac{E[(g_0 - g_{0s})^2 \mid D=0]}{E[(\Delta Y - g_{0s})^2 \mid D=0]}}_{=:C_{0\Delta Y}^2} \times \underbrace{\frac{E[(\alpha_0 - \alpha_{0s})^2 \mid D=0]}{E[\alpha_{0s}^2 \mid D=0]}}_{=:C_{0D}^2} \times S_0^2 \\
&= \rho_0^2 C_{0\Delta Y}^2 C_{0D}^2 S_0^2,
\end{aligned}
$$

where $S_0^2 := E[(\Delta Y - g_{0s})^2 \mid D=0] E[\alpha_{0s}^2 \mid D=0]$.

Now, we show that $C^2_{0\Delta Y} = R^2_{\Delta Y - g_{0s} \sim g_0 - g_{0s} | D=0}$:

$$\begin{aligned}
R^2_{\Delta Y - g_{0s} \sim g_0 - g_{0s} | D=0} &:= 1 - \frac{E[(\Delta Y - g_0)^2 \mid D=0]}{E[(\Delta Y - g_{0s})^2 \mid D=0]} \\
&= \frac{E[(\Delta Y - g_0 + g_0 - g_{0s})^2 \mid D=0] - E[(\Delta Y - g_0)^2 \mid D=0]}{E[(\Delta Y - g_{0s})^2 \mid D=0]} \\
&= \frac{E[(g_0 - g_{0s})^2 \mid D=0] + 2E[(\Delta Y - g_0)(g_0 - g_{0s}) \mid D=0]}{E[(\Delta Y - g_{0s})^2 \mid D=0]} \\
&\overset{\text{LTE}}{=} \frac{E[(g_0 - g_{0s})^2 \mid D=0] + 2E[(g_0 - g_0)(g_0 - g_{0s}) \mid D=0]}{E[(\Delta Y - g_{0s})^2 \mid D=0]} \\
&= C^2_{0\Delta Y}.
\end{aligned}$$

Next, we relate $C^2_{0\Delta Y}$ to the nonparametric partial R^2 measure,

$$\begin{aligned}
C^2_{0\Delta Y} &:= \frac{E[(g_0 - g_{0s})^2 \mid D=0]}{E[(\Delta Y - g_{0s})^2 \mid D=0]} \\
&\overset{\text{LTE}}{=} \frac{E[g_0^2 \mid D=0] - E[g_{0s}^2 \mid D=0]}{E[\Delta Y^2 \mid D=0] - E[g_{0s}^2 \mid D=0]} \\
&= \frac{\frac{E[g_0^2|D=0]}{E[\Delta Y^2|D=0]} - \frac{E[g_{0s}^2|D=0]}{E[\Delta Y^2|D=0]}}{1 - \frac{E[g_{0s}^2|D=0]}{E[\Delta Y^2|D=0]}} \\
&\overset{\text{def.}}{=} \eta^2_{\Delta Y \sim U | X, D=0},
\end{aligned}$$

Note that,

$$\begin{aligned}
R^2_{\alpha_0 \sim \alpha_{0s} | D=0} &:= 1 - \frac{E[(\alpha_0 - \alpha_{0s})^2 \mid D=0]}{E[\alpha_0^2 \mid D=0]} \\
&= \frac{E[\alpha_0^2 \mid D=0] - (E[\alpha_0^2 \mid D=0] + E[\alpha_{0s}^2 \mid D=0] - 2E[\alpha_0 \alpha_{0s} \mid D=0])}{E[\alpha_0^2 \mid D=0]} \\
&\overset{\text{LTE}}{=} \frac{E[\alpha_0^2 \mid D=0] - (E[\alpha_0^2 \mid D=0] + E[\alpha_{0s}^2 \mid D=0] - 2E[\alpha_{0s}^2 \mid D=0])}{E[\alpha_0^2 \mid D=0]} \\
&= \frac{E[\alpha_{0s}^2 \mid D=0]}{E[\alpha_0^2 \mid D=0]}.
\end{aligned}$$

Additionally,

$$\begin{aligned}
R^2_{O_{XU} \sim O_X | D=0} &:= 1 - \frac{E[(O_{XU} - O_X)^2 \mid D=0]}{E[O_{XU}^2 \mid D=0]} \\
&= \frac{E[O_{XU}^2 \mid D=0] - \big(E[O_{XU}^2 \mid D=0] + E[O_X^2 \mid D=0] - 2E[O_{XU} O_X \mid D=0]\big)}{E[O_{XU}^2 \mid D=0]} \\
&\overset{\text{LTE}}{=} \frac{E[O_{XU}^2 \mid D=0] - \big(E[O_{XU}^2 \mid D=0] + E[O_X^2 \mid D=0] - 2E[O_X^2 \mid D=0]\big)}{E[O_{XU}^2 \mid D=0]} \\
&= \frac{E[O_X^2 \mid D=0]}{E[O_{XU}^2 \mid D=0]} = \frac{E[\alpha_{0s}^2 \mid D=0]}{E[\alpha_0^2 \mid D=0]} \\
&= R^2_{\alpha_0 \sim \alpha_{0s} | D=0}.
\end{aligned}$$

Accordingly, we express C_{0D}^2 in terms of the nonparametric R^2 measure,

$$\begin{aligned}
C_{0D}^2 &= \frac{E[(\alpha_0 - \alpha_{0s})^2 \mid D=0]}{E[\alpha_{0s}^2 \mid D=0]} = \frac{E[\alpha_0^2 \mid D=0] + E[\alpha_{0s}^2 \mid D=0] - 2E[\alpha_0\alpha_{0s} \mid D=0]}{E[\alpha_{0s}^2 \mid D=0]} \\
&\overset{\text{LTE}}{=} \frac{E[\alpha_0^2 \mid D=0] - E[\alpha_{0s}^2 \mid D=0]}{E[\alpha_{0s}^2 \mid D=0]} \\
&= \frac{1 - R^2_{\alpha_0 \sim \alpha_{0s} \mid D=0}}{R^2_{\alpha_0 \sim \alpha_{0s} \mid D=0}} \\
&= \frac{1 - R^2_{O_{XU} \sim O_X \mid D=0}}{R^2_{O_{XU} \sim O_X \mid D=0}}.
\end{aligned}$$

Finally, we simplify the scaling factor S_0^2 as follows:

$$\begin{aligned}
\sigma_{0s}^2 &:= E[(\Delta Y - g_{0s})^2 \mid D=0] = \text{Var}(\Delta Y - g_{0s} \mid D=0) \\
&\overset{\text{LTV}}{=} \text{Var}\left(E\left[\Delta Y - g_{0s} \mid X, D=0\right] \mid D=0\right) + E\left[\text{Var}\left(\Delta Y - g_{0s} \mid X, D=0\right) \mid D=0\right] \\
&= E[\text{Var}(\Delta Y \mid X, D=0) \mid D=0], \text{ and} \\
\nu_{0s}^2 &:= E[\alpha_{0s}^2 \mid D=0] \overset{\text{def.}}{=} E\left[\left(\frac{O_X}{O}\right)^2 \middle| D=0\right], \\
\implies S_0^2 &:= \sigma_{0s}^2 \nu_{0s}^2 = E[\text{Var}(\Delta Y \mid X, D=0) \mid D=0] \times E\left[\left(\frac{O_X}{O}\right)^2 \middle| D=0\right].
\end{aligned}$$

□

Proof of Lemma 2. Parts 1 and *2* are immediate consequences of Proposition 2, items (1) and (3) respectively. *Part 3* follows directly from Proposition 3(2). □

Proof of Corollary 1.

Part 1. By Proposition 2(3), we transform $R^2_{O_{XU} \sim O_X \mid D=0}$ into an expression in terms of the average selection odds among the treated.

$$\begin{aligned}
R^2_{O_{XU} \sim O_X \mid D=0} &= \frac{E[O_X^2 \mid D=0]}{E[O_{XU}^2 \mid D=0]} = \frac{E[O_X \mid D=1]}{E[O_{XU} \mid D=1]}, \\
\implies C_{0D}^2 &= \frac{1 - R^2_{O_{XU} \sim O_X \mid D=0}}{R^2_{O_{XU} \sim O_X \mid D=0}} = \frac{E[O_{XU} \mid D=1] - E[O_X \mid D=1]}{E[O_X \mid D=1]}.
\end{aligned}$$

Correspondingly, we have

$$1 - R^2_{O_{XU} \sim O_X \mid D=0} = 1 - \frac{E[O_X \mid D=1]}{E[O_{XU} \mid D=1]} = \frac{E[O_{XU} \mid D=1] - E[O_X \mid D=1]}{E[O_{XU} \mid D=1]},$$

and

$$\nu_{0s}^2 = E\left[\left(\frac{O_X}{O}\right)^2 \middle| D=0\right] = E\left[\frac{O_X}{O} \middle| D=1\right].$$

Part 2. By Proposition 2(2) and Proposition 3(1), we have that,

$$\chi^2(P_{X|1} \| P_{X|0}) = E\left[\left(\frac{O_X}{O}\right)^2 \middle| D=0\right] - 1, \text{ and } \chi^2(P_{X,U|1} \| P_{X,U|0}) = E\left[\left(\frac{O_{XU}}{O}\right)^2 \middle| D=0\right] - 1.$$

Therefore,

$$\begin{aligned}
R^2_{O_{XU}\sim O_X|D=0} &= \frac{E[O_X^2 \mid D=0]}{E[O_{XU}^2 \mid D=0]} \\
&= \frac{(\chi^2(P_{X|1}\|P_{X|0})+1)\times O^2}{(\chi^2(P_{X,U|1}\|P_{X,U|0})+1)\times O^2} \\
&= \frac{\chi^2(P_{X|1}\|P_{X|0})+1}{\chi^2(P_{X,U|1}\|P_{X,U|0})+1}, \\
\implies C^2_{0D} &= \frac{1-R^2_{O_{XU}\sim O_X|D=0}}{R^2_{O_{XU}\sim O_X|D=0}} = \frac{1-\frac{\chi^2(P_{X|1}\|P_{X|0})+1}{\chi^2(P_{X,U|1}\|P_{X,U|0})+1}}{\frac{\chi^2(P_{X|1}\|P_{X|0})+1}{\chi^2(P_{X,U|1}\|P_{X,U|0})+1}} \\
&= \frac{\chi^2(P_{X,U|1}\|P_{X,U|0})-\chi^2(P_{X|1}\|P_{X|0})}{\chi^2(P_{X|1}\|P_{X|0})+1}.
\end{aligned}$$

Correspondingly, we have

$$1-R^2_{O_{XU}\sim O_X|D=0} = 1-\frac{\chi^2(P_{X|1}\|P_{X|0})+1}{\chi^2(P_{X,U|1}\|P_{X,U|0})+1} = \frac{\chi^2(P_{X,U|1}\|P_{X,U|0})-\chi^2(P_{X|1}\|P_{X|0})}{\chi^2(P_{X,U|1}\|P_{X,U|0})+1},$$

and

$$\nu^2_{0s} := E\left[\left(\frac{O_X}{O}\right)^2 \middle| D=0\right] = \chi^2(P_{X|1}\|P_{X|0})+1.$$

□

Proof of Lemma 3 and Theorem 2.

Lemma 3 is a direct application of Theorems 3.1 and 3.2 in Chernozhukov et al. (2018). Valid estimation of covariance follows similarly to the proof of Theorem 3.2 in Chernozhukov et al. (2018). The first claim of Theorem 2 then follows by the delta method (Van Der Vaart and Wellner, 1996), and the stated confidence intervals are justified by standard arguments on asymptotic normality.

Below, we verify that the scores in (7)–(9) are Neyman orthogonal, as required by Theorem 3.1 of Chernozhukov et al. (2018).

- **Score for θ_s.**

  Moment condition.

$$\begin{aligned}
&E[\psi_{\theta_s}(Z; g_{0s}, \pi, p)] \\
&= E[\Delta Y - g_{0s} \mid D=1] - E[(\Delta Y - g_{0s})\alpha_{0s} \mid D=0] - \theta_s \\
&\overset{\text{LTE}}{=} \underbrace{E[\Delta Y - g_{0s} \mid D=1]}_{\theta_s} - \theta_s \\
&= 0,
\end{aligned}$$

Orthogonality with respect to g_{0s}.

$$
\begin{aligned}
&\partial_r \left\{ E[\psi_{\theta_s}(Z; g_{0s} + r \underbrace{(g'_{0s} - g_{0s})}_{h}, \pi, p)] \right\} \Bigg|_{r=0} \\
&= \partial_r \left\{ E\left[\frac{D(\Delta Y - g_{0s} - rh)}{p} - \frac{\pi(1-D)(\Delta Y - g_{0s} - rh)}{p(1-\pi)} \right] \right\} \Bigg|_{r=0} \\
&= E\left[-\frac{Dh}{p} + \frac{\pi(1-D)h}{p(1-\pi)} \right] \Bigg|_{r=0} \\
&= \frac{1}{p} E\left[\frac{\pi - D}{1-\pi} (g'_{0s} - g_{0s}) \right] \\
&\overset{\text{LTE}}{=} \frac{1}{p} E\left[E\left[\frac{\pi - D}{1-\pi} (g'_{0s} - g_{0s}) \mid X \right] \right] \\
&= \frac{1}{p} E\left[E\left[\frac{\pi - \pi}{1-\pi} \right] (g'_{0s}(X) - g_{0s}(X)) \right] = 0,
\end{aligned}
$$

Orthogonality with respect to π.

$$
\begin{aligned}
&\partial_r \left\{ E[\psi_{\theta_s}(Z; \pi + r \underbrace{(\pi' - \pi)}_{h}, g_{0s}, p)] \right\} \Bigg|_{r=0} \\
&= \partial_r \left\{ E\left[-\frac{(\pi + rh)(1-D)(\Delta Y - g_{0s})}{p(1-\pi-rh)} \right] \right\} \Bigg|_{r=0} \\
&= E\left[-\frac{(1-D)(\Delta Y - g_{0s})h}{p(1-\pi)^2} \right] \\
&\overset{\text{LTE}}{=} E\left[-(1-D) E\left[\frac{(\Delta Y - g_{0s})h}{p(1-\pi)^2} \mid D \right] \right] \\
&= -E\left[\frac{(\Delta Y - g_{0s})h}{p(1-\pi)^2} \mid D = 0 \right] \times (1-p) \\
&\overset{\text{LTE}}{=} -E\left[\frac{h}{p(1-\pi)^2} E[\Delta Y - g_{0s} \mid X, D = 0] \mid D = 0 \right] \times (1-p) \\
&= 0,
\end{aligned}
$$

Orthogonality with respect to p.

$$
\begin{aligned}
&\partial_p \left\{ E[\psi_{\theta_s}(Z; \theta_s, g_{0s}, \pi, p)] \right\} \\
&= E\left[-\frac{D(\Delta Y - g_{0s})}{p^2} + \frac{\pi(1-D)(\Delta Y - g_{0s})}{p^2(1-\pi)} + \frac{D\theta_s}{p^2} \right] \\
&= \frac{1}{p^2} E\left[\frac{(\pi - D)(\Delta Y - g_{0s})}{1-\pi} + D\theta_s \right] \\
&\overset{\text{LTE}}{=} \frac{1}{p^2} \left(E\left[-(\Delta Y - g_{0s}) + \theta_s \mid D = 1 \right] p + E\left[\frac{\pi(\Delta Y - g_{0s})}{1-\pi} \mid D = 0 \right] \right) \\
&\overset{\text{LTE}}{=} 0 + 0 = 0.
\end{aligned}
$$

- **Score for σ_{0s}^2.**

  Moment condition.

$$E[\psi_{\sigma_{0s}^2}(Z; g_{0s}, p)] \overset{\text{def.}}{=} \sigma_{0s}^2 - \sigma_{0s}^2 = 0,$$

  Orthogonality with respect to g_{0s}.

$$\begin{aligned}
&\partial_r \left\{ E[\psi_{\sigma_{0s}^2}(Z; g_{0s} + r(g'_{0s} - g_{0s}), p)] \right\}\Big|_{r=0} \\
&= \partial_r \left\{ E\left[\frac{1-D}{1-p}(\Delta Y - g_{0s} - r(g'_{0s} - g_{0s}))^2 - \frac{1-D}{1-p}\sigma_{0s}^2 \right] \right\}\Big|_{r=0} \\
&= -2E\left[(\Delta Y - g_{0s} - r(g'_{0s} - g_{0s}))(g'_{0s} - g_{0s}) \mid D = 0 \right]\Big|_{r=0} \\
&= -2E\left[(\Delta Y - g_{0s})(g'_{0s} - g_{0s}) \mid D = 0 \right] \\
&\overset{\text{LTE}}{=} -2E[(g'_{0s} - g_{0s})E[(\Delta Y - g_{0s}) \mid D = 0, X] \mid D = 0] = 0,
\end{aligned}$$

  Orthogonality with respect to p.

$$\begin{aligned}
&\partial_p E[\psi_{\sigma_{0s}^2}(Z; g_{0s}, p)] \\
&= \frac{1}{1-p}\left(E[(\Delta Y - g_{0s})^2 \mid D = 0] - \sigma_{0s}^2 \right) \overset{\text{def.}}{=} 0.
\end{aligned}$$

- **Score for ν_{0s}^2.**

  Moment condition.

$$\begin{aligned}
&E[\psi_{\nu_{0s}^2}(Z; \pi, p)] \\
&= E\left[2\frac{D}{p}\left(\frac{O_X}{O} - \nu_{0s}^2 \right) - \frac{1-D}{1-p} \times \left[\left(\frac{O_X}{O} \right)^2 - \nu_{0s}^2 \right] \right] \\
&= 2E\left[\alpha_{0s} \mid D = 1 \right] - 2\nu_{0s}^2 - E[\alpha_{0s}^2 \mid D = 0] + \nu_{0s}^2 \\
&\overset{\text{def.}}{=} 0,
\end{aligned}$$

  Orthogonality with respect to π.

$$\begin{aligned}
&\partial_r \left\{ E[\psi_{\nu_{0s}^2}(Z; \pi + r(\pi' - \pi), p)] \right\}\Big|_{r=0} \\
&= \partial_r \left\{ E\left[2\frac{D}{p}\frac{\pi + rh}{1 - \pi - rh}\frac{1-p}{p} - \frac{1-D}{1-p}\frac{(\pi + rh)^2}{(1 - \pi - rh)^2}\frac{(1-p)^2}{p^2} \right] \right\}\Big|_{r=0} \\
&= E\left[\frac{2D(1-p)}{p^2}\frac{h}{(1 - \pi - rh)^2} - \frac{(1-D)(1-p)}{p^2}\frac{2h(\pi + rh)}{(1 - \pi - rh)^3} \right]\Big|_{r=0} \\
&= E\left[\frac{2D(1-p)}{p^2}\frac{h}{(1-\pi)^2} - \frac{(1-D)(1-p)}{p^2}\frac{2h\pi}{(1-\pi)^3} \right] \\
&= \frac{2(1-p)}{p^2} \times E\left[\frac{h(D - \pi)}{(1-\pi)^3} \right]
\end{aligned}$$

$$\overset{\text{LTE}}{=} \frac{2(1-p)}{p^2} \times E\left[h \times E\left[\frac{D-\pi}{(1-\pi)^3} \mid X\right]\right]$$
$$= \frac{2(1-p)}{p^2} \times E\left[h \times \frac{\pi-\pi}{(1-\pi)^3}\right] = 0,$$

Orthogonality with respect to p.

$$\begin{aligned}
&\partial_p E[\psi_{\nu_{0s}^2}(Z;\pi,p)] \\
&= E\left[\frac{-2D}{p^2}\alpha_{0s} - \frac{2D}{p^3}\frac{\pi}{1-\pi}\right] \\
&\qquad - E\left[\frac{1-D}{(1-p)^2}\alpha_{0s}^2 - \frac{1-D}{1-p}\frac{\pi^2}{(1-\pi)^2}\frac{2(1-p)}{p^3}\right] \\
&\qquad - E\left[\frac{-2D}{p^2} - \frac{1-D}{(1-p)^2}\right]\nu_{0s}^2 \\
&= \frac{-2}{p}\nu_{0s}^2 - \frac{1}{1-p}\nu_{0s}^2 + \frac{2}{p}\nu_{0s}^2 + \frac{1}{1-p}\nu_{0s}^2 \\
&\qquad - E\left[\frac{2D}{p^3}\frac{\pi}{1-\pi}\right] + E\left[\frac{\pi^2}{(1-\pi)^2}\frac{2(1-D)}{p^3}\right] \\
&= E\left[2\frac{\pi(\pi-D)}{p^3(1-\pi)^2}\right] \overset{\text{LTE}}{=} 0.
\end{aligned}$$

□

C Benchmarking analysis

C.1 Benchmarking against observed covariates

Our approach follows Cinelli and Hazlett (2020, 2025a) and Chernozhukov et al. (2026), extrapolating the strength of unobserved confounding from its relative strength compared to observed covariates. We begin by introducing notation and then define the benchmarking measures for each bias factor.

C.1.1 Notation

Suppose we use X_j, a subvector of the observed covariates X, as the benchmark covariate vector. Let X_{-j} denote the vector of observed covariates other than X_j. Correspondingly, we define $g_{0s,-j} := E[\Delta Y \mid X_{-j}, D=0]$, $\alpha_{0s,-j} = O_{X_{-j}}/O$, $\theta_{0s,-j} := E[g_{0s,-j} \mid D=1]$, and $\theta_{s,-j} := \theta_1 - \theta_{0s,-j}$. We define the bias factors that quantify the strength of X_j conditional on X_{-j} as follows:

$$\begin{aligned}
C_{0\Delta Y,j}^2 &:= \eta_{\Delta Y \sim X_j | X_{-j}, D=0}^2, \\
C_{0D,j}^2 &:= \frac{1 - R_{O_X \sim O_{X_{-j}} | D=0}^2}{R_{O_X \sim O_{X_{-j}} | D=0}^2} \\
&= \frac{E[O_X \mid D=1] - E[O_{X_{-j}} \mid D=1]}{E[O_{X_{-j}} \mid D=1]}
\end{aligned}$$

$$= \frac{\chi^2(P_{X|1}\|P_{X|0}) - \chi^2(P_{X_{-j}|1}\|P_{X_{-j}|0})}{\chi^2(P_{X_{-j}|1}\|P_{X_{-j}|0}) + 1}, \text{ and}$$

$$\rho_{0,j} := \text{Cor}(g_{0s} - g_{0s,-j}, O_X - O_{X_{-j}} \mid D = 0).$$

The corresponding scaling factors are $\sigma^2_{0s,-j} := E[\text{Var}(\Delta Y \mid X_{-j}, D = 0) \mid D = 0]$ and $\nu^2_{0s,-j} := E[\alpha^2_{0s,-j} \mid D = 0] = E[\alpha_{0s,-j} \mid D = 1]$.

