Potential weights and implicit causal designs in linear regression

Jiafeng Chen
Stanford University and SIEPR
jiafeng@stanford.edu

ABSTRACT. When we interpret linear regression as estimating causal effects justified by quasi-experimental treatment variation, what do we mean? This paper characterizes the necessary implications when researchers ascribe design-based interpretation to coefficients in a given regression. We find that a plausible design-based interpretation for a regression estimand necessarily implies linear restrictions on the true distribution of treatment. Solving these linear restrictions leads to a set of *implicit designs* that necessarily include the true design if the regression were to admit a design-based interpretation. These necessary implications lead to practical diagnostics that add transparency and robustness when design-based interpretation is invoked. They also serve as a framework that unifies and extends existing theoretical results, and they lead to new theoretical insights for widely used but less understood specifications.

Date: March 8, 2025. I thank Isaiah Andrews, Dmitry Arkhangelsky, Tim Armstrong, Kirill Borusyak, Ambarish Chattopadhyay, Avi Feller, Matthew Gentzkow, Peter Hull, Guido Imbens, Patrick Kline, Lihua Lei, Jimmy Lin, Yiqi Liu, José Luis Montiel Olea, Francesca Molinari, Chen Qiu, Ashesh Rambachan, David Ritzwoller, Brad Ross, Jonathan Roth, Yucheng Shang, Jesse Shapiro, Jann Spiess, Tymon Słocyński, Elie Tamer, Alex Torgovitsky, Winnie van Dijk, Davide Viviano, José Zubizarreta, and participants seminars and workshops at Berkeley, the Chinese University of Hong Kong, Cornell, Harvard, Stanford, UC Irvine, and SEA 2024 for helpful comments. I also thank Syed Ashal Ali for research assistance.

1. Introduction

Much of economic research, particularly in applied microeconomics, emphasizes causal identification from treatment variation. To defend causal interpretation of their estimates, researchers frequently appeal to as-if randomization of treatment and to judicious comparisons between units with different treatment statuses, rather than to functional-form restrictions on potential outcomes.[1] We call this type of identification argument *design-based*, for its emphasis on modeling the treatment assignment process (often called a *design* in experimental settings)—usually in the hopes of emulating some randomized control trial. Design-based identification has certain virtues: For instance, it allows causal identification for different outcomes subjected to the same intervention, without modeling each individually.

If one defends identification on design-based grounds, it seems natural that estimation should model the process of treatment assignment and target an estimand defined in terms of unrestricted potential outcomes. Indeed, many methods proposed in the causal inference literature estimate the propensity score (i.e., distribution of treatment) and target estimands like the average treatment effect (Zubizarreta *et al.*, 2014; Abadie and Imbens, 2011; Bang and Robins, 2005; Ben-Michael *et al.*, 2021; Li *et al.*, 2018).

However, rather than these more "design-native" methods, economic research overwhelmingly prefers linear regressions for estimation (Imbens, 2015). Linear regression neither explicitly models treatment assignment nor specifies a nonparametric estimand. Instead, regression resembles a model on the functional form of potential outcomes. Indeed, Freedman (2008) criticizes regressions in experimental settings precisely because treatment randomization does not imply the linear model.

To address this tension, a modern perspective views linear regressions as simply estimators to be analyzed under certain designs. In this view, linear regression is essentially a machine that compares different groups of units. Under the right design, these comparisons recover causal effects—often weighted average treatment effects with weights implicitly chosen by the regression specification. In this sense, linear regression can be thought to

[1] By May 2024, over 50% of papers in the NBER working paper series mention experimental or quasi-experimental methods (Goldsmith-Pinkham, 2024; Currie *et al.*, 2020).
We refer to "design-based" as a focus on modeling treatment assignment, in contrast to modeling potential outcomes, which is often called "model-based." This is consistent with the terminology adopted in Borusyak *et al.* (2024a), Borusyak and Hull (2024), Kolesár (2024), and Angrist and Pischke (2010). In particular, Angrist and Pischke (2010) observe, "In applied micro fields [...] researchers seek real experiments where feasible, and useful natural experiments if real experiments seem (at least for a time) infeasible. In either case, a hallmark of contemporary applied microeconometrics is a framework that highlights specific sources of variation. These studies can be said to be design based in that they give the research design underlying any sort of study the attention it would command in a real experiment."

implicitly model the treatment assignment process and implicitly choose a target parameter.

Much has been written on the design-based properties of regression[2] This literature examines whether particular specifications make sensible comparisons under a known set of designs. It often provides *sufficient conditions* for design-based interpretation of linear estimands—or cautionary tales for the lack thereof. A limitation is that this literature produces a hodge-podge of results: Certain specifications have design-based interpretations under certain restrictions, but other configurations of specification and design may produce nonsensical estimands (e.g., negatively weighted causal effects).