C.1.2 Relative bounds on $1 - R^2_{O_{XU} \sim O_X | D=0}$

We now discuss how to express the strength of unobserved confounders in explaining treatment assignment in terms of their relative strength as compared to X_j.

Definition 1 (Relative strength with treatment selection)**.**

$$k_{0D,j} := \frac{R^2_{O_X \sim O_{X_{-j}}|D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|D=0}}{1 - R^2_{O_X \sim O_{X_{-j}}|D=0}} \qquad \text{(Residual } R^2 \text{ in Selection Odds)}$$

$$= \frac{\frac{E[O_{XU}|D=1] - E[O_{X_{-j}}|D=1]}{E[O_{XU}|D=1]} - \frac{E[O_X|D=1] - E[O_{X_{-j}}|D=1]}{E[O_X|D=1]}}{\frac{E[O_X|D=1] - E[O_{X_{-j}}|D=1]}{E[O_X|D=1]}} \qquad \text{(Average Selection Odds)}$$

$$= \frac{\frac{\chi^2(P_{X,U|1}\|P_{X,U|0}) - \chi^2(P_{X_{-j}|1}\|P_{X_{-j}|0})}{\chi^2(P_{X,U|1}\|P_{X,U|0}) + 1} - \frac{\chi^2(P_{X|1}\|P_{X|0}) - \chi^2(P_{X_{-j}|1}\|P_{X_{-j}|0})}{\chi^2(P_{X|1}\|P_{X|0}) + 1}}{\frac{\chi^2(P_{X|1}\|P_{X|0}) - \chi^2(P_{X_{-j}|1}\|P_{X_{-j}|0})}{\chi^2(P_{X|1}\|P_{X|0}) + 1}}. \qquad \text{(Distributional Imbalance)}$$

Using this definition, we obtain the following reparameterization of the sensitivity parameter $1 - R^2_{O_{XU} \sim O_X | D=0}$.

Proposition 5 (Reparameterization of $1 - R^2_{O_{XU} \sim O_X | D=0}$ in terms of relative strength)**.**

$$1 - R^2_{O_{XU} \sim O_X | D=0} = k_{0D,j} \times C^2_{0D,j}, \text{ and } C^2_{0D} = \frac{1 - R^2_{O_{XU} \sim O_X | D=0}}{R^2_{O_{XU} \sim O_X | D=0}},$$

where $C^2_{0D,j} := \frac{1 - R^2_{O_X \sim O_{X_{-j}}|D=0}}{R^2_{O_X \sim O_{X_{-j}}|D=0}}$.

Proof.

$$k_{0D,j} \times C^2_{0D,j} = \frac{R^2_{O_X \sim O_{X_{-j}}|D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|D=0}}{1 - R^2_{O_X \sim O_{X_{-j}}|D=0}} \times \frac{1 - R^2_{O_X \sim O_{X_{-j}}|D=0}}{R^2_{O_X \sim O_{X_{-j}}|D=0}}$$

$$= 1 - \frac{R^2_{O_{XU} \sim O_{X_{-j}}|D=0}}{R^2_{O_X \sim O_{X_{-j}}|D=0}} = 1 - R^2_{O_{XU} \sim O_X | D=0}. \qquad \square$$

Remark 4. *Note that, since* $1 - R^2_{O_{XU}} \leq 1$*, we must have* $k_{0D,j} \leq 1/C^2_{0D,j}$*. Also,* $R^2_{O_X \sim O_{X_{-j}}|D=0} = \frac{E\left[O^2_{X_{-j}}|D=0\right]}{E\left[O^2_X|D=0\right]} = \frac{\nu^2_{0s,-j}}{\nu^2_{0s}}$*, so an alternative expression is*

$$1 - R^2_{O_{XU} \sim O_X | D=0} = k_{0D,j} \times \frac{\nu^2_{0s} - \nu^2_{0s,-j}}{\nu^2_{0s,-j}}.$$

This latter form is useful for estimation and inference.

C.1.3 Relative bounds on $\eta^2_{\Delta Y \sim U|X,D=0}$

We now discuss how to express the strength of unobserved confounders in explaining outcome evolution in terms of their relative strength as compared to X_j.

Definition 2 (Relative strength with outcome trend)**.**

$$k_{0\Delta Y,j} := \frac{\eta^2_{\Delta Y \sim U, X_j|X_{-j},D=0} - \eta^2_{\Delta Y \sim X_j|X_{-j},D=0}}{\eta^2_{\Delta Y \sim X_j|X_{-j},D=0}}.$$

Using this definition, we obtain the following reparameterization of the sensitivity parameter $\eta^2_{\Delta Y \sim U|X,D=0}$.

Proposition 6 (Reparameterization of $\eta^2_{\Delta Y \sim U|X,D=0}$ in terms of relative strength)**.**

$$\eta^2_{\Delta Y \sim U|X,D=0} = k_{0\Delta Y,j} \times \frac{C^2_{0\Delta Y,j}}{1 - C^2_{0\Delta Y,j}},$$

where $C^2_{0\Delta Y,j} := \eta^2_{\Delta Y \sim X_j|X_{-j},D=0}$

Proof.

$$\begin{aligned}
k_{0\Delta Y,j} \times \frac{C^2_{0\Delta Y,j}}{1 - C^2_{0\Delta Y,j}} &= \frac{\eta^2_{\Delta Y \sim U, X_j|X_{-j},D=0} - \eta^2_{\Delta Y \sim X_j|X_{-j},D=0}}{\eta^2_{\Delta Y \sim X_j|X_{-j},D=0}} \times \frac{\eta^2_{\Delta Y \sim X_j|X_{-j},D=0}}{1 - \eta^2_{\Delta Y \sim X_j|X_{-j},D=0}} \\
&= \frac{\eta^2_{\Delta Y \sim U, X_j|X_{-j},D=0} - \eta^2_{\Delta Y \sim X_j|X_{-j},D=0}}{1 - \eta^2_{\Delta Y \sim X_j|X_{-j},D=0}} \\
&= \eta^2_{\Delta Y \sim U|X,D=0}.
\end{aligned}$$

□

Remark 5. *Note that, since* $\eta^2_{\Delta Y \sim U|X,D=0} \leq 1$*, we must have* $k_{0\Delta Y,j} \leq \frac{1 - C^2_{0\Delta Y,j}}{C^2_{0\Delta Y,j}}$*. Also note that*

$$\begin{aligned}
\sigma^2_{0s,-j} - \sigma^2_{0s} &= E[(\Delta Y - g_{0s,-j})^2 \mid D = 0] - E[(\Delta Y - g_{0s})^2 \mid D = 0] \\
&\overset{LTE}{=} E[(g_{0s} - g_{0s,-j})^2 \mid D = 0],
\end{aligned}$$

which leads to the following alternative representation

$$\begin{aligned}
\eta^2_{\Delta Y \sim U|X,D=0} &= k_{0\Delta Y,j} \times \frac{\eta^2_{\Delta Y \sim X_j|X_{-j},D=0}}{1 - \eta^2_{\Delta Y \sim X_j|X_{-j},D=0}} \\
&= k_{0\Delta Y,j} \times \frac{E[(g_{0s} - g_{0s,-j})^2 \mid D = 0]}{E[(\Delta Y - g_{0s,-j})^2 \mid D = 0] - E[(g_{0s} - g_{0s,-j})^2 \mid D = 0]} \\
&\overset{LTE}{=} k_{0\Delta Y,j} \times \frac{E[(g_{0s} - g_{0s,-j})^2 \mid D = 0]}{E[(\Delta Y - g_{0s})^2 \mid D = 0]} \\
&= k_{0\Delta Y,j} \times \frac{\sigma^2_{0s,-j} - \sigma^2_{0s}}{\sigma^2_{0s}}.
\end{aligned}$$

This latter form is useful for estimation and inference.

C.1.4 Benchmarking $|\rho_0|$

Following Chernozhukov et al. (2026), we propose extrapolating ρ_0 using the observed alignment of X_j as given by $\rho_{0,j} := \text{Cor}(g_{0s} - g_{0s,-j}, O_X - O_{X_{-j}} \mid D = 0)$. Note that the bias decomposition with respect to X_j gives,

$$\theta_s - \theta_{s,-j} = -\rho_{0,j}\sqrt{(\sigma^2_{0s,-j} - \sigma^2_{0s})(\nu^2_{0s} - \nu^2_{0s,-j})}$$

This leads to the following debiased representation of $\rho_{0,j}$,

$$\rho_{0,j} = \frac{-(\theta_s - \theta_{s,-j})}{\sqrt{(\sigma^2_{0s,-j} - \sigma^2_{0s})(\nu^2_{0s} - \nu^2_{0s,-j})}}.$$

This latter form is useful for estimation and inference, which we discuss below.

C.1.5 Statistical inference for benchmark components

Statistical inference for our benchmarking analysis follows the procedure introduced in Appendices E.5 and E.6 of Chernozhukov et al. (2026). Specifically, we define the estimable "gain" metrics with the debiased representations derived above,

$$\begin{aligned}
G_{0D,j} &:= C^2_{0D,j} = \frac{\nu^2_{0s} - \nu^2_{0s,-j}}{\nu^2_{0s,-j}}, \\
G_{0\Delta Y,j} &:= \frac{C^2_{0\Delta Y,j}}{1 - C^2_{0\Delta Y,j}} = \frac{\sigma^2_{0s,-j} - \sigma^2_{0s}}{\sigma^2_{0s}}, \text{ and} \\
\rho_{0,j} &= \frac{-(\theta_s - \theta_{s,-j})}{\sqrt{(\sigma^2_{0s,-j} - \sigma^2_{0s}) \times (\nu^2_{0s} - \nu^2_{0s,-j})}}.
\end{aligned}$$

We have shown that $1 - R^2_{O_{XU} \sim O_X | D=0} = k_{0D,j} \times G_{0D,j}$ and $\eta^2_{\Delta Y \sim U | X, D=0} = k_{0\Delta Y,j} \times G_{0\Delta Y,j}$. Therefore, given the influence functions of θ_s, $\theta_{s,-j}$, σ^2_{0s}, $\sigma^2_{0s,-j}$, ν^2_{0s}, and $\nu^2_{0s,-j}$ obtained from DML, we apply the delta method to obtain the following debiased influence functions for the "gain" metrics.

$$\begin{aligned}
\varphi^0_{G_{0D,j}}(Z) &= \frac{\nu^2_{0s,-j}\varphi^0_{\nu^2_{0s}}(Z) - \nu^2_{0s}\varphi^0_{\nu^2_{0s,-j}}(Z)}{\nu^4_{0s,-j}}, \\
\varphi^0_{G_{0\Delta Y,j}}(Z) &= \frac{\sigma^2_{0s}\varphi^0_{\sigma^2_{0s,-j}}(Z) - \sigma^2_{0s,-j}\varphi^0_{\sigma^2_{0s}}(Z)}{\sigma^4_{0s}}, \text{ and} \\
\varphi^0_{\rho_{0,j}}(Z) &= \frac{\varphi^0_{\theta_{s,-j}}(Z) - \varphi^0_{\theta_s}(Z)}{(\sigma^2_{0s,-j} - \sigma^2_{0s})^{1/2}(\nu^2_{0s} - \nu^2_{0s,-j})^{1/2}} - \frac{(\theta_{s,-j} - \theta_s)(\varphi^0_{\sigma^2_{0s,-j}}(Z) - \varphi^0_{\sigma^2_{0s}}(Z))}{2(\sigma^2_{0s,-j} - \sigma^2_{0s})^{3/2}(\nu^2_{0s} - \nu^2_{0s,-j})^{1/2}} \\
&\quad - \frac{(\theta_{s,-j} - \theta_s)(\varphi^0_{\nu^2_{0s}}(Z) - \varphi^0_{\nu^2_{0s,-j}}(Z))}{2(\sigma^2_{0s,-j} - \sigma^2_{0s})^{1/2}(\nu^2_{0s} - \nu^2_{0s,-j})^{3/2}}.
\end{aligned}$$

The plug-in estimator of the bias and the corresponding influence function for $\theta_\pm$ follow directly from the discussion in Appendix E.6 of Chernozhukov et al. (2026), after replacing the "unconditional"

components with their corresponding "conditional" counterparts in our results.

C.2 Benchmarking against pre-trends

Here we let $Z := (\Delta Y, D, X, \Delta Y^{\text{pre}}, D^{\text{pre}}, X^{\text{pre}})$. In this section we provide the deferred influence functions for the pre-trend extrapolation bounds. The first approach yields the following bounds on the target estimand in the post-treatment period:

$$\theta_{\pm} = \theta_s \pm k \cdot |\theta_s^{\text{pre}}|,$$

where θ_s and θ_s^{pre} are estimable from the data, and $\theta_s^{\text{pre}} \neq 0$. Under standard DML conditions, by applying the delta method, the plug-in estimator $\widehat{\theta}_{\pm}$ is asymptotically linear and Gaussian with influence function

$$\varphi^0_{\theta_{\pm}}(Z) := \varphi^0_{\theta_s}(Z) \pm k \cdot \text{sign}(\theta_s^{\text{pre}}) \cdot \varphi^0_{\theta_s^{\text{pre}}}(Z).$$

As for the second approach, recall it yields the bounds:

$$\theta_{\pm} = \theta_s \pm k \cdot \left(\frac{S_0}{S_0^{\text{pre}}}\right) \cdot |\theta_s^{\text{pre}}|.$$

The influence function for S_0 is:

$$\varphi^0_{S_0}(Z) = \frac{1}{2S_0}\left(\sigma^2_{0s}\,\varphi^0_{\nu^2_{0s}}(Z) + \nu^2_{0s}\,\varphi^0_{\sigma^2_{0s}}(Z)\right),$$

with a similar influence function for S_0^{pre}. By the delta method, the plug-in estimator $\widehat{\theta}_{\pm}$ is then asymptotically linear and Gaussian with influence function

$$\varphi^0_{\theta_{\pm}}(Z) = \varphi^0_{\theta_s}(Z) \pm k\,|\theta_s^{\text{pre}}|\frac{S_0}{S_0^{\text{pre}}}\left(\frac{\varphi^0_{\theta_s^{\text{pre}}}(Z)}{\theta_s^{\text{pre}}} + \frac{\varphi^0_{S_0}(Z)}{S_0} - \frac{\varphi^0_{S_0^{\text{pre}}}(Z)}{S_0^{\text{pre}}}\right).$$

D Sensitivity statistics

D.1 Definitions

Here we provide a precise definition of the robustness value and extreme robustness value.

Definition 3 (Robustness Value (RV))**.** *For a fixed θ^*, the robustness value, $RV_{\theta^*,\alpha}(\theta)$, is the minimum strength of confounding needed for the null hypothesis $H_0 : \theta = \theta^*$ not to be rejected at the significance level α, when setting $|\rho_0| \leq 1$. Formally,*

$$RV_{\theta^*,\alpha}(\theta) := \inf\left\{RV : \theta^* \in CI^{max}_{1-\alpha,RV,RV}(\theta)\right\},$$

where $CI^{\max}_{1-\alpha,RV,RV}(\theta)$ is the widest confidence bound constructed from Theorem 2 under the restrictions that $\eta^2_{\Delta Y \sim U|X,D=0} \leq RV$ and $1 - R^2_{O_{XU} \sim O_X|D=0} \leq RV$. To simplify notation, we denote

$\eta^2_{\Delta Y\sim U|X,D=0}$ *and* $1-R^2_{O_{XU}\sim O_X|D=0}$ *by* η^2 *and* $1-R^2$ *respectively. Then, formally,*

$$CI^{\max}_{1-\alpha,RV,RV}(\theta)=[L_{RV},U_{RV}],$$

$$\textit{with } L_{RV}=\min_{\eta^2,R^2}\left(\widehat{\theta}_-(\eta^2,R^2)-\Phi^{-1}(1-\alpha)\sqrt{\frac{E[(\varphi^0_{\theta_-}(Z;\eta^2,R^2))^2]}{n}}\right),\ \textit{s.t. } \eta^2\leq RV, 1-R^2\leq RV,$$

$$U_{RV}=\max_{\eta^2,R^2}\left(\widehat{\theta}_+(\eta^2,R^2)+\Phi^{-1}(1-\alpha)\sqrt{\frac{E[(\varphi^0_{\theta_+}(Z;\eta^2,R^2))^2]}{n}}\right),\ \textit{s.t. } \eta^2\leq RV, 1-R^2\leq RV.$$

Definition 4 (Extreme Robustness Value (XRV))**.** *For a fixed* θ^**, the extreme robustness value,* $XRV_{\theta^*,\alpha}(\theta)$*, is the minimum strength of confounding needed on selection such that the null hypothesis* $H_0:\theta=\theta^*$ *is not rejected at the significance level* α*, given that the residual variation of the outcome evolution explained by the latent confounders is unrestricted (i.e.,* $\eta^2_{\Delta Y\sim U|X,D=0}\leq 1$*), when setting* $|\rho_0|\leq 1$*. Formally,*

$$XRV_{\theta^*,\alpha}(\theta):=\inf\left\{XRV:\theta^*\in CI^{\max}_{1-\alpha,1,XRV}(\theta)\right\},$$

where $CI^{\max}_{1-\alpha,1,XRV}(\theta)$ *is the widest confidence bound constructed from Theorem 2 under the restrictions that* $\eta^2_{\Delta Y\sim U|X,D=0}\leq 1$ *and* $1-R^2_{O_{XU}\sim O_X|D=0}\leq XRV$*. To simplify notation, we denote* $\eta^2_{\Delta Y\sim U|X,D=0}$ *and* $1-R^2_{O_{XU}\sim O_X|D=0}$ *by* η^2 *and* $1-R^2$ *respectively. Then, formally,*

$$CI^{\max}_{1-\alpha,1,XRV}(\theta)=[L_{XRV},U_{XRV}],$$

$$\textit{with } L_{XRV}=\min_{\eta^2,R^2}\left(\widehat{\theta}_-(\eta^2,R^2)-\Phi^{-1}(1-\alpha)\sqrt{\frac{E[(\varphi^0_{\theta_-}(Z;\eta^2,R^2))^2]}{n}}\right),\ \textit{s.t. } \eta^2\leq 1, 1-R^2\leq XRV,$$

$$U_{XRV}=\max_{\eta^2,R^2}\left(\widehat{\theta}_+(\eta^2,R^2)+\Phi^{-1}(1-\alpha)\sqrt{\frac{E[(\varphi^0_{\theta_+}(Z;\eta^2,R^2))^2]}{n}}\right),\ \textit{s.t. } \eta^2\leq 1, 1-R^2\leq XRV.$$

D.2 Interpretation in terms of increase in average odds or covariate imbalance

For $k\geq 0$, note that

$$1-R^2_{O_{XU}\sim O_X|D=0}=k\iff C^2_{0D}=\frac{k}{1-k}.$$

Then,

- If k is the value of $\text{XRV}_{0,\alpha}$: if unobserved confounders increase treatment odds or covariate imbalance by less than $\frac{k}{1-k}$ then such confounders are not capable of overturning the original results, at the significance level of α, *regardless* of how much variation such confounders explain of the outcome trend.

- If k is the value of $\text{RV}_{0,\alpha}(\theta)$: unobserved confounders explaining less than k of the residual variation in untreated trend and inducing less than $\frac{k}{1-k}$ increase in average odds or covariate imbalance cannot overturn the conclusions of the study, at the significance level of α.

E Simulations

This section discusses the use of Monte Carlo simulations to examine the finite-sample properties of the proposed OVB results in the canonical DiD setting. Our data simulation strategy allows us to derive explicit expressions for the bias factors, as well as to identify the correctly specified models for estimating the nuisance parameters. This enables us to attribute the source of bias solely to the omission of covariates, which leads to the violation of the conditional PTA.

E.1 Data generating process

The data are generated according to the following steps:

Step 1: Given $p \geq \epsilon$, let $D_i \overset{i.i.d.}{\sim} \text{Bernoulli}\ (p)$. Then, generate the covariates as follows,

$$(X_i, U_i) \mid D_i = d \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\mu}_d, \Sigma_d), \text{ with}$$
$$\boldsymbol{\mu}_d = (\mu_{dX}, \mu_{dU}) \text{ and } \Sigma_d = \text{diag}(\sigma^2_{dX}, \sigma^2_{dU}), \text{ for } d \in \{0, 1\}.$$

Step 2: For each unit i, calculate the potential outcomes at time period $t \in \{1, 2\}$ using

$$Y_{i,t}(0) = \alpha_i + \rho_t + X_i\beta_{xt} + U_i\beta_{ut} + \epsilon_{i,t},$$
$$Y_{i,t}(1) = Y_{i,t}(0) + \mathbb{1}\{t \geq 2\} \cdot \theta + (\nu_{i,t} - \epsilon_{i,t}),$$

where $\rho_t = \beta_{xt} = \beta_{ut} = t$, $\alpha_i \mid D_i = d \overset{i.i.d.}{\sim} \mathcal{N}(2d, 1)$, and $\epsilon_{i,t}, \nu_{i,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$.

Step 3: Calculate the observed outcomes using consistency, which leads to the following model for outcome evolution:

$$\Delta Y_i = (\Delta\rho + X_i\Delta\beta_x + U_i\Delta\beta_u) + \theta D_i + \Delta\tilde{\epsilon}_i,$$

where $\Delta\rho = \Delta\beta_x = \Delta\beta_u = 1$, and $\Delta\tilde{\epsilon}_i = D_i\Delta\nu_i + (1 - D_i)\Delta\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 2)$.

The dependence of covariates (X_i, U_i) on D_i reflects selection into treatment. The covariates represent pre-treatment characteristics and are fixed over time.

In this simulation setting, the propensity score model is correctly specified as a logistic regression with quadratic terms in the covariates for both the "long" and "short" estimations. From Step 3,

the outcome evolution model is also correctly specified as a linear regression for both estimations. The true bias factors admit the following closed-form expressions:

$$C_{0\Delta Y}^2 = \frac{\sigma_{0U}^2}{\sigma_{0U}^2 + 2}, \quad C_{0D}^2 = \chi^2(P_{U|1} \| P_{U|0}), \text{ and}$$

$$\rho_0^2 = \frac{(\mu_{1U} - \mu_{0U})^2}{\sigma_{0U}^2 \times (\chi^2(P_{X|1} \| P_{X|0}) + 1)\chi^2(P_{U|1} \| P_{U|0})},$$

where $\Delta\theta_s := \theta - \theta_s = -(\mu_{1U} - \mu_{0U})$, and

$$\chi^2(P_{U|1} \| P_{U|0}) = \frac{\sigma_{0U}^2}{\sigma_{1U}\sqrt{2\sigma_{0U}^2 - \sigma_{1U}^2}} \exp\left(\frac{(\mu_{1U} - \mu_{0U})^2}{2\sigma_{0U}^2 - \sigma_{1U}^2}\right) - 1,$$

$$\chi^2(P_{X|1} \| P_{X|0}) = \frac{\sigma_{0X}^2}{\sigma_{1X}\sqrt{2\sigma_{0X}^2 - \sigma_{1X}^2}} \exp\left(\frac{(\mu_{1X} - \mu_{0X})^2}{2\sigma_{0X}^2 - \sigma_{1X}^2}\right) - 1.$$

These divergences are finite if and only if $2\sigma_{0X}^2 > \sigma_{1X}^2$ and $2\sigma_{0U}^2 > \sigma_{1U}^2$. Intuitively, when the treated distribution has substantially heavier tails, π_{XU} approaches one in the tails, undermining overlap and causing the density ratio to grow explosively, leading to infinite divergence.

Remark 6. *By the Riesz-Frechet representation theorem, the linear functional $\theta_0(g_0)$ is continuous on $L^2(P_{(X,U)|D=0})$ if and only if there exists a unique representer $\alpha_0 \in \Gamma$ such that $\theta_0(g_0) = E[g_0\alpha_0 \mid D = 0]$. Lemma 1 shows that this representer is $\alpha_0 = O_{XU}/O$. Therefore, continuity of θ_0 requires that $E[\alpha_0^2 \mid D = 0] < \infty$. Moreover, as shown in Corollary 1, $\chi^2(P_{X,U|1} \| P_{X,U|0}) = E[\alpha_0^2 \mid D = 0] - 1$, so in our simulation setting, requiring $2\sigma_{0X}^2 > \sigma_{1X}^2$ and $2\sigma_{0U}^2 > \sigma_{1U}^2$ ensures $\chi^2(P_{X,U|1} \| P_{X,U|0}) < \infty$ and equivalently $E[\alpha_0^2 \mid D = 0] < \infty$. This condition is automatically satisfied under the usual strong overlap assumption (Assumption A.4), which enforces uniform boundedness of the propensity score and therefore implies square-integrability. Formally,*

$$\alpha_0 = \frac{O_{XU}}{O} = \frac{\pi_{XU}}{1 - \pi_{XU}} \times \frac{1-p}{p} \leq \left(\frac{1-\epsilon}{\epsilon}\right)^2 \text{ by Assumption A.4,}$$

$$\implies \quad \chi^2(P_{X,U|1} \| P_{X,U|0}) + 1 = E[\alpha_0^2 \mid D = 0] < \infty.$$

E.2 Results

We evaluate the performance of our proposed approach from the following perspectives:

1. **Coverage of Confidence Bounds.** We evaluate the coverage of the confidence bound for θ, after plugging in the true bias factors.

   We set $\theta = 2$, $\boldsymbol{\mu}_0 = (0.3, 0.3)$, $\boldsymbol{\mu}_1 = (0, 0)$, $(\sigma_{0X}^2, \sigma_{0U}^2) = (6, 6)$, and $(\sigma_{1X}^2, \sigma_{1U}^2) = (3, 3)$. We consider scenarios with $p \in \{0.2, 0.5, 0.8\}$ and sample sizes $n \in \{500, 2,000, 5,000\}$, at signifi-

cance levels $\alpha \in \{0.01, 0.05, 0.1\}$. The nuisance parameters are estimated using the correctly specified model with ten-fold cross-fitting.[7] We report results from $B = 5,000$ repetitions.

The results are visualized in Figure 4, which shows that empirical coverage closely tracks nominal levels, indicating well-calibrated confidence bounds with stable finite-sample performance.

2. **Sensitivity Statistics.** Following the same scenarios as in our first coverage experiment, we compare the empirical values of the sensitivity statistics with their population values. Following arguments similar to those in Cinelli and Hazlett (2020, 2025a), define $f_{\theta^*} := \frac{|\theta^* - \theta_s|}{|\rho_0| S_0}$. The analytical solutions for the population values of $\text{RV}_{\theta^*=0}$ and $\text{XRV}_{\theta^*=0}$ are given by

$$\text{RV}_{\theta^*=0}(\theta) = \frac{1}{2}\left(\sqrt{f_{\theta^*}^4 + 4f_{\theta^*}^2} - f_{\theta^*}^2\right), \text{ and } \text{XRV}_{\theta^*=0}(\theta) = \frac{f_{\theta^*}^2}{1 + f_{\theta^*}^2}.$$

Note that the sensitivity statistics we define are conservative since we set $|\rho_0| = 1$ rather than using the true population value of ρ_0.

Figure 5 visualizes the results. The red dashed lines correspond to the population value of the *conservative* sensitivity statistics obtained by setting $|\rho_0| = 1$, which aligns with our definition. The blue dashed lines represent the *exact* sensitivity statistics obtained using the true population value of $|\rho_0|$, which should by definition be less conservative and higher than the red line. The boxplots show the empirical sensitivity statistics as they would be computed in applied settings.

As expected, the empirical values of $\text{RV}_{\theta^*=0,\alpha=0}$ and $\text{XRV}_{\theta^*=0,\alpha=0}$ concentrate around their population values (as indicated by the red dashed lines), with tighter concentration as the sample size increases. Moreover, these conservative estimates lie below the exact sensitivity statistics (blue dashed lines), confirming that our estimates are conservative in practice.

We additionally assess the finite-sample performance at the significance level $\alpha = 0.05$, with empirical estimates computed via grid search. Figure 6 shows that the estimated sensitivity statistics generally lie below the true conservative values (red dashed lines). This behavior is expected, since the 95% confidence bounds are wider than point estimates and therefore require a smaller confounding strength to cover θ^*. As the sample size increases, confidence bounds shrink, narrowing this gap and causing the estimates to move closer to the conservative population values.

[7] To ensure overlap, we set $\sigma_{0\cdot}^2$ to be sufficiently larger than $\sigma_{1\cdot}^2$. We use more than the default five cross-fitting folds to stabilize estimation of ν_{0s}^2.

3. **Model Misspecification.** In the presence of model misspecification, we evaluate the performance of our proposed approach by following the pipeline of Chernozhukov et al. (2018, 2024) for first-stage machine learning estimation of nuisance parameters, which we refer to as the "nonparametric" model. Detailed information on the learners is summarized in Table 6. We then compare these results with those obtained under parametric nuisance estimation, namely linear regression for the outcome evolution model and logistic regression for the propensity score model, which we refer to as the "parametric" model hereafter.

   Specifically, we maintain the DGP described in Section D.1 but fit the nuisance models using a one-to-one transformation of the covariates, $(X^*, U^*) = (\exp(X/2), U)$, thereby inducing misspecification of the parametric model. The remaining design follows the first coverage experiment, under which the true bias factors remain unchanged.

   Figures 7 and 8 show that confidence bounds obtained from the parametric model tend to be wider and, consequently, exhibit over-coverage relative to those from the nonparametric model. In our simulation, we focus on the upper confidence bound, which corresponds to the direction of bias relevant for overturning the conclusion. Accordingly, Figures 9 and 10 show that the upper bounds from the nonparametric model tend to lie closer to the true effect than those from the parametric model. These findings are consistent with our expectations: under model misspecification, the nonparametric model better approximates the true nuisance functions, leading to less biased first-stage estimates and, consequently, tighter confidence bounds with coverage closer to the nominal level and upper bounds closer to the true effect.

   In addition, Figure 9 shows that in some repetitions the parametric model produces excessively wide confidence bounds. This occurs because its estimates of ν_{0s}^2 are highly sensitive to the estimated propensity scores, which can exhibit greater dispersion under model misspecification. By contrast, the nonparametric model is less sensitive to extreme propensity score values.

   Finally, Figure 11 visualizes the sensitivity statistics at the significance level $\alpha = 0.05$. Overall, the sensitivity statistics estimated from the nonparametric model are more tightly concentrated and are closer to the true values than those from the parametric model. This indicates improved finite-sample stability and accuracy of the nonparametric model under model misspecification.

E.3 Deferred derivations

We first verify correct specification of the propensity score model by deriving the expression for the propensity score and showing that it takes the form of a logistic regression with quadratic terms.

(a) **Short:**

$$\begin{aligned}\frac{P(X=x|D=0)}{P(X=x|D=1)} &= \frac{\prod_{i=1}^{q_X} \frac{1}{\sqrt{2\pi\sigma_{0X}^2(i)}} \exp\left(-\frac{1}{2\sigma_{0X}^2(i)}(x_i-\mu_{0X}(i))^2\right)}{\prod_{i=1}^{q_X} \frac{1}{\sqrt{2\pi\sigma_{1X}^2(i)}} \exp\left(-\frac{1}{2\sigma_{1X}^2(i)}(x_i-\mu_{1X}(i))^2\right)} \\ &= \sqrt{\prod_{i=1}^{q_X} \frac{\sigma_{1X}^2(i)}{\sigma_{0X}^2(i)}} \exp\left(\sum_{i=1}^{q_X}(C_{X1i}x_i^2 + C_{X2i}x_i + C_{X3i})\right) \\ &= \exp\left(C_{X4} + \left(\sum_{i=1}^{q_X}(C_{X1i}x_i^2 + C_{X2i}x_i)\right)\right),\end{aligned}$$

for some constants $C_{X4}, C_{X1i}, C_{X2i} \in \mathbb{R}$. Therefore, by Bayes' rule, the following holds,

$$\begin{aligned}P(D=1|X=x) &= \frac{1}{1+\frac{P(D=0)}{P(D=1)}\frac{P(X=x|D=0)}{P(X=x|D=1)}} \\ &= \frac{1}{1+\exp\left(-\left(C_{X5} + \left(\sum_{i=1}^{q_X}(-C_{X1i}x_i^2 - C_{X2i}x_i)\right)\right)\right)} \\ &= \text{expit}\left(C_{X5} - \left(\sum_{i=1}^{q_X}(C_{X1i}x_i^2 + C_{X2i}x_i)\right)\right),\end{aligned}$$

for some constants $C_{X5}, C_{X1i}, C_{X2i} \in \mathbb{R}$, which aligns with the functional form of a logistic regression model with quadratic terms.

(b) **Long:** We proceed analogously to the long case.

$$\begin{aligned}\frac{P(X=x,U=u|D=0)}{P(X=x,U=u|D=1)} &= \frac{P(X=x|D=0)P(U=u|D=0)}{P(X=x|D=1)P(U=u|D=1)} \text{ by conditional independence,} \\ &= \exp\left(C_{XU4} + \left(\sum_{i=1}^{q_X}(C_{X1i}x_i^2 + C_{X2i}x_i)\right) + \left(\sum_{i=1}^{q_U}(C_{U1i}u_i^2 + C_{U2i}u_i)\right)\right),\end{aligned}$$

for some constants $C_{XU4}, C_{X1i}, C_{X2i}, C_{U1i}, C_{U2i} \in \mathbb{R}$. Therefore, by Bayes' rule,

$$\begin{aligned}P(D=1|X=x,U=u) &= \frac{1}{1+\frac{P(D=0)}{P(D=1)}\frac{P(X=x,U=u|D=0)}{P(X=x,U=u|D=1)}} \\ &= \text{expit}\left(C_{XU5} - \left(\sum_{i=1}^{q_X}(C_{X1i}x_i^2 + C_{X2i}x_i) + \sum_{i=1}^{q_U}(C_{U1i}u_i^2 + C_{U2i}u_i)\right)\right),\end{aligned}$$

for some constants $C_{XU5}, C_{X1i}, C_{X2i}, C_{U1i}, C_{U2i} \in \mathbb{R}$, which aligns with the functional form of a logistic regression model with quadratic terms.

Next, we derive closed-form expressions for the bias factors in our simulation setting.

(a) $\boldsymbol{C^2_{0\Delta Y}}$ **:**

$$\begin{aligned}
E[(g_0 - g_{0s})^2|D=0] &= (\Delta\beta_u)^2 E[(U - E[U|D=0])^2|D=0] \text{ by } U \perp\!\!\!\perp X \mid D=0,\\
&= (\Delta\beta_u)^2 \operatorname{Var}(U|D=0) = 1 \times \sigma^2_{0U},\\
E[(\Delta Y - g_{0s})^2|D=0] &= E[(\Delta\beta_u(U - E[U|D=0]) + \Delta\tilde{\epsilon})^2|D=0] \text{ by } U \perp\!\!\!\perp X \mid D=0,\\
&= E[(g_0 - g_{0s})^2|D=0] + E[\Delta\tilde{\epsilon}^2|D=0] + 2E[(g_0 - g_{0s})\Delta\tilde{\epsilon}|D=0]\\
&= E[(g_0 - g_{0s})^2|D=0] + E[\Delta\tilde{\epsilon}^2|D=0] \text{ by independence},\\
&= \sigma^2_{0U} + 2,\\
\implies \quad C^2_{0\Delta Y} &= \frac{E[(g_0 - g_{0s})^2|D=0]}{E[(\Delta Y - g_{0s})^2|D=0]} = \frac{\sigma^2_{0U}}{\sigma^2_{0U}+2}.
\end{aligned}$$

(b) $\boldsymbol{C^2_{0D}}$ **:**

$$C^2_{0D} = \frac{\chi^2(P_{X,U|1}\|P_{X,U|0}) - \chi^2(P_{X|1}\|P_{X|0})}{\chi^2(P_{X|1}\|P_{X|0}) + 1},$$

which follows directly from Corollary 1. To compute this quantity in our simulation, we plug in the conditional density functions of X and U given D, and use conditional independence to simplify the calculation. Specifically,

$$\begin{aligned}
&\chi^2(P_{X|1}\|P_{X|0})\\
&= \int \frac{\mathrm{d}P_{X|1}(x)}{\mathrm{d}P_{X|0}(x)}\,\mathrm{d}P_{X|1}(x) - 1\\
&= \frac{\sigma_{0X}}{\sqrt{2\pi}\sigma^2_{1X}} \int \exp\left(-\frac{(x-\mu_{1X})^2}{\sigma^2_{1X}} + \frac{(x-\mu_{0X})^2}{2\sigma^2_{0X}}\right) dx - 1\\
&= \frac{\sigma_{0X}}{\sqrt{2\pi}\sigma^2_{1X}} \int \exp\left(- \underbrace{\left(\frac{1}{\sigma^2_{1X}} - \frac{1}{2\sigma^2_{0X}}\right)}_{a} x^2 + \underbrace{\left(\frac{2\mu_{1X}}{\sigma^2_{1X}} - \frac{\mu_{0X}}{\sigma^2_{0X}}\right)}_{b} x - \underbrace{\left(\frac{\mu^2_{1X}}{\sigma^2_{1X}} - \frac{\mu^2_{0X}}{2\sigma^2_{0X}}\right)}_{c} \right) dx - 1,
\end{aligned}$$

(1) If $a > 0$ (i.e., $2\sigma^2_{0X} > \sigma^2_{1X}$):

$$\begin{aligned}
\chi^2(P_{X|1}\|P_{X|0}) &= \frac{\sigma_{0X}}{\sqrt{2\pi}\sigma^2_{1X}} \exp\left(\frac{b^2}{4a} - c\right) \int \exp\left(-\frac{1}{2}\left(\sqrt{2a}x - \frac{b}{\sqrt{2a}}\right)^2\right) dx - 1\\
&= \frac{\sigma_{0X}}{\sqrt{2\pi}\sigma^2_{1X}} \exp\left(\frac{b^2}{4a} - c\right) \sqrt{\frac{\pi}{a}} - 1\\
&= \frac{\sigma^2_{0X}}{\sigma_{1X}\sqrt{2\sigma^2_{0X} - \sigma^2_{1X}}} \exp\left(\frac{(\mu_{1X} - \mu_{0X})^2}{2\sigma^2_{0X} - \sigma^2_{1X}}\right) - 1;
\end{aligned}$$

(2) Otherwise:

$$\chi^2(P_{X|1}\|P_{X|0}) = \infty.$$

Similarly,

$$\chi^2(P_{X,U|1}\|P_{X,U|0})$$
$$= \left(\chi^2(P_{X|1}\|P_{X|0})+1\right)\left(\chi^2(P_{U|1}\|P_{U|0})+1\right)-1 \text{ by conditional independence,}$$

(1) If $2\sigma_{0X}^2 > \sigma_{1X}^2$ and $2\sigma_{0U}^2 > \sigma_{1U}^2$:

$$\chi^2(P_{X,U|1}\|P_{X,U|0}) = \frac{\sigma_{0X}^2\sigma_{0U}^2}{\sigma_{1X}\sigma_{1U}\sqrt{(2\sigma_{0X}^2-\sigma_{1X}^2)(2\sigma_{0U}^2-\sigma_{1U}^2)}}\exp\left(\frac{(\mu_{1X}-\mu_{0X})^2}{2\sigma_{0X}^2-\sigma_{1X}^2}+\frac{(\mu_{1U}-\mu_{0U})^2}{2\sigma_{0U}^2-\sigma_{1U}^2}\right)-1;$$

(2) Otherwise:

$$\chi^2(P_{X,U|1}\|P_{X,U|0}) = \infty.$$

Accordingly, when $2\sigma_{0X}^2 > \sigma_{1X}^2$ and $2\sigma_{0U}^2 > \sigma_{1U}^2$,

$$\begin{aligned}
C_{0D}^2 &= \frac{\chi^2(P_{X,U|1}\|P_{X,U|0})-\chi^2(P_{X|1}\|P_{X|0})}{\chi^2(P_{X|1}\|P_{X|0})+1} \\
&= \frac{\left(\chi^2(P_{X|1}\|P_{X|0})+1\right)\left(\chi^2(P_{U|1}\|P_{U|0})+1\right)-1-\chi^2(P_{X|1}\|P_{X|0})}{\chi^2(P_{X|1}\|P_{X|0})+1} \\
&= \chi^2(P_{U|1}\|P_{U|0}) \\
&= \frac{\sigma_{0U}^2}{\sigma_{1U}\sqrt{2\sigma_{0U}^2-\sigma_{1U}^2}}\exp\left(\frac{(\mu_{1U}-\mu_{0U})^2}{2\sigma_{0U}^2-\sigma_{1U}^2}\right)-1.
\end{aligned}$$

(c) $\boldsymbol{\rho_0^2}$ **:**

$$\begin{aligned}
\theta_s &= E\left[(E[\Delta Y|D=1,X]-E[\Delta Y|D=0,X]) \mid D=1\right] \\
&= E\left[\frac{D}{p}\left(\theta+E[U\Delta\beta_u|D=1]-E[U\Delta\beta_u|D=0]\right)\right] \text{ by } U \perp\!\!\!\perp X \mid D=0, \\
&= \theta+(\mu_{1U}-\mu_{0U})\Delta\beta_u = \theta+(\mu_{1U}-\mu_{0U}).
\end{aligned}$$

Accordingly,

$$\begin{aligned}
&\mathrm{Cov}(g_0-g_{0s}, O_{XU}-O_X|D=0) \\
&= E[(g_0-g_{0s})(O_{XU}-O_X)|D=0] \\
&= O\times(\theta_0-\theta_{0s}) \\
&= O\times(\theta_s-\theta) \\
&= O\times(\mu_{1U}-\mu_{0U}),
\end{aligned}$$

where the second equality follows from the proof of Theorem 1. Additionally,

$$\mathrm{Var}(O_{XU}-O_X|D=0) = E[(O_{XU}-O_X)^2|D=0]$$

$$= E[O_{XU}^2|D=0] - E[O_X^2|D=0]$$
$$= O^2 \times (\chi^2(P_{X,U|1}\|P_{X,U|0}) - \chi^2(P_{X|1}\|P_{X|0})) \text{ by Proposition 3 (1),}$$

$$\text{Var}(g_0 - g_{0s} \mid D=0) = E[(g_0 - g_{0s})^2|D=0] = \sigma_{0U}^2.$$

Finally, we have

$$\begin{aligned}\rho_0^2 &= \frac{\text{Cov}^2(g_0 - g_{0s}, O_{XU} - O_X|D=0)}{\text{Var}(O_{XU} - O_X|D=0)\,\text{Var}(g_0 - g_{0s} \mid D=0)} \\ &= \frac{(\mu_{1U} - \mu_{0U})^2}{\sigma_{0U}^2 \times (\chi^2(P_{X,U|1}\|P_{X,U|0}) - \chi^2(P_{X|1}\|P_{X|0}))} \\ &= \frac{(\mu_{1U} - \mu_{0U})^2}{\sigma_{0U}^2 \times (\chi^2(P_{X|1}\|P_{X|0}) + 1)\chi^2(P_{U|1}\|P_{U|0})}.\end{aligned}$$