Complementary to this literature, this paper instead asks what are the *necessary implications* when researchers ascribe a design-based interpretation to a particular regression specification. Suppose a researcher presents a regression and argues that certain coefficients estimate *some* causal comparisons of potential outcomes. Suppose further that the researcher defends this interpretation by arguing that the treatment is randomly assigned under *some* design. However, as is frequently the case, the researcher neither explicitly specifies the treatment randomization model nor the target estimand, hoping instead that the regression specification makes reasonable implicit choices. What are these choices that the researcher should be prepared to defend?

Specifically, this paper provides generic tools that answer the following questions:

(Q1) Is *any* design-based interpretation plausible given the specification?

(Q2) If yes for (Q1), what are the corresponding designs (i.e., the probability distributions of treatment for each unit)?

(Q3) If yes for (Q1), what are the corresponding estimands?

Our approach is constructive, computational, and general: It computes concrete designs and estimands that answer (Q2) and (Q3) without assuming much structure on the specification. We allow for panels, cross-sections, binary, or multi-valued treatments, but we do restrict treatments to be finitely valued. In leading cases, our approach reveals at most one design and at most one estimand consistent with the data—design-based interpretation is indefensible (Q1) if no design is revealed. As a result, our diagnostics are highly informative on what design and economic target that a regression implicitly chooses and that its user should defend.

[2]Among others, Imbens and Angrist (1994); Angrist (1998); Lin (2013); Słoczyński (2022, 2020); Blandhol *et al.* (2025); Aronow and Samii (2016); Goldsmith-Pinkham *et al.* (2024); Borusyak and Hull (2024); Athey and Imbens (2022); Kline (2011); Bugni *et al.* (2023); Mogstad and Torgovitsky (2024); Arkhangelsky and Imbens (2023); Arkhangelsky *et al.* (2021); Chetverikov *et al.* (2023); Kolesár and Plagborg-Møller (2024); Zhao *et al.* (2025).

To fix ideas, on well-understood specifications, our approach produces intuitively reasonable answers. Suppose the researcher simply regresses a scalar outcome on a binary treatment. Our approach finds that a design-based interpretation of this regression coefficient is tenable *only if* the design is *the* randomized experiment in which every unit has the same probability of being treated; this probability is consistently estimated by the treatment's empirical frequency. The estimand under this design is the average treatment effect. When covariates are added to the specification, the unique tenable propensity score may vary linearly with covariates, and the estimand features individual treatment effects weighted by the conditional variance of the treatment (Angrist, 1998; Blandhol *et al.*, 2025).

This approach arises from two simple observations. First, linear regression estimands $\boldsymbol{\tau}$ under a given design are linear aggregations of individual *potential* outcomes $\mathbf{y}_i(\cdot), i = 1, \ldots, n$. In this aggregation, the weight placed on $\mathbf{y}_i(\mathbf{w})$, for a treatment value $\mathbf{w}$, can be written in the form $\pi_i^*(\mathbf{w})\ \boldsymbol{\rho}_i(\mathbf{w})$:

$$\underbrace{\boldsymbol{\tau}}_{\text{Regression estimand, } \mathbb{R}^k} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\sum_{\mathbf{w}}}_{\substack{\text{Summing over} \\ \text{treatment values } \mathbf{w}}} \underbrace{\pi_i^*(\mathbf{w})}_{\text{true design, } \mathbb{R}} \cdot \underbrace{\boldsymbol{\rho}_i(\mathbf{w})}_{\text{potential weights, } \mathbb{R}^{k\times T}} \cdot \underbrace{\mathbf{y}_i(\mathbf{w})}_{\text{potential outcomes, } \mathbb{R}^T} . \tag{1.1}$$

The weight put on $\mathbf{y}_i(\mathbf{w})$ depends on the true design $\pi_i^*(\mathbf{w})$, which is the unknown probability that individual i receives treatment $\mathbf{w}$. It also depends on certain numbers $\boldsymbol{\rho}_i(\mathbf{w})$, which we call *potential weights*, associated with *potential* outcomes. These potential weights encode how a regression specification makes comparisons over counterfactual treatment assignments. Importantly, $\boldsymbol{\rho}_i(\mathbf{w})$ are *known* up to the population Gram matrix of the linear regression (i.e., the "$\mathbb{E}[x_i x_i']$" matrix), which is consistently estimable.