E.4 Simulation: figures and tables

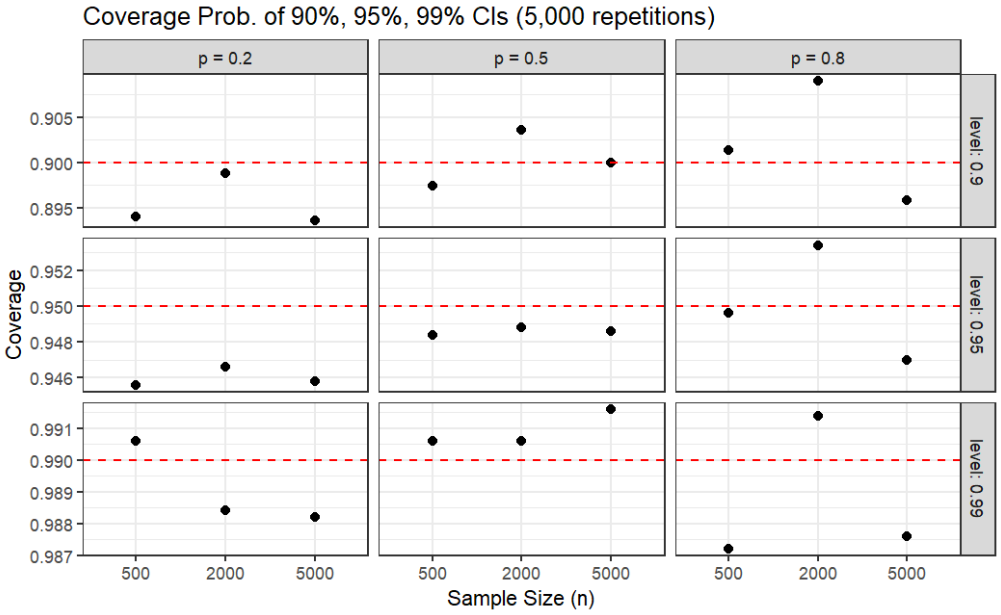


Figure 4: Empirical coverage probabilities of 90%, 95%, and 99% confidence bounds across sample sizes n and treatment probabilities p, using the true bias factors; red dashed lines indicate nominal coverage levels.

p	n	Coverage
0.2	500	0.894
	2000	0.899
	5000	0.894
0.5	500	0.897
	2000	0.904
	5000	0.900
0.8	500	0.901
	2000	0.909
	5000	0.896

(a) Confidence level = 0.90

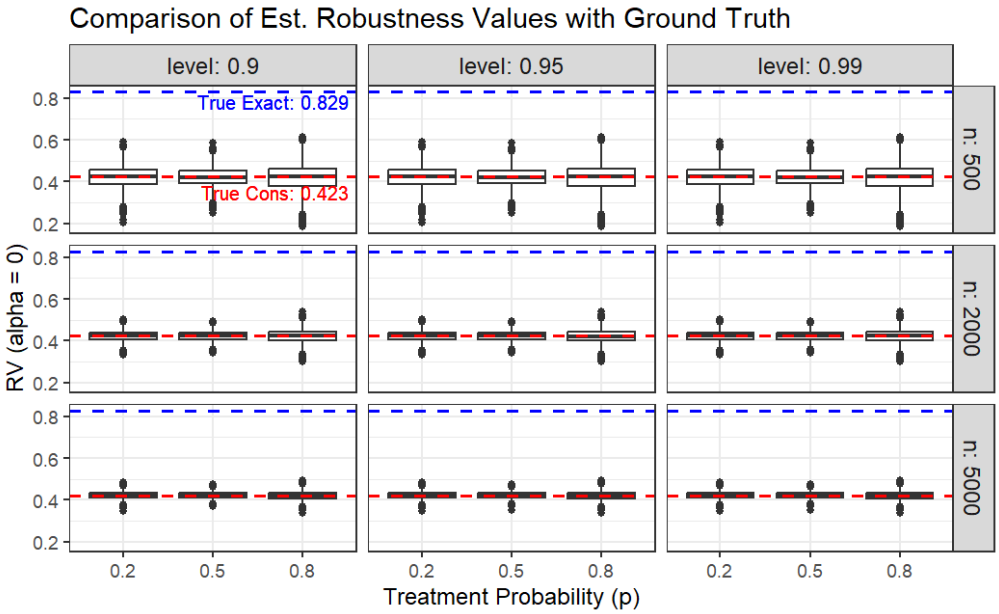
p	n	Coverage
0.2	500	0.946
	2000	0.947
	5000	0.946
0.5	500	0.948
	2000	0.949
	5000	0.949
0.8	500	0.950
	2000	0.953
	5000	0.947

(b) Confidence level = 0.95

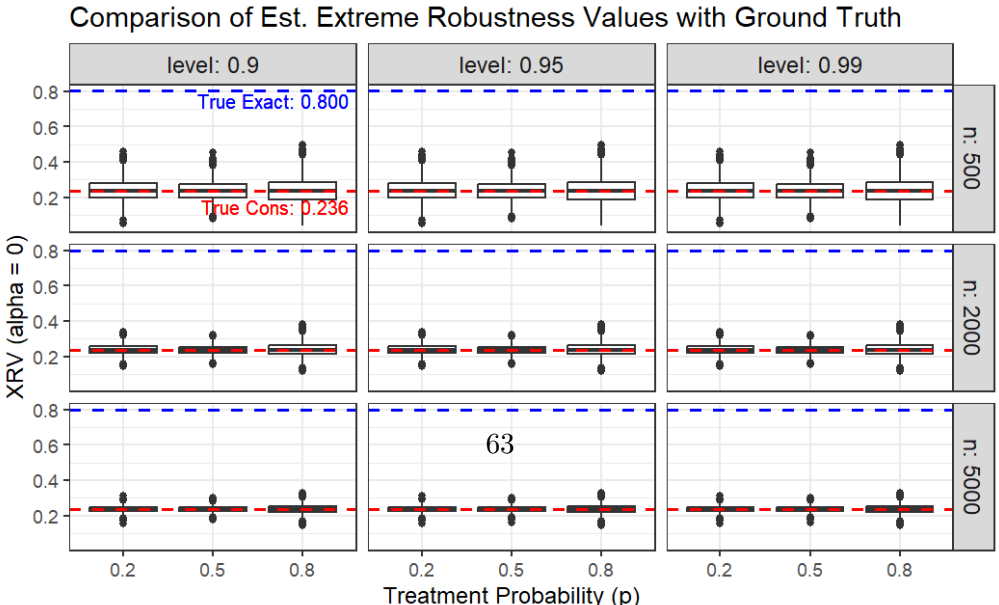
p	n	Coverage
0.2	500	0.991
	2000	0.988
	5000	0.988
0.5	500	0.991
	2000	0.991
	5000	0.992
0.8	500	0.987
	2000	0.991
	5000	0.988

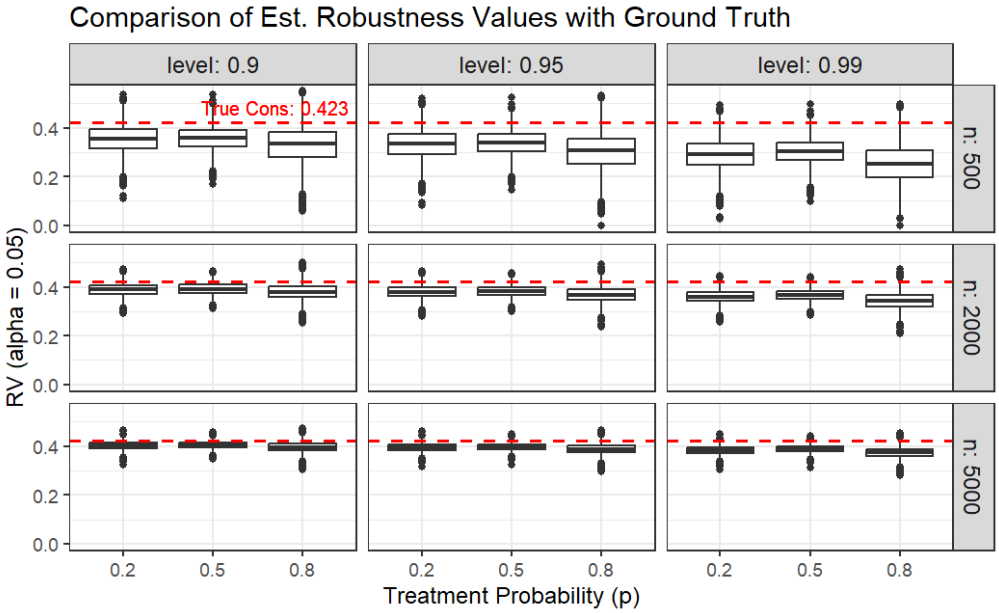
(c) Confidence level = 0.99

Table 5: Empirical coverage probabilities of 90%, 95%, and 99% confidence bounds across sample sizes n and treatment probabilities p, using the true bias factors.

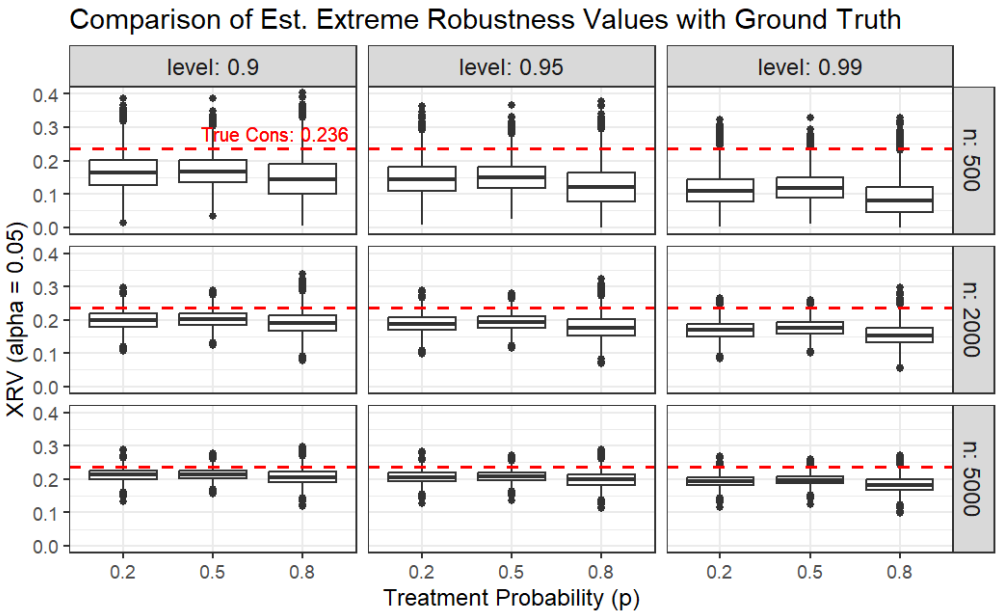


(a) $\mathrm{RV}_{\theta^*=0,\alpha=0}$.





(a) $\text{RV}_{\theta^*=0,\alpha=0.05}$.



(b) $\text{XRV}_{\theta^*=0,\alpha=0.05}$.

Figure 6: Empirical estimates of $\text{RV}_{\theta^*=0,\alpha=0.05}$ and $\text{XRV}_{\theta^*=0,\alpha=0.05}$ compared with their true values (red dashed line) across sample sizes and treatment probabilities (5,000 repetitions). Red dashed lines indicate the conservative true reference values by setting $|\rho_0| = 1$.

Learner	**Variables**	**Package**	**Tuning grid**
Parametric: Linear for g_{0s} Logistic for π	$X^*, U^*, (X^*)^2, (U^*)^2$	`stats` (`lm`) `stats` (`glm`)	N/A
Ridge: Linear for g_{0s} Logistic for π	$X^*, U^*, (X^*)^2, (U^*)^2,$ $(X^*)^3, (U^*)^3$	`glmnet`	$\alpha_{\text{EN}} = 0$ (ridge) λ selected from `cv.glmnet`
Lasso: Linear for g_{0s} Logistic for π	$X^*, U^*, (X^*)^2, (U^*)^2,$ $(X^*)^3, (U^*)^3$	`glmnet`	$\alpha_{\text{EN}} = 1$ (lasso) λ selected from `cv.glmnet`
Random Forest (`num.trees` $= 500$)	X^*, U^*	`ranger`	`mtry` $\in \{2, 3\}$ `min.node.size` $\in \{50, 100, \ldots, 300\}$ `splitrule` $\in \{$`variance`, `extratrees`$\}$

Table 6: Machine learning methods used for first-stage nuisance estimation. Covariates enter linearly. Models are selected by minimizing the RMSE. Across 500 repetitions, random forest is selected in 87% of replications for the propensity score π and in 98.8% of replications for the outcome evolution model g_{0s}.

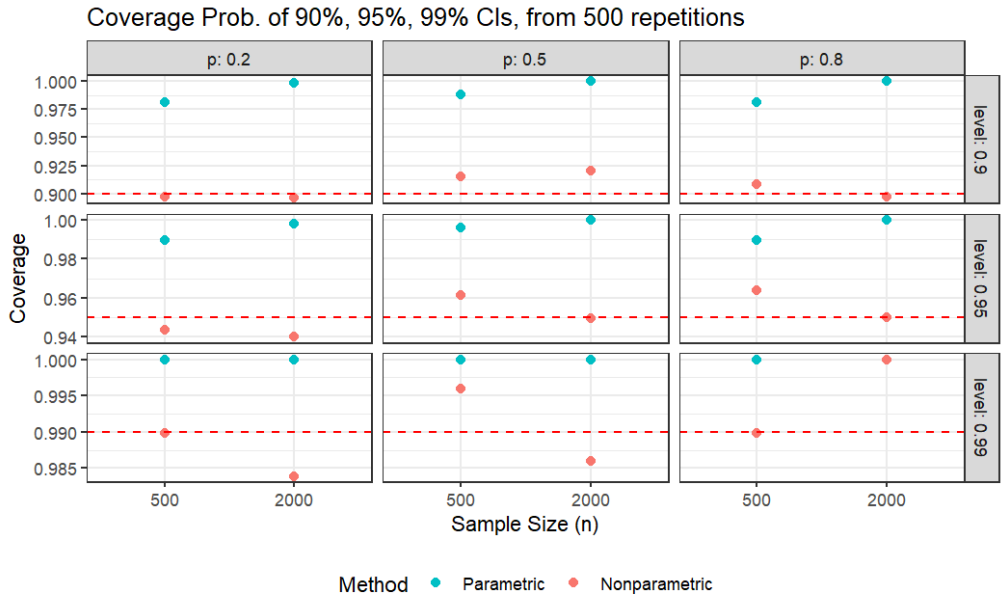


Figure 7: Comparison of empirical coverage probabilities between parametric and nonparametric confidence bounds, using the true bias factors; red dashed lines indicate nominal coverage levels.

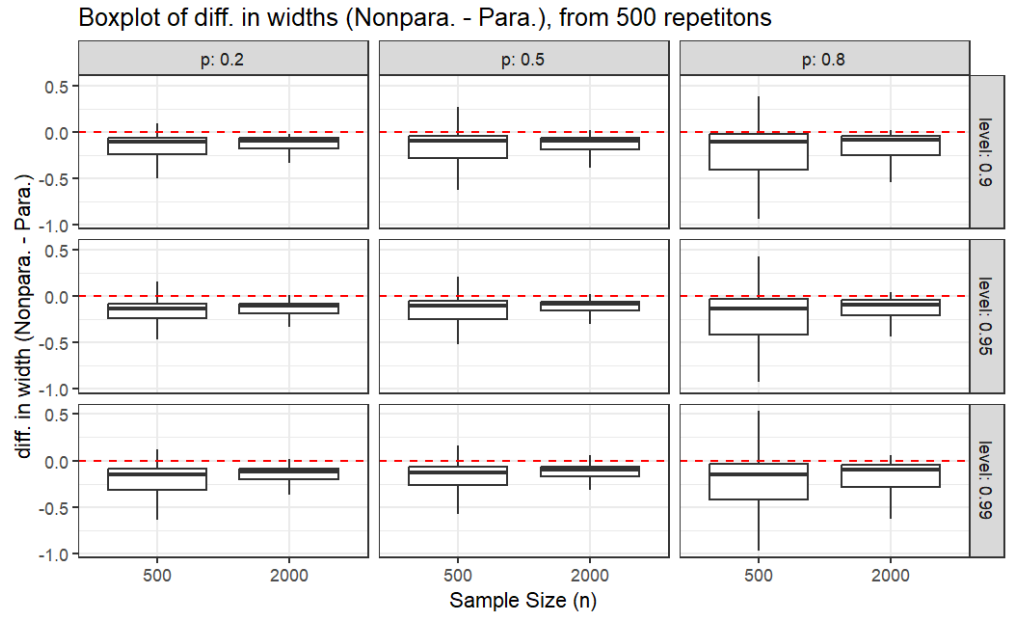


Figure 8: Boxplots of the difference in confidence-interval widths (nonparametric minus parametric) across sample sizes and treatment probabilities; the red dashed line denotes zero difference.

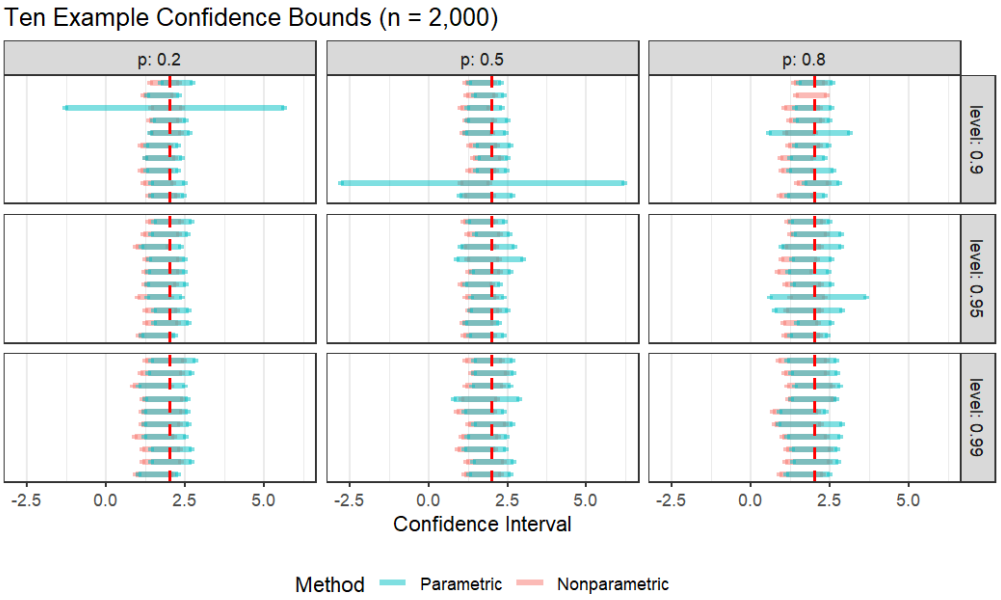


Figure 9: Comparison of parametric and nonparametric example confidence bounds under model misspecification; red dashed lines indicate true θ value.

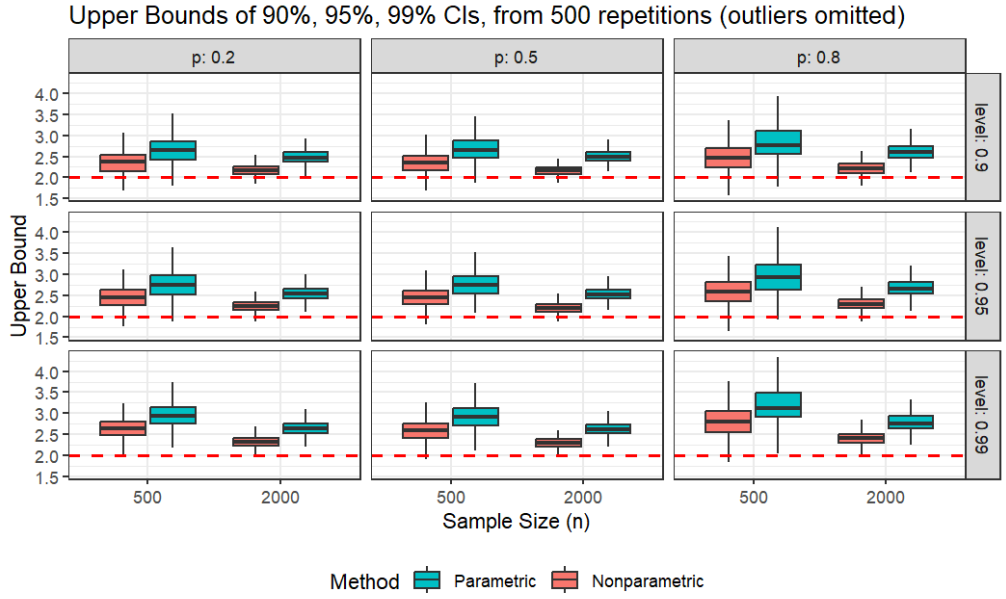
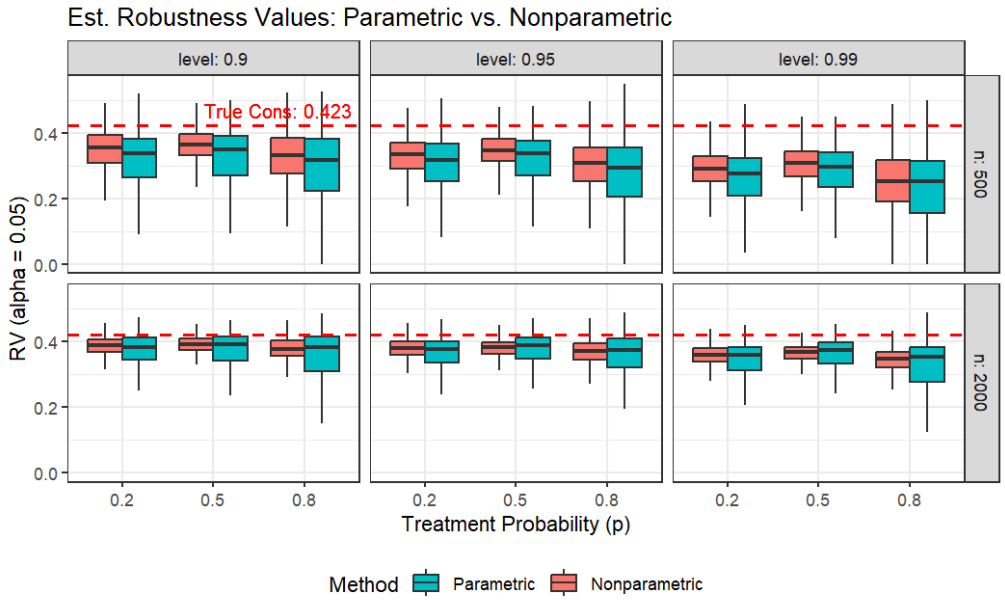
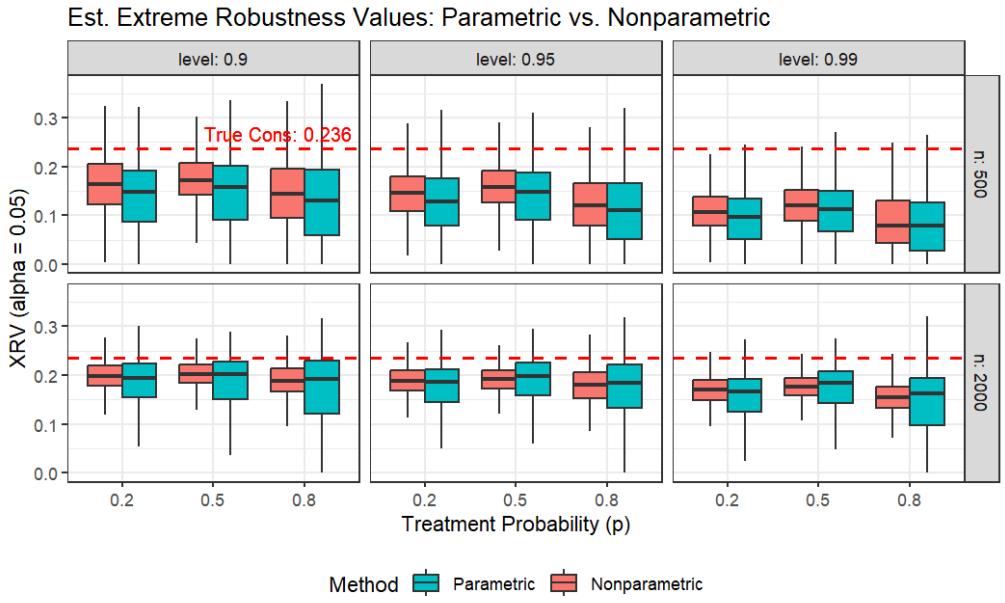


Figure 10: Comparison of parametric and nonparametric upper confidence bounds under model misspecification; outliers are omitted for clarity. Red dashed lines indicate the true θ value.



(a) $\text{RV}_{\theta^*=0,\alpha=0.05}$.



(b) $\text{XRV}_{\theta^*=0,\alpha=0.05}$.

Figure 11: Finite-sample comparison of parametric and nonparametric estimators for $\text{RV}_{\theta^*=0,\alpha=0.05}$ (left) and $\text{XRV}_{\theta^*=0,\alpha=0.05}$ (right) across sample sizes and treatment probabilities (500 repetitions). Red dashed lines indicate the conservative true reference values by setting $|\rho_0| = 1$.

F Comparison with related literature

F.1 Relationship to the "unconditional" OVB analysis

In this section, we compare our OVB analysis for the ATT with the results of Chernozhukov et al. (2026) as applied to the ATT, which were used by Bach et al. (2025). We refer to that result as "unconditional." Both our results and the unconditional results target the same causal estimand and characterize the same total bias. The distinction lies in how confounding strength is parameterized and interpreted.

F.1.1 Review of the unconditional results

We begin by briefly reviewing the results of Chernozhukov et al. (2026). Under the same identification assumptions as those imposed in our main analysis, they start with the following parameterization of the ATT:

$$\theta = E[g(D,X,U)\alpha(D,X,U)], \text{ and } \theta_s = E[g_s(D,X)\alpha_s(D,X)] \text{ with}$$
$$g(d,X,U) := E[\Delta Y \mid D=d,X,U], \text{ and } g_s(d,X) := E[\Delta Y \mid D=d,X],$$
$$\alpha(d,X,U) := \frac{D}{p} - \frac{1-D}{1-p} \times \frac{O_{XU}}{O}, \text{ and } \alpha_s(d,X) := \frac{D}{p} - \frac{1-D}{1-p} \times \frac{O_X}{O}.$$

This leads to the following OVB result for the ATT:

$$\theta - \theta_s = \rho \times \underbrace{\sqrt{R^2_{\Delta Y - g_s \sim g - g_s}}}_{=:C_{\Delta Y}} \times \underbrace{\sqrt{\frac{1-R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}}}}_{=:C_D} \times \underbrace{\sqrt{E[(\Delta Y - g_s)^2] \times E[\alpha_s^2]}}_{=:S},$$

$$\text{where } \rho = \operatorname{Cor}\left(g - g_s, \alpha - \alpha_s\right),$$

which can be further specialized as:

$$\theta - \theta_s = \rho \times \underbrace{\sqrt{\eta^2_{\Delta Y \sim U \mid X,D}}}_{=:C_{\Delta Y}} \times \underbrace{\sqrt{\frac{E[O_{XU}] - E[O_X]}{E[O_X]}}}_{=:C_D} \times \underbrace{\sqrt{E[\operatorname{Var}(\Delta Y \mid X, D)] \times \left(\frac{E[O_X]}{p^2}\right)}}_{=:S},$$

$$\text{where } \rho = \operatorname{Cor}\left(E[\Delta Y \mid D,X,U] - E[\Delta Y \mid D,X], -(1-D) \times (O_{XU} - O_X)\right).$$

F.1.2 Alternative ways to quantify selection strength for the unconditional analysis

For completeness, we now provide alternative characterizations of selection strength for the unconditional results, expressed in terms of R^2 measures in odds or χ^2 divergence. These characterizations show that their selection strength measure increases with the treatment probability.

Proposition 7 (Alternative characterization of C_D^2)**.**

$$C_D^2 = \frac{1 - R^2_{O_{XU} \sim O_X | D=0}}{\frac{1}{O} R^2_{O_{XU} \sim O | D=0} + R^2_{O_{XU} \sim O_X | D=0}} \qquad \text{(Residual } R^2 \text{ in Selection Odds)}$$

$$= \frac{\chi^2(P_{X,U|1} \| P_{X,U|0}) - \chi^2(P_{X|1} \| P_{X|0})}{\chi^2(P_{X|1} \| P_{X|0}) + \frac{1}{p}}. \qquad \text{(Distributional Imbalance)}$$

F.1.3 Difference in parameterization

For the ATT, the only unobserved component is the untreated counterfactual trend for the treated units. The treated trend for treated units is observed and therefore does not need a bias analysis. This allows us to localize the bias conditionally on the untreated group. Note that the unconditional result does not take this information into account.

We show that $C^2_{\Delta Y}$ is a weighted average of the strength of confounding on trends in the treated and control groups. In particular, as treatment becomes more likely, the weight placed on the control-group component, where the bias originates, decreases. Consequently, practitioners must rely on a pooled measure of confounding, which obscures the source of the bias and may complicate interpretation. Similarly, we show that C_D^2 is a weighted version of C^2_{0D}. In particular, $C_D^2 < C^2_{0D}$, with a weight on C^2_{0D} that decreases as treatment becomes less likely. Likewise, ρ^2 is a downweighted version of ρ_0^2, with the weight decreasing as the treatment probability increases.

Proposition 8 (Relationship with unconditional)**.**

(i) $C^2_{0\Delta Y}$ ***vs*** C^2_Y**.**

$$C^2_{\Delta Y} = W_{0\Delta Y} C^2_{0\Delta Y} + (1 - W_{0\Delta Y}) C^2_{1\Delta Y},$$

$$\textit{with } C^2_{0\Delta Y} = \eta^2_{\Delta Y \sim U | X, D=0},\ C^2_{1\Delta Y} = \eta^2_{\Delta Y \sim U | X, D=1},\ \textit{and}$$

$$W_{0\Delta Y} = \frac{(1-p) \times \sigma^2_{0s}}{p \times \sigma^2_{1s} + (1-p) \times \sigma^2_{0s}},$$

where $\sigma^2_{ds} = E[\mathrm{Var}(\Delta Y \mid X, D = d) \mid D = d]$ *for* $d \in \{0, 1\}$.

(ii) C^2_{0D} ***vs*** C^2_D**.**

$$C^2_D = W_{0D} C^2_{0D},$$

$$\textit{with } W_{0D} = \frac{O \times (\chi^2(P_{X|1} \| P_{X|0}) + 1)}{O \times (\chi^2(P_{X|1} \| P_{X|0}) + 1) + 1}.$$

(iii) ρ_0 ***vs*** ρ**.** *Define* $g_d := E[\Delta Y \mid X, U, D = d]$ *and* $g_{ds} := E[\Delta Y \mid X, D = d]$.

$$\rho^2 = W_{0\rho} \rho_0^2,$$

$$with\ W_{0\rho} = \frac{1}{\frac{\mathrm{Var}(g_1 - g_{1s}|D=1)}{\mathrm{Var}(g_0 - g_{0s}|D=0)} \times O + 1}.$$

F.1.4 Simulation: comparison with unconditional

In this section, we use Monte Carlo simulations to compare the finite-sample properties of our results with the unconditional results of Chernozhukov et al. (2026). We follow the same data-generating process as in Section E.1.

1. **Coverage and width of confidence bounds.**

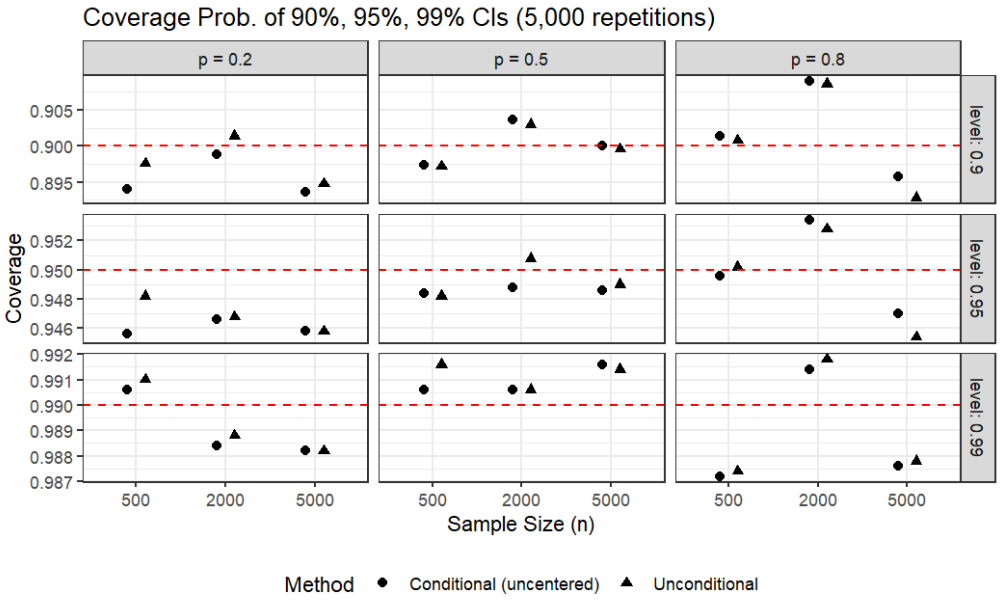


Figure 12: Empirical coverage probabilities of 90%, 95%, and 99% confidence bounds across sample sizes n and treatment probabilities p, using the true bias factors; red dashed lines indicate nominal coverage levels.

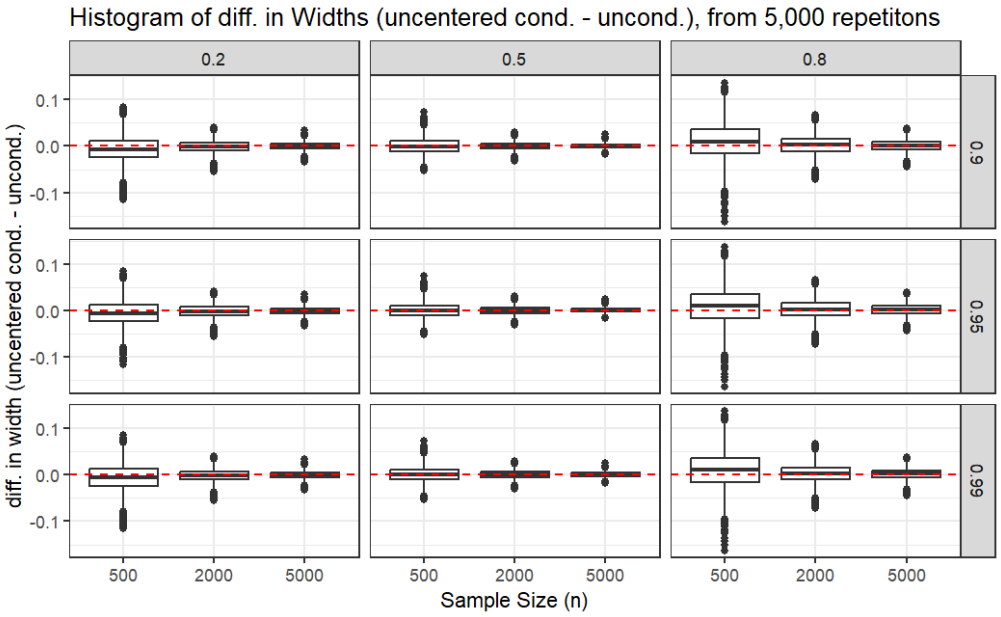
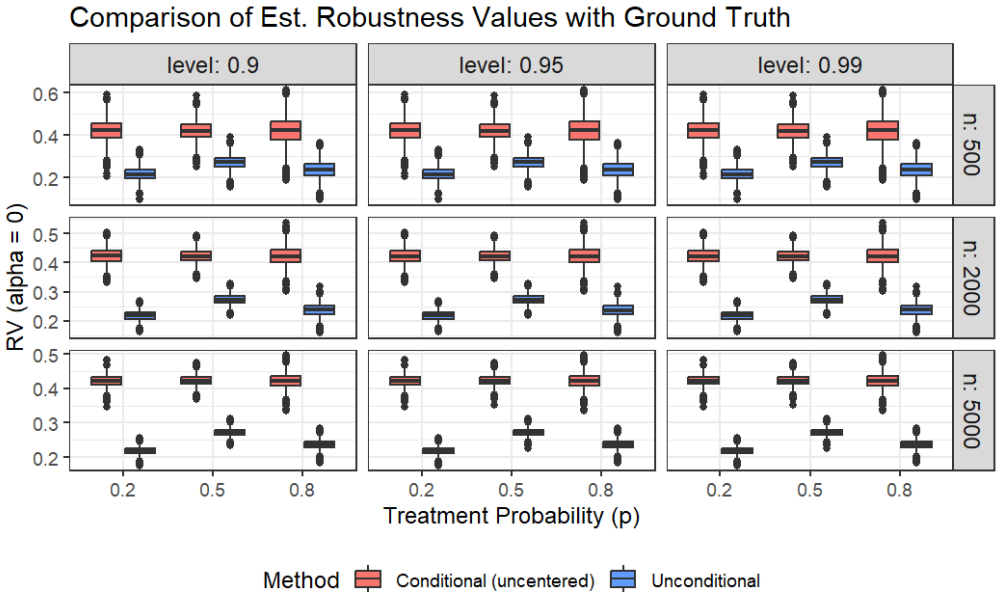
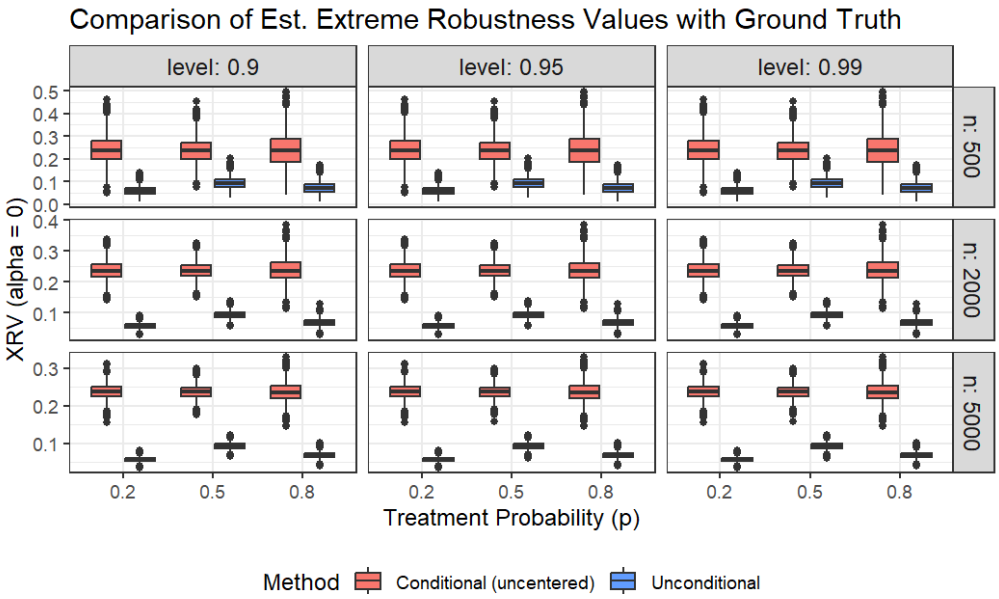


Figure 13: Boxplots of the difference in confidence-interval widths (uncentered conditional minus unconditional) across sample sizes and treatment probabilities; the red dashed line denotes zero difference.

2. **Sensitivity statistics.**

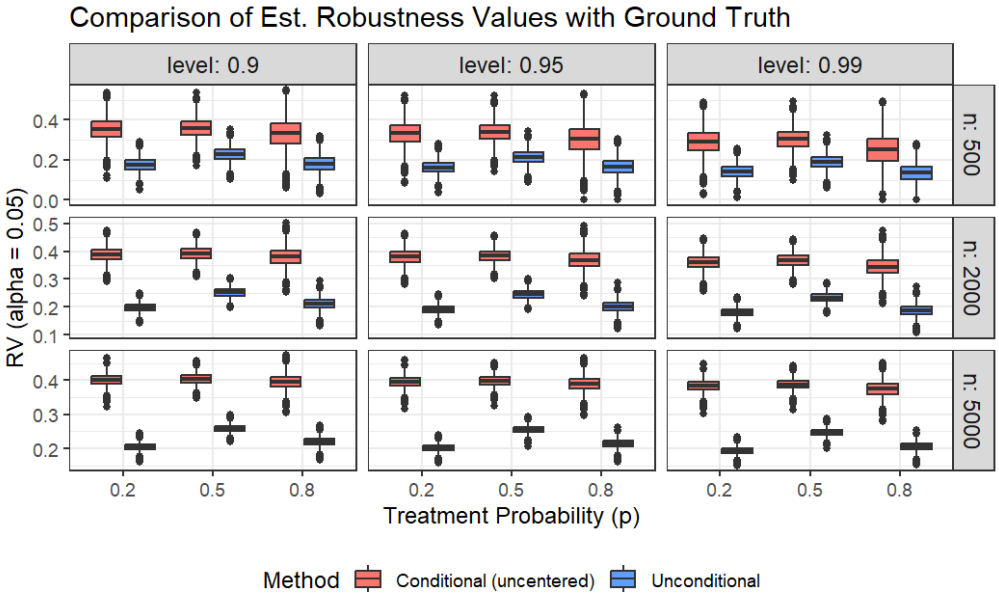


(a) $\text{RV}_{\theta^*=0,\alpha=0}$.

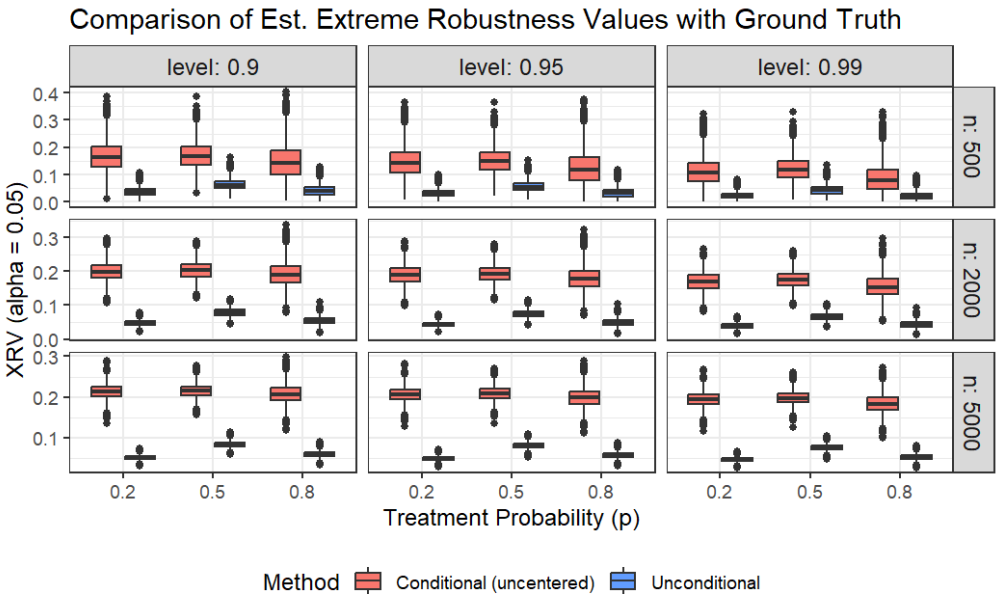


(b) $\text{XRV}_{\theta^*=0,\alpha=0}$.

Figure 14: Empirical estimates of $\text{RV}_{\theta^*=0,\alpha=0}$ and $\text{XRV}_{\theta^*=0,\alpha=0}$ across sample sizes and treatment probabilities (5,000 repetitions).



(a) $\text{RV}_{\theta^*=0,\alpha=0.05}$.



(b) $\text{XRV}_{\theta^*=0,\alpha=0.05}$.

Figure 15: Empirical estimates of $\text{RV}_{\theta^*=0,\alpha=0.05}$ and $\text{XRV}_{\theta^*=0,\alpha=0.05}$ across sample sizes and treatment probabilities (5,000 repetitions).

F.2 Relationship to variance-based sensitivity analysis for weighting estimators

This section compares the results of Huang and Pimentel (2025), hereafter HP, with our approach. As before, both results target the same causal estimand and characterize the same total bias. They

differ in (1) how confounding strength is parameterized, (2) how benchmarking is performed, and (3) how statistical inference is conducted.

F.2.1 Review of HP

We begin by briefly reviewing the IPW formulation of Huang and Pimentel (2025) (hereafter, HP). To facilitate comparison, we express their characterization in terms of ΔY, using our notation and indexing HP's bias factors by w to distinguish them from ours.

To begin with, they parameterize the long and short parameters of interest in the following way:

$$\theta_0 = E[\alpha_0(X,U)\Delta Y \mid D=0], \text{ and } \theta_{0s} = E[\alpha_{0s}(X)\Delta Y \mid D=0] \text{ with}$$
$$\alpha_0(X,U) = \frac{O_{XU}}{O}, \text{ and } \alpha_{0s}(X) = \frac{O_X}{O}.$$

This leads to the following OVB result for the ATT:

$$\theta - \theta_s = -\rho_{w0} \times \underbrace{\sqrt{\frac{1 - R^2_{\alpha_0 \sim \alpha_{0s}|1,D=0}}{R^2_{\alpha_0 \sim \alpha_{0s}|1,D=0}}}}_{=:C_{w0D}} \times \underbrace{\sqrt{\text{Var}(\Delta Y \mid D=0) \times \text{Var}(\alpha_{0s} \mid D=0)}}_{=:S_{w0}},$$

where $\rho_{w0} = \text{Cor}(\Delta Y, \alpha_0 - \alpha_{0s} \mid D=0)$.

We define $\sigma^2_{w0s} := \text{Var}(\Delta Y \mid D=0)$ and $\nu^2_{w0s} := \text{Var}(\alpha_{0s} \mid D=0)$. Additionally, the correlation can be further bounded as $\rho_{w0} \leq \sqrt{1 - \text{Cor}^2(\alpha_{0s}, \Delta Y \mid D=0)}$.

F.2.2 Alternative ways to quantify selection strength for HP

HP do not explore alternative parameterizations of C^2_{w0D} in terms of odds or χ^2-divergence. For completeness, we provide these representations here.

Proposition 9 (Alternative characterizations of C^2_{w0D})**.**

$$\begin{aligned} C^2_{w0D} &= \frac{1 - R^2_{O_{XU} \sim O_X|1,D=0}}{R^2_{O_{XU} \sim O_X|1,D=0}} && \text{(Residual } R^2 \text{ in Selection Odds)} \\ &= \frac{E[O_{XU}] - E[O_X]}{E[O_X] - O} && \text{(Average Selection Odds)} \\ &= \frac{\chi^2(P_{X,U|1} \| P_{X,U|0}) - \chi^2(P_{X|1} \| P_{X|0})}{\chi^2(P_{X|1} \| P_{X|0})}. && \text{(Distributional Imbalance)} \end{aligned}$$

Proof. The corresponding proof follows steps similar to those in Corollary 1. □

Remark 7. *For the scaling factors, note that* $\nu^2_{w0s} = \chi^2(P_{X|1} \| P_{X|0}) < \chi^2(P_{X|1} \| P_{X|0}) + 1 = \nu^2_{0s}$. *Therefore, when the variance of observable weights is small (i.e.,* ν^2_{w0s} *is close to zero), estimation*

noise, particularly from propensity score estimation, can lead to negative estimates of ν^2_{w0s}. This issue is mitigated in the uncentered parameterization.

F.2.3 Difference in parameterization

Note that our definition of C^2_{0D}, which measures the strength of confounding on selection, takes a form similar to HP's C^2_{w0D}. The main distinction is that we use the uncentered definition of R^2, while HP use a centered version that partials out the constant. As a result, HP's approach does not accommodate settings in which observed covariates do not induce any group imbalance, or in the case without covariates. In these cases, C^2_{w0D} suffers from a zero-denominator issue. HP recognize this issue in their paper, *"Since our sensitivity analysis is based on variances of weights, it is not equipped to address situations in which either the observable or ideal weights are all identical, with zero variance."*

We now provide connections between HP's result and ours.

Proposition 10 (Relationship with HP)**.**

(i) C^2_{0D} ***vs*** C^2_{0wD}**.**

$$C^2_{0wD} = C^2_{0D} \times \frac{E[O_X \mid D=1]}{E[O_X \mid D=1] - O} \left(= C^2_{0D} \times \frac{1}{1 - R^2_{\alpha_{0s} \sim 1 \mid D=0}} \right).$$

(ii) ρ_0 ***vs*** ρ_{w0}**.**

$$\rho^2_{w0} = \rho^2_0 \times C^2_{0\Delta Y} \times (1 - R^2_{\Delta Y \sim g_{0s} \mid 1, D=0}).$$

Remark 8. *Since* $|\rho_{w0}| \leq \sqrt{1 - \mathrm{Cor}^2(\alpha_{0s}, \Delta Y \mid D=0)}$*, it follows that*

$$\rho^2_0 \times C^2_{0\Delta Y} \leq \frac{1 - \mathrm{Cor}^2(\alpha_{0s}, \Delta Y \mid D=0)}{1 - R^2_{\Delta Y \sim g_{0s} \mid 1, D=0}} \left(= \frac{\textit{residual alignment room}}{\textit{unexplained trend variation}} \right),$$

where the upper bound on the RHS is estimable from data.

Next, we show how their upper bound on $|\rho_{w0}|^2$ relates to our results. Recall that $g_{0s}(X) = E[\Delta Y \mid X, D=0]$ attains the maximal correlation with ΔY among all square-integrable functions of X within the control group. This characterization follows from classical results on maximal correlation (see, e.g., Rényi (1959), Theorem 1). Formally,

$$g_{0s}(X) = \arg \max_{f \in L^2(P_{X \mid D=0})} \mathrm{Cor}^2(f(X), \Delta Y \mid D=0).$$

Using this result, we obtain the following relationship.

Proposition 11 (Relationship with HP's upper bound on alignment)**.**

$$\rho_{w0}^2 = 1 - \mathrm{Cor}^2(\alpha_{0s}, \Delta Y \mid D=0) \quad \textit{implies} \quad \rho_0^2 C_{0\Delta Y}^2 = 1.$$

Proof. By Proposition 10,

$$\rho_0^2 C_{0\Delta Y}^2 = \frac{\rho_{w0}^2}{1 - R^2_{\Delta Y \sim g_{0s}|1,D=0}}.$$

Therefore, when ρ_{w0}^2 equals $1 - \mathrm{Cor}^2(\alpha_{0s}, \Delta Y \mid D=0)$, this implies the following

$$\rho_0^2 C_{0\Delta Y}^2 = \frac{1 - \mathrm{Cor}^2(\alpha_{0s}, \Delta Y \mid D=0)}{1 - R^2_{\Delta Y \sim g_{0s}|1,D=0}}.$$

Since $g_{0s}(X) = \arg\max_{f \in L^2(P_{X|D=0})} \mathrm{Cor}^2(f(X), \Delta Y \mid D=0)$,

$$1 = \frac{1 - \mathrm{Cor}^2(g_{0s}, \Delta Y \mid D=0)}{1 - R^2_{\Delta Y \sim g_{0s}|1,D=0}} \leq \frac{1 - \mathrm{Cor}^2(\alpha_{0s}, \Delta Y \mid D=0)}{1 - R^2_{\Delta Y \sim g_{0s}|1,D=0}} = \rho_0^2 C_{0\Delta Y}^2 \leq 1,$$

$$\implies \quad \rho_0^2 C_{0\Delta Y}^2 = 1.$$

□

F.2.4 Difference in benchmarking

We now show how the benchmarking approach proposed by Huang and Pimentel (2025) is anti-conservative compared with ours.

First, note we can write their benchmarking multiple in the following way.

Proposition 12.

$$k_{0D,j}^{HP} = \frac{1 - R^2_{O_{XU} \sim O_X|1,D=0}}{R^2_{O_{XU} \sim O_X|1,D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|1,D=0}} = \frac{1 - R^2_{O_{XU} \sim O_X|D=0}}{R^2_{O_{XU} \sim O_X|D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|D=0}}.$$

Proof.

$$\begin{aligned}
k_{0D,j}^{\mathrm{HP}} &= \frac{1 - R^2_{O_{XU} \sim O_X|1,D=0}}{R^2_{O_{XU} \sim O_X|1,D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|1,D=0}} \\
&= \frac{1 - \frac{\mathrm{Var}(O_X|D=0)}{\mathrm{Var}(O_{XU}|D=0)}}{\frac{\mathrm{Var}(O_X|D=0)}{\mathrm{Var}(O_{XU}|D=0)} - \frac{\mathrm{Var}(O_{X_{-j}}|D=0)}{\mathrm{Var}(O_{XU}|D=0)}} \\
&= \frac{\mathrm{Var}(O_{XU} \mid D=0) - \mathrm{Var}(O_X \mid D=0)}{\mathrm{Var}(O_X \mid D=0) - \mathrm{Var}(O_{X_{-j}} \mid D=0)} \\
&= \frac{(E[O_{XU}^2 \mid D=0] - O^2) - (E[O_X^2 \mid D=0] - O^2)}{(E[O_X^2 \mid D=0] - O^2) - (E[O_{X_{-j}}^2 \mid D=0] - O^2)} \text{ by Proposition 2(2),} \\
&= \frac{E[O_{XU}^2 \mid D=0] - E[O_X^2 \mid D=0]}{E[O_X^2 \mid D=0] - E[O_{X_{-j}}^2 \mid D=0]} \\
&= \frac{1 - R^2_{O_{XU} \sim O_X|D=0}}{R^2_{O_{XU} \sim O_X|D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|D=0}}.
\end{aligned}$$

□

Our benchmarking metric is,

$$k_{0D,j} := \frac{R^2_{O_X \sim O_{X_{-j}}|D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|D=0}}{1 - R^2_{O_X \sim O_{X_{-j}}|D=0}},$$

while their benchmarking metric, by Proposition 12, is

$$k^{\text{HP}}_{0D,j} = \frac{1 - R^2_{O_{XU} \sim O_X|D=0}}{R^2_{O_{XU} \sim O_X|D=0} - R^2_{O_{XU} \sim O_{X_{-j}}|D=0}}.$$

Note that

$$k_{0D,j} = k^{\text{HP}}_{0D,j} \times R^2_{O_{XU} \sim O_{X_{-j}}|D=0},$$

which shows that their benchmarking exercise will be anti-conservative.

To illustrate, consider the following example.

Example 1. *We consider an example based on Cinelli and Hazlett (2025b) and generate data as follows. Let the observed covariate $X \sim \mathcal{N}(0,1)$ and let the unobserved covariate $U \sim \mathcal{N}(0,1)$ be independent of X. The treatment assignment is determined by*

$$D = \mathbf{1}\left\{X/2 + U/2 + \varepsilon_D > 0\right\},$$

where ε_D is independent and standard normal. The observed outcome evolution is given by

$$\Delta Y = X + U + \varepsilon_Y,$$

where ε_Y is independent and standard normal.

The propensity score is correctly specified by a probit model for D on (X, U), and the outcome evolution model is correctly specified by a linear regression of ΔY on (X, U). Note that in this design, $\theta = 0$, and U is exactly like X in terms of its strength of association with the treatment selection and the outcome evolution.

We simulate a sample of size $n = 10^6$, fit correctly specified nuisance models, and set the benchmarking metric to one (i.e., $k_{0D,j} = k^{\text{HP}}_{0D,j} = 1$), reflecting the belief that U and X are exchangeable. We plug in the empirically estimated true strength of U on outcome evolution. We then examine whether the resulting confidence bounds implied by HP's benchmarking proposal for treatment selection cover the true parameter value $\theta = 0$.

The results are visualized in Figure 16, which shows that HP's benchmark point remains far from zero. This would lead a practitioner to incorrectly conclude that an unobserved confounder U comparable to X cannot explain away the observed effect. By contrast, our benchmark point crosses

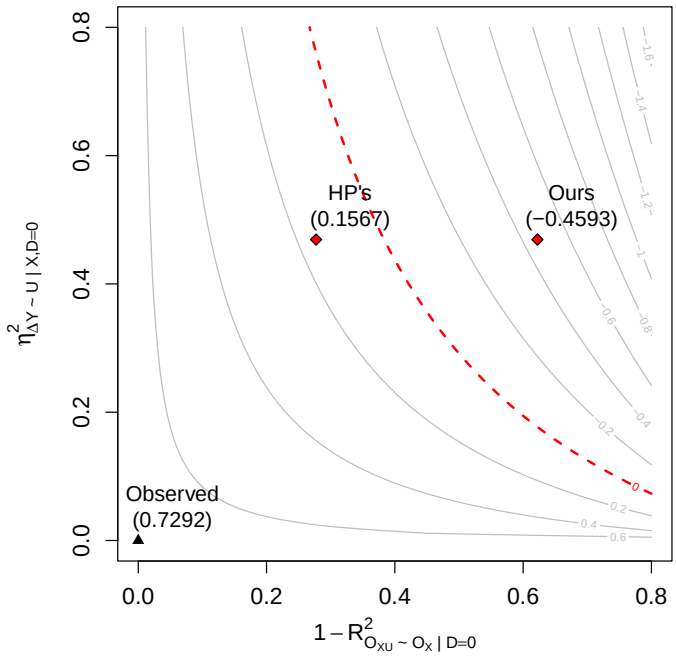


Figure 16: Sensitivity contour plots at the significance level of $\alpha = 0.05$.

zero, and correctly indicates that a confounder as strong as X could explain away the observed effect.

F.2.5 Difference in statistical inference

Huang and Pimentel (2025) focus on parametric nuisance estimators, and recommend using the percentile bootstrap to construct asymptotically valid confidence intervals for θ, under pre-specified restrictions on confounding strength. For completeness, here, we provide additional results for the centered OVB formula, which allows inference with DML. In particular,

$$\theta - \theta_s = -\rho_0 C_{0\Delta Y} C_{w0D} \sqrt{\sigma^2_{0s} \nu^2_{w0s}},$$

where,

$$\rho_0 := \mathrm{Cor}\left(g_0 - g_{0s}, O_{XU} - O_X \mid D = 0\right), \quad C^2_{0\Delta Y} := \eta^2_{\Delta Y \sim U \mid X, D=0}, \quad C^2_{w0D} := \frac{1 - R^2_{O_{XU} \sim O_X \mid 1, D=0}}{R^2_{O_{XU} \sim O_X \mid 1, D=0}},$$

and

$$\sigma^2_{0s} := E[\mathrm{Var}(\Delta Y \mid X, D = 0) \mid D = 0], \qquad \nu^2_{w0s} := E\left[\left(\frac{O_X}{O}\right)^2 \middle| D = 0\right] - 1.$$

In this characterization, θ_s and σ_{0s}^2 can be estimated using the same scores as in (7)–(8). We derive the following debiased score for ν_{w0s}^2:

$$\psi_{\nu_{w0s}^2}(Z;\pi,p) = 2\frac{D}{p}\left(\frac{O_X}{O} - (\nu_{w0s}^2+1)\right) - \frac{1-D}{1-p}\left(\left(\frac{O_X}{O}\right)^2 - (\nu_{w0s}^2+1)\right).$$

F.2.6 Simulation: comparison with HP

In this section, we use Monte Carlo simulations to compare the finite-sample properties of our approach with those of the "centered" approach introduced in Section F.2.5, both implemented via DML. We follow the same data-generating process as in Section E.1.

1. **Coverage and width of confidence bounds.**

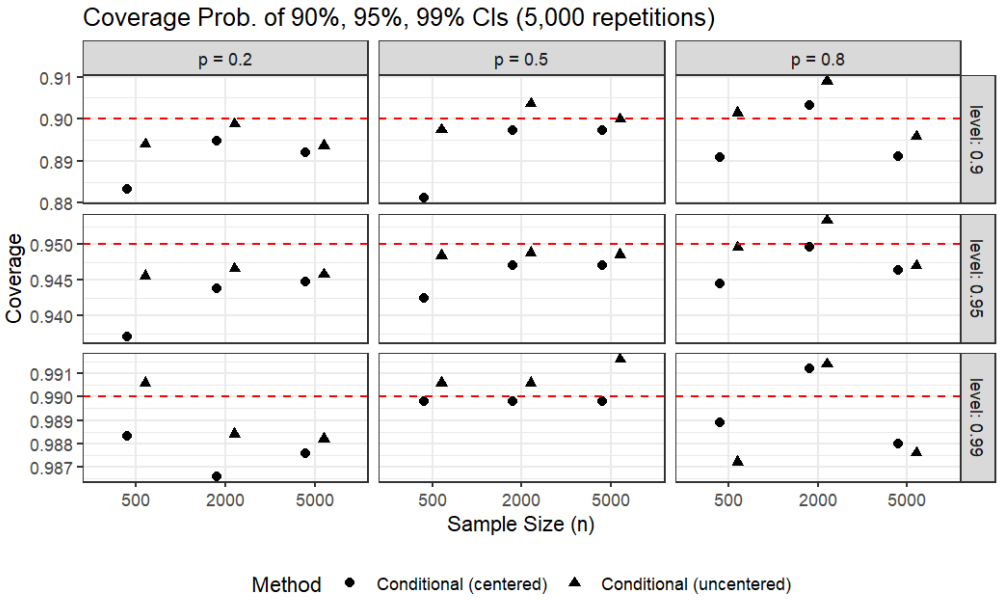


Figure 17: Empirical coverage probabilities of 90%, 95%, and 99% confidence bounds across sample sizes n and treatment probabilities p, using the true bias factors; red dashed lines indicate nominal coverage levels. For $n = 500$, the centered conditional yields 34 negative estimates of ν_{0s}^2 when $p = 0.2$ and approximately 46 when $p = 0.8$. We omit these cases when calculating the coverage.

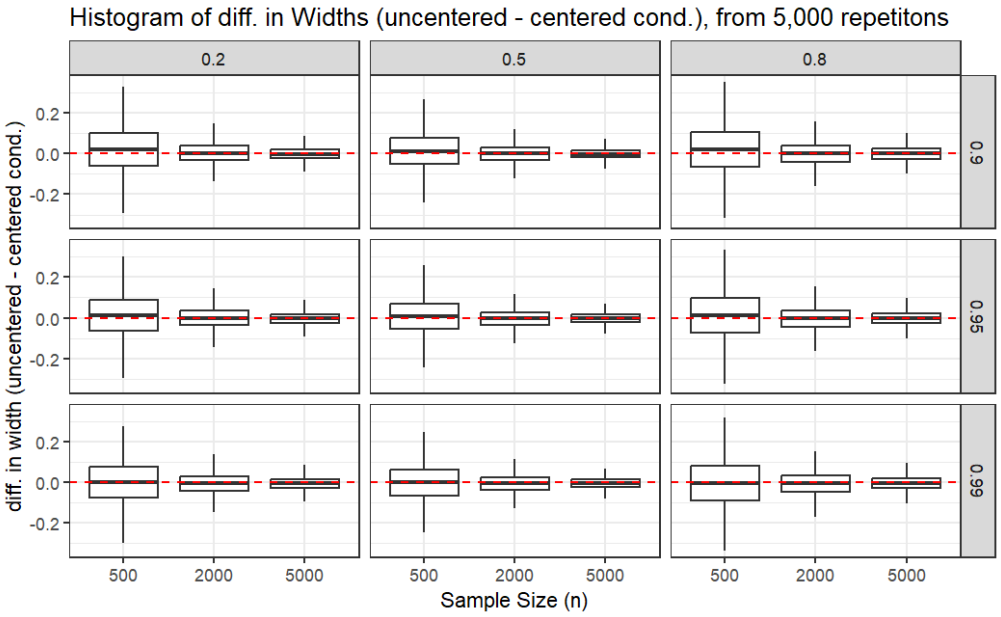
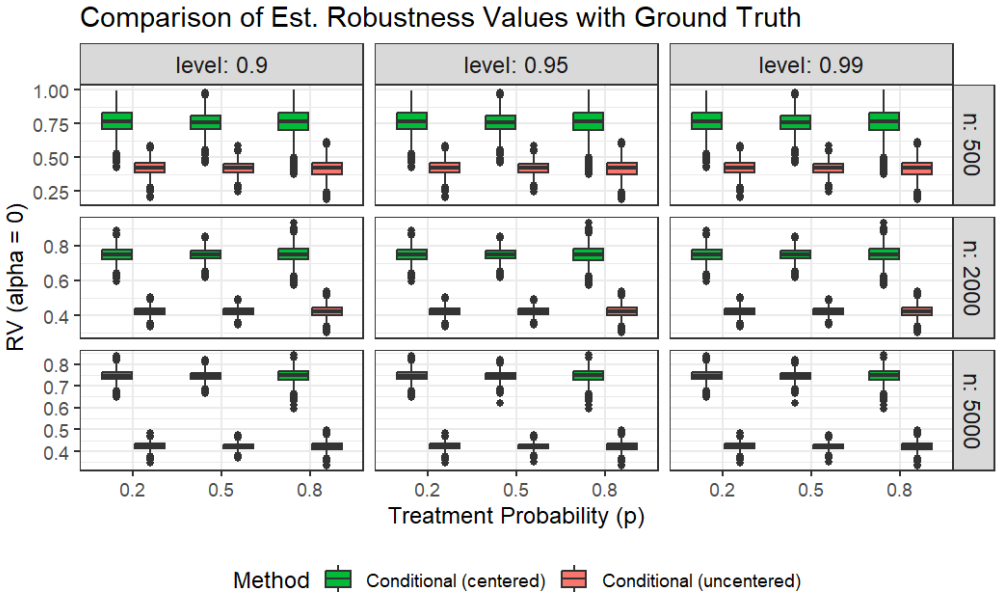
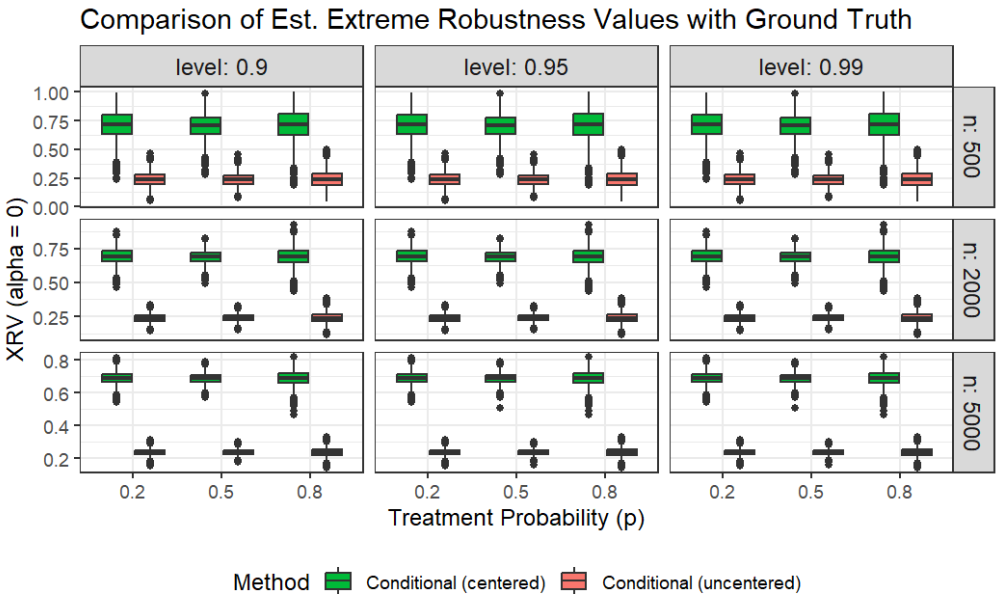


Figure 18: Boxplots of the difference in confidence-interval widths (uncentered conditional minus centered conditional) across sample sizes and treatment probabilities; the red dashed line denotes zero difference.

2. **Sensitivity statistics.**

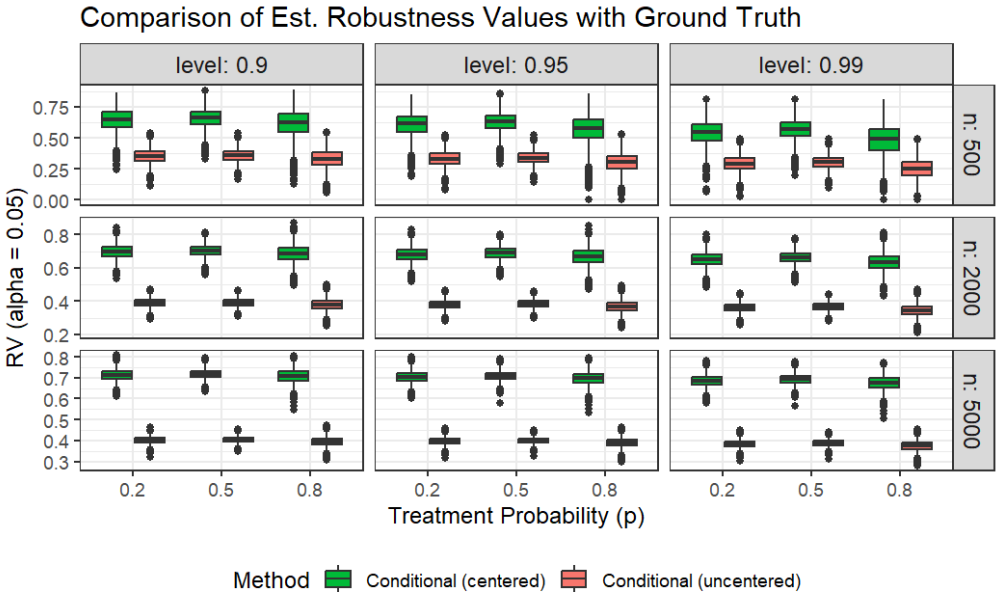


(a) $\text{RV}_{\theta^*=0,\alpha=0}$.

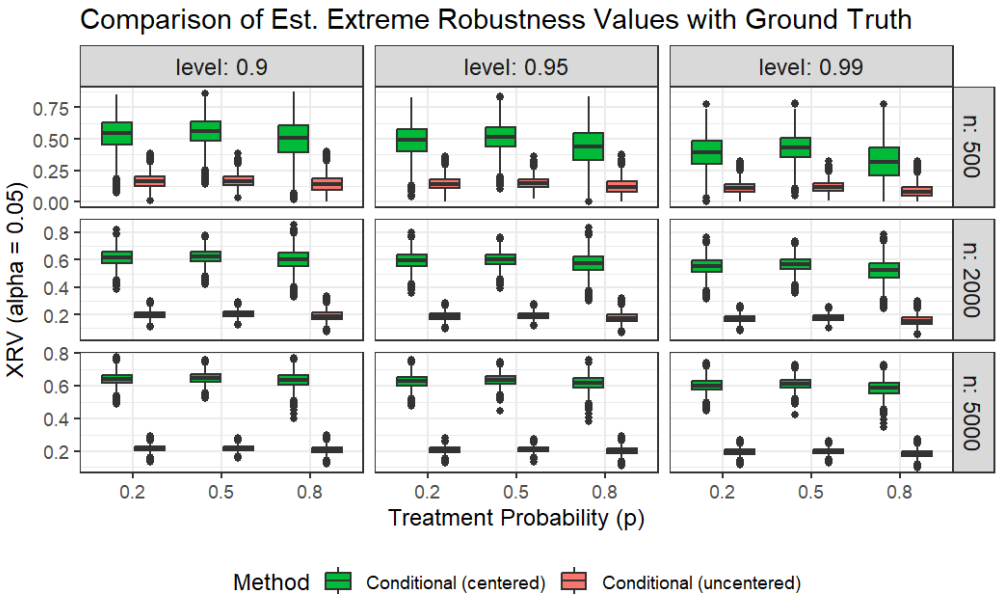


(b) $\text{XRV}_{\theta^*=0,\alpha=0}$.

Figure 19: Empirical estimates of $\text{RV}_{\theta^*=0,\alpha=0}$ and $\text{XRV}_{\theta^*=0,\alpha=0}$ across sample sizes and treatment probabilities (5,000 repetitions).



(a) $\text{RV}_{\theta^*=0,\alpha=0.05}$.



(b) $\text{XRV}_{\theta^*=0,\alpha=0.05}$.

Figure 20: Empirical estimates of $\text{RV}_{\theta^*=0,\alpha=0.05}$ and $\text{XRV}_{\theta^*=0,\alpha=0.05}$ across sample sizes and treatment probabilities (5,000 repetitions).

F.3 Relationship to pre-trend–based extrapolation approaches

This section compares pre-trend extrapolation (Rambachan and Roth, 2023) with the OVB formula. The main idea of the former is to extrapolate post-treatment deviations from parallel trends using

pre-treatment information. Under standard DiD identification assumptions, the following result provides the bridge.

Proposition 13 (Parallel trends violation as omitted variable bias)**.**

Under no anticipation and consistency,

$$ATT - \theta_s = -\delta, \text{ with } \delta := E[E[\Delta Y(0) \mid X, D=1] - E[\Delta Y(0) \mid X, D=0] \mid D=1],$$

where δ is the violation of the conditional parallel trends assumption, averaged over the X distribution of the treated group.

If, in addition, the conditional parallel trends assumption holds given (X, U), then

$$\underbrace{E\left[E[\Delta Y(0) \mid X, D=1] - E[\Delta Y(0) \mid X, D=0] \mid D=1\right]}_{=:\delta \text{ (deviation from parallel trends)}} = \underbrace{E[g_0(X,U) \mid D=1] - E[g_{0s}(X) \mid D=1]}_{=\theta_0 - \theta_{0s} \text{ (bias from omitting } U)}.$$

Proof. By the proof of Proposition 4, if no anticipation holds, we have

$$\text{ATT} := E\left[Y_2(1) - Y_2(0) | D=1\right] = E[\Delta Y(1) \mid D=1] - E[\Delta Y(0) \mid D=1].$$

Then, by consistency,

$$\begin{aligned}
\text{ATT} - \theta_s &= E[\Delta Y(1) \mid D=1] - E[\Delta Y(0) \mid D=1] - E[\Delta Y - g_{0s} \mid D=1] \\
&= E[\Delta Y \mid D=1] - E[\Delta Y(0) \mid D=1] - E[\Delta Y \mid D=1] + E[E[\Delta Y(0) \mid X, D=0] \mid D=1] \\
&= -E[E[\Delta Y(0) \mid X, D=1] \mid D=1] + E[E[\Delta Y(0) \mid X, D=0] \mid D=1] \text{ by LTE,} \\
&= -E\left[E[\Delta Y(0) \mid X, D=1] - E[\Delta Y(0) \mid X, D=0] \mid D=1\right] \\
&= -\delta.
\end{aligned}$$

This matches Example 2.2.3 of Rambachan and Roth (2023) where δ is the violation of conditional parallel trends, averaged over the X distribution of the treated group. As a special case, if one instead uses the unconditional untreated mean trend $E[\Delta Y \mid D=0]$ (in place of g_{0s}) as a proxy for the treated group's counterfactual trend, the deviation from parallel trends reduces to $\delta = E[\Delta Y(0) \mid D=1] - E[\Delta Y(0) \mid D=0]$, which matches Example 2.2.1 of Rambachan and Roth (2023).

Assume further that the parallel trends assumption additionally holds conditional on (X, U),

$$\begin{aligned}
\delta &:= E\left[E[\Delta Y(0) \mid X, D=1] - E[\Delta Y(0) \mid X, D=0] \mid D=1\right] \\
&= E[\Delta Y(0) \mid D=1] - E[g_{0s}(X) \mid D=1] \text{ by consistency,} \\
&= E[\Delta Y(0) \mid D=1] - E[g_0(X,U) \mid D=1] + E[g_0(X,U) \mid D=1] - E[g_{0s}(X) \mid D=1] \\
&= E[\Delta Y(0) \mid D=1] - E[g_0(X,U) \mid D=1] + (\theta_0 - \theta_{0s}) \\
&= E[E[\Delta Y(0) \mid X, U, D=1] \mid D=1] - E[g_0(X,U) \mid D=1] + (\theta_0 - \theta_{0s}) \text{ by LTE,}
\end{aligned}$$

$= E[E[\Delta Y(0) \mid X, U, D = 0] \mid D = 1] - E[g_0(X, U) \mid D = 1] + (\theta_0 - \theta_{0s})$ by conditional PTA,

$= E[g_0(X, U) \mid D = 1] - E[g_0(X, U) \mid D = 1] + (\theta_0 - \theta_{0s})$ by consistency,

$= \theta_0 - \theta_{0s}$.

□

Note the pre-trend extrapolation approach is a partial identification exercise where confounding restrictions are imposed using pre-treatment information. In contrast, we treat confounding strength as a sensitivity parameter and explicitly trace how identification and inference change as this strength varies. These two approaches are complementary. Pre-trend extrapolation can be used to motivate a range of plausible confounding strengths based on restrictions on deviations from parallel trends, while our results help interpret and translate a given confounding strength into the corresponding magnitude of deviations from parallel trends.

We now highlight two primary limitations of the extrapolation-based approach.

- **Not applicable to the canonical two-group, two-period DiD:** The extrapolation-based approach relies on pre-treatment information to infer post-treatment deviations from parallel trends. In contrast, our results do not require pre-trend information and can serve as a general building block for sensitivity analysis of the ATT in both two-period and multi-period designs (Callaway and Sant'Anna, 2021). In this way, we can see pre-trend extrapolation as one additional way of benchmarking confounding strength, as discussed in Section 4.

- **Sensitive to mismatches between pre- and post-treatment deviations:** Even when pre-treatment trend information is available, the extrapolation-based approach may not correctly restrict deviations from parallel trends in the post-treatment period. Pre-treatment trends may fail to detect post-treatment violations of parallel trends, and conversely, pre-treatment deviations may exist even when post-treatment parallel trends hold. Our approach provides additional ways to gauge the magnitude of the bias, such as, for example, comparing it with the bias induced by observed covariates. This provides a more complete picture of the robustness of the result, and also allows us to put the pre-trend bias itself into better context.

F.4 Deferred proofs

Proof of Proposition 7. We first construct some important relationships:

$$
\begin{aligned}
(1):\ R^2_{\alpha\sim\alpha_s} &:= \frac{\mathrm{Var}(\alpha_s)}{\mathrm{Var}(\alpha)} = \frac{E[\alpha_s^2]}{E[\alpha^2]} \\
&\overset{\text{LTE}}{=} \frac{E[E[\alpha_s^2 \mid D]]}{E[E[\alpha^2 \mid D]]} = \frac{pE[\alpha_s^2 \mid D=1] + (1-p)E[\alpha_s^2 \mid D=0]}{pE[\alpha^2 \mid D=1] + (1-p)E[\alpha^2 \mid D=0]} \\
&= \frac{\frac{1}{p} + \frac{1}{1-p} \times E[\alpha_{0s}^2 \mid D=0]}{\frac{1}{p} + \frac{1}{1-p} \times E[\alpha_0^2 \mid D=0]} \\
&= \frac{R^2_{O_{XU}\sim O|D=0} + O \times R^2_{O_{XU}\sim O_X|D=0}}{R^2_{O_{XU}\sim O|D=0} + O} \\
&= \frac{R^2_{O_{XU}\sim O|D=0}}{R^2_{O_{XU}\sim O|D=0} + O} \times 1 + \frac{O}{R^2_{O_{XU}\sim O|D=0} + O} \times R^2_{O_{XU}\sim O_X|D=0}, \\
(2):\ R^2_{\alpha\sim\alpha_s} &= \frac{\frac{1}{p} + \frac{1}{1-p} \times (E[\alpha_{0s}^2 \mid D=0] - 1) + \frac{1}{1-p}}{\frac{1}{p} + \frac{1}{1-p} \times (E[\alpha_0^2 \mid D=0] - 1) + \frac{1}{1-p}} \\
&= \frac{\frac{1}{p} + \mathrm{Var}(\alpha_{0s}|D=0)}{\frac{1}{p} + \mathrm{Var}(\alpha_0|D=0)} \\
&= \frac{\frac{1}{p} + \frac{(1-p)^2}{p^2}\mathrm{Var}(O_X \mid D=0)}{\frac{1}{p} + \frac{(1-p)^2}{p^2}\mathrm{Var}(O_{XU} \mid D=0)} \\
&= \frac{p + (1-p)^2\mathrm{Var}(O_X \mid D=0)}{p + (1-p)^2\mathrm{Var}(O_{XU} \mid D=0)} \\
(3):\ R^2_{\alpha\sim\alpha_s} &= \frac{\frac{1}{p} + \frac{1}{1-p} \times E[\alpha_{0s}^2 \mid D=0]}{\frac{1}{p} + \frac{1}{1-p} \times E[\alpha_0^2 \mid D=0]} \\
&\overset{\text{def}}{=} \frac{O + E[O_X^2 \mid D=0]}{O + E[O_{XU}^2 \mid D=0]} \\
&= \frac{O + \frac{1}{1-p}(E[O_X] - p)}{O + \frac{1}{1-p}(E[O_{XU}] - p)} \text{ by Proposition 2 (4)}, \\
&= \frac{E[O_X]}{E[O_{XU}]}.
\end{aligned}
$$

Accordingly,

$$
\begin{aligned}
\text{By (1): } C_D^2 &:= \frac{1 - R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}} \\
&= \frac{1 - R^2_{O_{XU}\sim O_X|D=0}}{\frac{1}{O}R^2_{O_{XU}\sim O|D=0} + R^2_{O_{XU}\sim O_X|D=0}}.
\end{aligned}
$$

$$
\text{Additionally, } 1 - R^2_{\alpha\sim\alpha_s} = \frac{1 - R^2_{O_{XU}\sim O_X|D=0}}{\frac{1}{O}R^2_{O_{XU}\sim O|D=0} + 1}.
$$

$$
\begin{aligned}
\text{By (2): } C_D^2 &:= \frac{1 - R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}} \\
&= \frac{1 - \frac{p+(1-p)^2\,\mathrm{Var}(O_X|D=0)}{p+(1-p)^2\,\mathrm{Var}(O_{XU}|D=0)}}{\frac{p+(1-p)^2\,\mathrm{Var}(O_X|D=0)}{p+(1-p)^2\,\mathrm{Var}(O_{XU}|D=0)}} \\
&= \frac{\mathrm{Var}(O_{XU} \mid D=0) - \mathrm{Var}(O_X \mid D=0)}{\frac{1}{p}\times O^2 + \mathrm{Var}(O_X \mid D=0)} \\
&= \frac{1 - R^2_{O_{XU}\sim O_X|1,D=0}}{\frac{1}{p\times CV^2_{O_{XU}}} + R^2_{O_{XU}\sim O_X|1,D=0}}.
\end{aligned}
$$

$$
\text{Additionally, } 1 - R^2_{\alpha\sim\alpha_s} = \frac{1 - R^2_{O_{XU}\sim O_X|1,D=0}}{\frac{1}{p\times CV^2_{O_{XU}}} + 1}.
$$

$$
\begin{aligned}
\text{By (3): } C_D^2 &:= \frac{1 - R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}} \\
&= \frac{1 - \frac{E[O_X]}{E[O_{XU}]}}{\frac{E[O_X]}{E[O_{XU}]}} = \frac{E[O_{XU}] - E[O_X]}{E[O_X]} \\
&= \frac{E[O_{XU} \mid D=1] - E[O_X \mid D=1]}{E[O_X \mid D=1] + 1} \text{ by Proposition 2 (4),} \\
&= \frac{\chi^2(P_{X,U|1}\|P_{X,U|0}) - \chi^2(P_{X|1}\|P_{X|0})}{\chi^2(P_{X|1}\|P_{X|0}) + \frac{1}{p}} \text{ by Proposition 3 (2).}
\end{aligned}
$$

$$
\text{Additionally, } 1 - R^2_{\alpha\sim\alpha_s} = \frac{\chi^2(P_{X,U|1}\|P_{X,U|0}) - \chi^2(P_{X|1}\|P_{X|0})}{\chi^2(P_{X,U|1}\|P_{X,U|0}) + \frac{1}{p}}.
$$

□

Proof of Proposition 8.

(i) Define $g_d := E[\Delta Y \mid X, U, D=d]$ and $g_{ds} := E[\Delta Y \mid X, D=d]$. Note that $g_d(X,U) = g(D=d, X, U)$ and $g_{ds}(X) = g_s(D=d, X)$, for $d \in \{0,1\}$.

$$
\begin{aligned}
C^2_{\Delta Y} &:= \frac{\mathrm{Var}(g - g_s)}{\mathrm{Var}(\Delta Y - g_s)} \\
&\overset{\text{LTV}}{=} \frac{E\left[\mathrm{Var}(g - g_s \mid D)\right] + \mathrm{Var}(E[g - g_s \mid D])}{E\left[\mathrm{Var}(\Delta Y - g_s \mid D)\right] + \mathrm{Var}(E[\Delta Y - g_s \mid D])} \\
&\overset{\text{LTE}}{=} \frac{E\left[\mathrm{Var}(g - g_s \mid D)\right]}{E\left[\mathrm{Var}(\Delta Y - g_s \mid D)\right]} \\
&= \frac{p \times \mathrm{Var}(g - g_s \mid D=1) + (1-p) \times \mathrm{Var}(g - g_s \mid D=0)}{p \times \mathrm{Var}(\Delta Y - g_s \mid D=1) + (1-p) \times \mathrm{Var}(\Delta Y - g_s \mid D=0)} \\
&= \frac{p \times \mathrm{Var}(g_1 - g_{1s} \mid D=1) + (1-p) \times \mathrm{Var}(g_0 - g_{0s} \mid D=0)}{p \times \mathrm{Var}(\Delta Y - g_{1s} \mid D=1) + (1-p) \times \mathrm{Var}(\Delta Y - g_{0s} \mid D=0)} \\
&= \frac{p \times \mathrm{Var}(g_1 - g_{1s} \mid D=1)}{p \times \sigma^2_{1s} + (1-p) \times \sigma^2_{0s}} + \frac{(1-p) \times \mathrm{Var}(g_0 - g_{0s} \mid D=0)}{p \times \sigma^2_{1s} + (1-p) \times \sigma^2_{0s}}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{p \times \text{Var}(\Delta Y - g_{1s} \mid D=1)}{p \times \sigma_{1s}^2 + (1-p) \times \sigma_{0s}^2} \times \frac{\text{Var}(g_1 - g_{1s} \mid D=1)}{\text{Var}(\Delta Y - g_{1s} \mid D=1)} + \\
&\qquad \frac{(1-p) \times \text{Var}(g_0 - g_{0s} \mid D=0)}{p \times \sigma_{1s}^2 + (1-p) \times \sigma_{0s}^2} \times \frac{\text{Var}(g_0 - g_{0s} \mid D=0)}{\text{Var}(\Delta Y - g_{0s} \mid D=0)} \\
&= \frac{p \times \sigma_{1s}^2}{p \times \sigma_{1s}^2 + (1-p) \times \sigma_{0s}^2} \times C_{1\Delta Y}^2 + \frac{(1-p) \times \sigma_{0s}^2}{p \times \sigma_{1s}^2 + (1-p) \times \sigma_{0s}^2} \times C_{0\Delta Y}^2 \\
&= W_{0\Delta Y} C_{0\Delta Y}^2 + (1 - W_{0\Delta Y}) C_{1\Delta Y}^2.
\end{aligned}
$$

The 6th and 8th equalities are established by observing that for $d \in \{0, 1\}$,

$$
\begin{aligned}
\text{Var}(\Delta Y - g_{ds} \mid D=d) &\overset{\text{LTV}}{=} E[\text{Var}(\Delta Y \mid X, D=d) \mid D=d] + \text{Var}(E[\Delta Y - g_{ds} \mid X, D=d] \mid D=d) \\
&\overset{\text{LTE}}{=} E[\text{Var}(\Delta Y \mid X, D=d) \mid D=d] \\
&\overset{\text{def.}}{=} \sigma_{ds}^2.
\end{aligned}
$$

(ii)

$$
\begin{aligned}
C_D^2 &:= \frac{1 - R_{\alpha \sim \alpha_s}^2}{R_{\alpha \sim \alpha_s}^2} = \frac{E[O_{XU}] - E[O_X]}{E[O_X]} \\
&= \frac{p \times (E[O_{XU} \mid D=1] - E[O_X \mid D=1])}{p \times (E[O_X \mid D=1] + 1)} \text{ by Proposition 2 (4)}, \\
&= \frac{E[O_{XU} \mid D=1] - E[O_X \mid D=1]}{E[O_X \mid D=1] + 1} \\
&= \frac{E[O_{XU} \mid D=1] - E[O_X \mid D=1]}{E[O_X \mid D=1]} \times \frac{E[O_X \mid D=1]}{E[O_X \mid D=1] + 1} \\
&\overset{\text{def.}}{=} C_{0D}^2 \times \frac{E[O_X \mid D=1]}{E[O_X \mid D=1] + 1} \\
&= C_{0D}^2 \times \frac{O \times \nu_{0s}^2}{O \times \nu_{0s}^2 + 1},
\end{aligned}
$$

where the last equality uses $\nu_{0s}^2 = \frac{E[O_X|D=1]}{O}$, which was established in Corollary 1. Furthermore, to better separate the confounding strength of interest from the sampling mechanism, we use $\nu_{0s}^2 = \chi^2(P_{X|1} \| P_{X|0}) + 1$, which was established in Corollary 1. Then, the following relationship holds:

$$
C_D^2 = C_{0D}^2 \times \frac{O \times (\chi^2(P_{X|1} \| P_{X|0}) + 1)}{O \times (\chi^2(P_{X|1} \| P_{X|0}) + 1) + 1}.
$$

(iii) We begin by constructing several key relationships between the covariances in the conditional and unconditional results.

$$
\begin{aligned}
\text{Cov}(g - g_s, \alpha - \alpha_s) &= E[(g - g_s)(\alpha - \alpha_s)] \overset{\text{LTE}}{=} E\left[E[(g - g_s)(\alpha - \alpha_s) \mid D]\right] \\
&= pE[(g - g_s)(\alpha - \alpha_s) \mid D=1] + (1-p)E[(g - g_s)(\alpha - \alpha_s) \mid D=0] \\
&= (1-p)E\left[(g_0 - g_{0s}) \times (-O_{XU} + O_X) \times \frac{1}{p} \mid D=0\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{O} \times E\left[(g_0 - g_{0s})\left(-O_{XU} + O_X\right) \mid D = 0\right] \\
&= -E[(g_0 - g_{0s})(\alpha_0 - \alpha_{0s}) \mid D = 0] \\
&= -\operatorname{Cov}(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D = 0), \\
\operatorname{Var}(g - g_s) &\overset{\text{LTV}}{=} E[\operatorname{Var}(g - g_s \mid D)] + \operatorname{Var}(E[g - g_s \mid D]) \\
&\overset{\text{LTE}}{=} E[\operatorname{Var}(g - g_s \mid D)] \\
&= \operatorname{Var}(g - g_s \mid D = 1)p + \operatorname{Var}(g - g_s \mid D = 0)(1 - p) \\
&= \operatorname{Var}(g_1 - g_{1s} \mid D = 1)p + \operatorname{Var}(g_0 - g_{0s} \mid D = 0)(1 - p), \\
\operatorname{Var}(\alpha - \alpha_s) &\overset{\text{LTV}}{=} E[\operatorname{Var}(\alpha - \alpha_s \mid D)] + \operatorname{Var}(E[\alpha - \alpha_s \mid D]) \\
&\overset{\text{LTE}}{=} E[\operatorname{Var}(\alpha - \alpha_s \mid D)] \\
&= 0 \times p + \operatorname{Var}(\alpha - \alpha_s \mid D = 0) \times (1 - p) \\
&= \operatorname{Var}\left(\frac{-O_{XU} + O_X}{p} \mid D = 0\right) \times (1 - p) \\
&= \operatorname{Var}\left(\frac{O_{XU} - O_X}{O} \mid D = 0\right) \times \frac{1}{1 - p} \\
&\overset{\text{def.}}{=} \frac{1}{1 - p} \operatorname{Var}(\alpha_0 - \alpha_{0s} \mid D = 0),
\end{aligned}
$$

Accordingly, the following holds:

$$
\begin{aligned}
\rho^2 &:= \frac{\operatorname{Cov}^2(g - g_s, \alpha - \alpha_s)}{\operatorname{Var}(g - g_s)\operatorname{Var}(\alpha - \alpha_s)} \\
&= \frac{\operatorname{Cov}^2(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D = 0)}{(\operatorname{Var}(g_1 - g_{1s} \mid D = 1)p + \operatorname{Var}(g_0 - g_{0s} \mid D = 0)(1 - p)) \times \left(\frac{1}{1-p} \operatorname{Var}(\alpha_0 - \alpha_{0s} \mid D = 0)\right)} \\
&\overset{\text{def.}}{=} \frac{\rho_0^2}{\frac{\operatorname{Var}(g_1 - g_{1s} \mid D=1)}{\operatorname{Var}(g_0 - g_{0s} \mid D=0)} \times O + 1} \left(= \frac{\rho_0^2}{\frac{C_{1\Delta Y}^2 \sigma_{1s}^2}{C_{0\Delta Y}^2 \sigma_{0s}^2} \times O + 1}\right).
\end{aligned}
$$

We implicitly assume that $\operatorname{Var}(g_0 - g_{0s} \mid D = 0), \operatorname{Var}(\alpha_0 - \alpha_{0s} \mid D = 0) > 0$; otherwise, the bias is zero.

(iv) **Supplement:** Finally, combining the results above, we show that the unconditional result reduces to the conditional result. We begin by deriving two important relationships for the scaling factors:

$$
\begin{aligned}
\text{(1):}\ \operatorname{Var}(\Delta Y - g_s) &= E[(\Delta Y - g_s)^2] \overset{\text{LTE}}{=} E\left[E[(\Delta Y - g_s)^2 \mid D]\right] \\
&= pE[(\Delta Y - g_{1s})^2 \mid D = 1] + (1 - p)E[(\Delta Y - g_{0s})^2 \mid D = 0] \\
&= p \operatorname{Var}(\Delta Y - g_{1s} \mid D = 1) + (1 - p) \operatorname{Var}(\Delta Y - g_{0s} \mid D = 0),
\end{aligned}
$$

$$
\begin{aligned}
\text{(2):}\ \operatorname{Var}(\alpha_s) = E[\alpha_s^2] &\overset{\text{LTE}}{=} E\left[E[\alpha_s^2 \mid D]\right] \\
&= p \times E[\alpha_s^2 \mid D=1] + (1-p) \times E[\alpha_s^2 \mid D=0] \\
&\overset{\text{def.}}{=} \frac{1}{p} + \frac{1}{1-p} \times E[\alpha_{0s}^2 \mid D=0].
\end{aligned}
$$

The second equality implies that $\operatorname{Var}(\alpha_s) = E[\alpha_s^2] > E[\alpha_{0s}^2 \mid D=0]$, with the gap increasing as p approaches zero or one. Accordingly, we have the following:

$$
\begin{aligned}
&\rho^2 C_{\Delta Y}^2 C_D^2 S^2 \\
&= \frac{\rho_0^2}{\frac{\operatorname{Var}(g_1 - g_{1s} \mid D=1) O}{\operatorname{Var}(g_0 - g_{0s} \mid D=0)} + 1} \times \left((1 - W_{0\Delta Y}) C_{1\Delta Y}^2 + W_{0\Delta Y} C_{0\Delta Y}^2\right) \times \left(W_{0D} C_{0D}^2\right) \times \operatorname{Var}(\Delta Y - g_s) \operatorname{Var}(\alpha_s) \\
&= \rho_0^2 \times \operatorname{Var}(g_0 - g_{0s} \mid D=0) \times \frac{O \times \operatorname{Var}(g_1 - g_{1s} \mid D=1) + \sigma_{0s}^2 C_{0\Delta Y}^2}{O \times \operatorname{Var}(g_1 - g_{1s} \mid D=1) + \operatorname{Var}(g_0 - g_{0s} \mid D=0)} \times C_{0D}^2 \times \nu_{0s}^2 \\
&= \rho_0^2 \times \sigma_{0s}^2 C_{0\Delta Y}^2 \times \frac{O \operatorname{Var}(g_1 - g_{1s} \mid D=1) + \sigma_{0s}^2 C_{0\Delta Y}^2}{O \operatorname{Var}(g_1 - g_{1s} \mid D=1) + \sigma_{0s}^2 C_{0\Delta Y}^2} \times C_{0D}^2 \times \nu_{0s}^2 \\
&= \rho_0^2 C_{0\Delta Y}^2 C_{0D}^2 \sigma_{0s}^2 \nu_{0s}^2 = \rho_0^2 C_{0\Delta Y}^2 C_{0D}^2 S_0^2,
\end{aligned}
$$

where the first equality follows from (i),(ii),(iii), the second equality follows from (1) and (2), and the third equality uses $\operatorname{Var}(g_0 - g_{0s} \mid D=0) = \sigma_{0s}^2 C_{0\Delta Y}^2$.

An alternative way to establish this equivalence is to use $\operatorname{Cov}(g - g_s, \alpha - \alpha_s) = -\operatorname{Cov}(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D=0)$ directly, which is shown as part of the proof of (iii).

□

Proof of Proposition 10.

(i) Note that

$$
\begin{aligned}
R_{\alpha_{0s} \sim 1 \mid D=0}^2 &= 1 - \frac{E[(\alpha_{0s} - 1)^2 \mid D=0]}{E[\alpha_{0s}^2 \mid D=0]} \\
&= \frac{1}{E[\alpha_{0s}^2 \mid D=0]} = \frac{O}{E[O_X \mid D=1]}.
\end{aligned}
$$

Accordingly, we have

$$
\begin{aligned}
C_{0wD}^2 := \frac{1 - R_{\alpha_0 \sim \alpha_{0s} \mid 1, D=0}^2}{R_{\alpha_0 \sim \alpha_{0s} \mid 1, D=0}^2} &= \frac{\operatorname{Var}(\alpha_0 \mid D=0) - \operatorname{Var}(\alpha_{0s} \mid D=0)}{\operatorname{Var}(\alpha_{0s} \mid D=0)} \\
&= \frac{E[\alpha_0^2 \mid D=0] - E[\alpha_{0s}^2 \mid D=0]}{E[\alpha_{0s}^2 \mid D=0] - 1} \\
&= \frac{1 - R_{\alpha_0 \sim \alpha_{0s} \mid D=0}^2}{R_{\alpha_0 \sim \alpha_{0s} \mid D=0}^2 - \frac{1}{E[\alpha_0^2 \mid D=0]}} \\
&\overset{\text{def.}}{=} C_{0D}^2 \times \frac{R_{\alpha_0 \sim \alpha_{0s} \mid D=0}^2}{R_{\alpha_0 \sim \alpha_{0s} \mid D=0}^2 - \frac{1}{E[\alpha_0^2 \mid D=0]}}
\end{aligned}
$$

$$= C_{0D}^2 \times \frac{\frac{E[O_X|D=1]}{E[O_{XU}|D=1]}}{\frac{E[O_X|D=1]}{E[O_{XU}|D=1]} - \frac{O}{E[O_{XU}|D=1]}}$$

$$= C_{0D}^2 \times \frac{E[O_X \mid D=1]}{E[O_X \mid D=1] - O} \left(= C_{0D}^2 \times \frac{1}{1 - R^2_{\alpha_{0s} \sim 1 | D=0}} \right).$$

These equalities follow from the results established in the proof of Theorem 1.

(ii) We begin by constructing several important relationships:

(1): $\text{Cov}(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D=0) = E[(g_0 - g_{0s})(\alpha_0 - \alpha_{0s}) \mid D=0]$

$= E[g_0(X,U)\alpha_0(X,U) \mid D=0] - E[g_{0s}(X)\alpha_{0s}(X) \mid D=0]$

$= E[E[\Delta Y \mid X, U, D=0]\,\alpha_0(X,U) \mid D=0] - E[E[\Delta Y \mid X, D=0]\,\alpha_{0s}(X) \mid D=0]$

$= E[E[\alpha_0(X,U)\Delta Y \mid X, U, D=0] \mid D=0] - E[E[\alpha_{0s}(X)\Delta Y \mid X, D=0] \mid D=0]$

$= E[\alpha_0(X,U)\Delta Y \mid D=0] - E[\alpha_{0s}(X)\Delta Y \mid D=0]$

$= E[(\alpha_0 - \alpha_{0s})\Delta Y \mid D=0] - \underbrace{E[(\alpha_0 - \alpha_{0s}) \mid D=0]E[\Delta Y \mid D=0]}_{=0}$

$= \text{Cov}((\alpha_0 - \alpha_{0s}), \Delta Y \mid D=0),$

where the second equality follows from the proof of Theorem 1. Moreover, we showed in the proof of Theorem 1 that $\sigma_{0s}^2 = \text{Var}(\Delta Y - g_{0s} \mid D=0) = E[\text{Var}(\Delta Y \mid X, D=0) \mid D=0]$, so we have

(2): $\sigma_{w0s}^2 := \text{Var}(\Delta Y \mid D=0)$

$\overset{\text{LTV}}{=} E[\text{Var}(\Delta Y \mid X, D=0) \mid D=0] + \text{Var}(E[\Delta Y \mid X, D=0] \mid D=0)$

$= \sigma_{0s}^2 + \text{Var}(g_{0s} \mid D=0),$

which suggests that $\sigma_{w0s}^2 \geq \sigma_{0s}^2$. Additionally, we have

$$R^2_{g_0 \sim g_{0s}|1,D=0} := 1 - \frac{\text{Var}(g_0 - g_{0s} \mid D=0)}{\text{Var}(g_0 \mid D=0)}$$

$$= \frac{-\text{Var}(g_{0s} \mid D=0) + 2(E[g_0 g_{0s} \mid D=0] - E[g_0 \mid D=0]E[g_{0s} \mid D=0])}{\text{Var}(g_0 \mid D=0)}$$

$$\overset{\text{LTE}}{=} \frac{\text{Var}(g_{0s} \mid D=0)}{\text{Var}(g_0 \mid D=0)}.$$

Then, we define the corresponding Cohen's partial f^2 as:

$$f^2_{g_0 \sim g_{0s}|1,D=0} := \frac{R^2_{g_0 \sim g_{0s}|1,D=0}}{1 - R^2_{g_0 \sim g_{0s}|1,D=0}}$$

$$= \frac{\text{Var}(g_{0s} \mid D=0)}{\text{Var}(g_0 \mid D=0) - \text{Var}(g_{0s} \mid D=0)}.$$

Similarly, we have that

$$\begin{aligned} R^2_{\Delta Y \sim g_{0s}|1,D=0} &:= 1 - \frac{\text{Var}(\Delta Y - g_{0s} \mid D=0)}{\text{Var}(\Delta Y \mid D=0)} \\ &= \frac{-\text{Var}(g_{0s} \mid D=0) + 2(E[\Delta Y g_{0s} \mid D=0] - E[\Delta Y \mid D=0]E[g_{0s} \mid D=0])}{\text{Var}(\Delta Y \mid D=0)} \\ &\overset{\text{LTE}}{=} \frac{\text{Var}(g_{0s} \mid D=0)}{\text{Var}(\Delta Y \mid D=0)}. \end{aligned}$$

Accordingly, we have the following:

$$\begin{aligned} \rho^2_{w0} &= \frac{\text{Cov}^2(\Delta Y, \alpha_0 - \alpha_{0s} \mid D=0)}{\text{Var}(\Delta Y \mid D=0)\,\text{Var}(\alpha_0 - \alpha_{0s} \mid D=0)} \\ &\overset{(1)}{=} \frac{\text{Cov}^2(g_0 - g_{0s}, \alpha_0 - \alpha_{0s} \mid D=0)}{\text{Var}(g_0 - g_{0s} \mid D=0)\,\text{Var}(\alpha_0 - \alpha_{0s} \mid D=0)} \times \frac{\text{Var}(g_0 - g_{0s} \mid D=0)}{\text{Var}(\Delta Y \mid D=0)} \\ &\overset{(2)}{=} \rho^2_0 \times \frac{\text{Var}(g_0 - g_{0s} \mid D=0)}{\sigma^2_{0s} + \text{Var}(g_{0s} \mid D=0)} \\ &\overset{\text{def.}}{=} \rho^2_0 \times \frac{1}{\frac{1}{C^2_{0\Delta Y}} + \frac{\text{Var}(g_{0s}|D=0)}{\text{Var}(g_0 - g_{0s}|D=0)}} = \rho^2_0 \times \frac{1}{\frac{1}{C^2_{0\Delta Y}} + \frac{\text{Var}(g_{0s}|D=0)}{\text{Var}(g_0|D=0) - \text{Var}(g_{0s}|D=0)}} \\ &= \rho^2_0 \times \frac{1}{\frac{1}{C^2_{0\Delta Y}} + f^2_{g_0 \sim g_{0s}|1,D=0}} \left(= \rho^2_0 \times \frac{1}{\frac{1}{C^2_{0\Delta Y}} + \frac{1}{1 - R^2_{g_0 \sim g_{0s}|1,D=0}} - 1} \right). \end{aligned}$$

Note that $\frac{1}{C^2_{0\Delta Y}} + \frac{1}{1 - R^2_{g_0 \sim g_{0s}|1,D=0}} - 1 \geq 1$, so we have $\rho^2_{w0} \leq \rho^2_0$. Alternatively, we have that

$$\begin{aligned} \rho^2_{w0} &\overset{(1)}{=} \rho^2_0 \times \frac{\text{Var}(g_0 - g_{0s} \mid D=0)}{\text{Var}(\Delta Y - g_{0s} \mid D=0)} \times \frac{\text{Var}(\Delta Y - g_{0s} \mid D=0)}{\text{Var}(\Delta Y \mid D=0)} \\ &\overset{\text{def.}}{=} \rho^2_0 \times C^2_{0\Delta Y} \times (1 - R^2_{\Delta Y \sim g_{0s}|1,D=0}). \end{aligned}$$

□

G Additional results for the empirical application

Our empirical application studies the effect of the minimum wage on teen employment. This section presents additional results from our main analysis (which uses the never-treated comparison group), as well as results from the not-yet-treated comparison group.

G.1 Nuisance learners: specification and RMSE

Table 7 presents the specifications of the machine learning methods used to estimate the first-stage nuisance functions. We consider four learners—a parametric model, ridge regression, lasso, and random forest—using the covariates from Callaway and Sant'Anna (2021). Table 8 reports the

corresponding cross-fitted root mean squared errors (RMSE) for predicting D and ΔY. The random forest attains the lowest RMSE for both, and we therefore use it as the learner for our main analysis.

Learner	**Variables**	**Package**	**Tuning grid**
Parametric: Linear for g_{0s} Logistic for π	*region, white, hs, pov,* *lpop, lpop*^{2} *lmedinc, lmedinc*^{2}	`stats` (`lm`) `stats` (`glm`)	N/A
Ridge: Linear for g_{0s} Logistic for π	*region*; cubic polynomials in *lpop, white, pov, hs, lmedinc*; *region*-polynomial interactions	`glmnet`	$\alpha_{\text{EN}} = 0$ (ridge) λ selected from `cv.glmnet`
Lasso: Linear for g_{0s} Logistic for π	*region*; cubic polynomials in *lpop, white, pov, hs, lmedinc*; *region*-polynomial interactions	`glmnet`	$\alpha_{\text{EN}} = 1$ (lasso) λ selected from `cv.glmnet`
Random Forest (`num.trees` $= 1,000$)	*region, white, hs, pov,* *lpop, lmedinc*	`ranger`	`mtry` $\in \{2, 4, 6\}$ `min.node.size` $\in \{10, 15, 25, 50, 75, \ldots, 150\}$ `splitrule` $\in \{$`variance`, `extratrees`$\}$

Table 7: Machine learning methods used for first-stage nuisance estimation in the minimum wage example. Covariates enter linearly. Models are selected by minimizing the RMSE.

Methods	RMSE(D)	RMSE(ΔY)
Parametric	0.420	0.1548
Lasso	0.3803	0.1540
Ridge	0.3859	0.1539
Random Forest	**0.3719**	**0.1529**
Best	0.3719	0.1529

Table 8: Cross-fitted RMSEs for predicting ΔY and D in the minimum wage application. The random forest achieves the lowest RMSE for both the outcome evolution and the propensity score and is used as the learner in our main analysis.

G.2 Multiplier bootstrap for uniform confidence bands

Algorithm 2 presents the multiplier bootstrap procedure of Callaway and Sant'Anna (2021), which we use to construct the uniform 95% confidence bands shown in Figure 1. The procedure is com-

putationally fast and, by reweighting rather than resampling observations, guarantees that every bootstrap iteration retains units from both treated and control groups.

Algorithm 2 Multiplier Bootstrap: MBoot($\boldsymbol{\varphi}^0_{\boldsymbol{\theta}_s}$)

Input: An $n \times T$ influence-function matrix $\boldsymbol{\varphi}^0_{\boldsymbol{\theta}_s}$, with columns corresponding to time periods ($T = 1$ in the canonical setup). Significance level α, and number of bootstrap draws B.

for $b = 1, \ldots, B$ **do**

Draw weights $\{V_i\}_{i=1}^n$ from Mammen's distribution: let $\kappa = (\sqrt{5}+1)/2$,

$$V_i \overset{\text{i.i.d.}}{\sim} \begin{cases} 1-\kappa & \text{with probability } \kappa/\sqrt{5}, \\ \kappa & \text{with probability } 1-\kappa/\sqrt{5}. \end{cases}$$

Compute $\widehat{R}^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(V_i \times \boldsymbol{\varphi}^0_{\boldsymbol{\theta}_s}(Z_i)\right)$. Denote the t-th element by $\widehat{R}^*(t)$ for $t = 1, \ldots, T$.

end for

Estimate bootstrap standard deviations: $\widehat{\sigma}_{\boldsymbol{\varphi}^0_{\boldsymbol{\theta}_s,t}} = \text{IQR}(\widehat{R}^*(t)) / \left(\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\right)$.

For each bootstrap draw compute t-test$_b = \max_t \left(\left| \widehat{R}^*(t) \right| / \widehat{\sigma}_{\boldsymbol{\varphi}^0_{\boldsymbol{\theta}_s,t}} \right)$.

Calculate the critical value as $\hat{c}_{1-\alpha} = \text{Quantile}_{1-\alpha}(\{\text{t-test}_b\}_{b=1}^B)$.

Return: Construct confidence band for $\boldsymbol{\theta}_s(t)$ as $\widehat{\text{CI}}_{1-\alpha}(t) = \left[\widehat{\boldsymbol{\theta}_s}(t) \pm \left(\hat{c}_{1-\alpha} \widehat{\sigma}_{\boldsymbol{\varphi}^0_{\boldsymbol{\theta}_s,t}} \Big/ \sqrt{n}\right)\right]$.

G.3 Additional sensitivity results

Table 9 reports additional empirical benchmarking results for the minimum wage application. The gain metrics $G_{0\Delta Y,j}$ and $G_{0D,j}$, defined in Section C.1.5, measure the explanatory power of each observed covariate for the outcome evolution and the treatment indicator, respectively; these are the gain metrics used in the contour plots in the main text. The correlations $\rho_{0,j}$, defined in Section C.1.4 for benchmarking ρ_0, are all bounded above by 0.3 in absolute value.

	Gain Metrics		Correlation	Change in estimate
Observed covariate	$G_{0\Delta Y,j}$	$G_{0D,j}$	$\rho_{0,j}$	$\theta_s - \theta_{s,-j}$
lmedinc	0.0051	0.3701	-0.2296	0.0036
region	0.0049	0.7389	0.1327	-0.0025
white	0.0021	0.5190	-0.0544	0.0006
pov	0.0047	0.2392	0.0834	-0.0011

Table 9: Explanatory power of observed covariates in the minimum wage example. All estimates are debiased and cross-fitted.