Second, since not all aggregations of potential outcomes are bona fide causal effects, interpreting (1.1) as causal necessarily restricts $\pi_i^*(\mathbf{w})\boldsymbol{\rho}_i(\mathbf{w})$. This in turn implies restrictions on $\pi_i^*(\mathbf{w})$ since $\boldsymbol{\rho}_i(\mathbf{w})$ is estimable. To operationalize this, we consider a weak restriction that all causal comparisons should satisfy. Suppose all valid causal effects $\boldsymbol{\tau} = \boldsymbol{\tau}(\mathbf{y}_1(\cdot), \ldots, \mathbf{y}_n(\cdot))$ at least satisfy the following property, called *level irrelevance* by Blandhol *et al.* (2025): $\boldsymbol{\tau}$, as a function of potential outcomes, is invariant to shifting *every* potential outcome of any unit by the same number.[3] We call estimands satisfying this property *causal contrasts*. All reasonable causal effects should at least be causal contrasts, but being a causal contrast does not automatically guarantee economic relevance.

[3]That is, $\boldsymbol{\tau}$ is not changed by replacing, for any $i \in [n]$ and any conformable vector $\mathbf{c}_i$, *all* potential outcomes $\mathbf{y}_i(\mathbf{w})$ with $\mathbf{y}_i(\mathbf{w}) + \mathbf{c}_i$.

Crucially, it turns out that if $\boldsymbol{\tau}$ were to be a causal contrast under $\boldsymbol{\pi}^*$, then for every individual, over all treatment values $\mathbf{w}$, the weights on $\mathbf{y}_i(\mathbf{w})$ necessarily sum to zero

$$\text{For every } i = 1, \ldots, n, \sum_{\mathbf{w}} \pi_i^*(\mathbf{w}) \boldsymbol{\rho}_i(\mathbf{w}) = 0. \tag{1.2}$$

Since the potential weights $\boldsymbol{\rho}_i(\mathbf{w})$ are known, these restrictions are linear restrictions for the true design $\pi_i^*(\mathbf{w})$. Treating (1.2) as linear equations in the unknown $\pi_i^*(\mathbf{w})$ with known coefficients $\boldsymbol{\rho}_i(\mathbf{w})$, we call the solutions to these equations the *implicit design* of the specification, and distinguish them from the true assignment process $\pi_i^*(\mathbf{w})$ that Nature chooses. In leading settings, there are more equations in (1.2) than there are unknowns, meaning that the set of implicit designs is often either empty or a singleton.

Implied by these observations, our core observation is simply: The estimand $\boldsymbol{\tau}$ is a vector of causal contrasts under the true design $\{\pi_i^*(\cdot)\}$ if and *only if* the true design is equal to some implicit design $\{\pi_i(\cdot)\}$. Because the potential weights in (1.2) are known up to an estimable Gram matrix, we can simply compute the set of implicit designs $\pi_i(\mathbf{w})$. This in turn is informative of the true design $\pi_i^*(\mathbf{w})$ when the estimand $\boldsymbol{\tau}$ is thought to be causal contrasts. Thus, returning to (Q1)–(Q3):

(A1) No design-based interpretation if (1.2) cannot be solved.

(A2) Any design-based interpretation must be under some implicit design.

(A3) Plugging plausible implicit designs as $\pi_i^*(\cdot)$ into (1.1) reveals the causal estimands targeted by the regression.

Our results inform the debate between critics of regression and those who are more sanguine on regression's utility.[4] The heart of this debate is whether it is harmless or reckless for researchers to let regression make certain choices on autopilot. This paper advances this debate with the simple observation that these implicit choices—how a regression models the assignment process and chooses the economic parameter—are in part knowable and computable. Depending on the context, these choices may be difficult to defend or they

[4]At risk of oversimplification, the critics argue that empiricists should target clearly defined economic parameters; thus, the practice of fixing an estimation method (typically regression) and defining the estimand through the estimation method is dubious (This sentiment is present in, e.g., Blandhol *et al.*, 2025; Mogstad and Torgovitsky, 2024; Yitzhaki, 1996; Freedman, 2008; Imbens, 2015; Heckman and Urzua, 2010; Manski and Tamer, 2019; Haile, 2018). In contrast, others argue that regressions provide practical approximations to structural quantities—whether as outcome models or as implicit models of design—and in many cases target reasonable and nonparametrically defined economic quantities; moreover, understanding the properties of regression without imposing a linear model on the outcomes is valuable (This sentiment is present in, e.g., Goldsmith-Pinkham *et al.*, 2024; Angrist and Pischke, 2008; Angrist, 2001; Imbens and Angrist, 1994; Kline, 2011; Imbens, 2010; Lin, 2013; Borusyak and Hull, 2024).

may be reasonable and robust to minor changes. At a minimum—and especially since regressions are so popular—these choices should be made transparent to scrutiny.

To that end, we propose a set of practical tools to assess and improve potential weights, implicit designs, and implicit estimands, which we illustrate with an application to Blakeslee *et al.* (2020). If a plausible implicit design is found and the corresponding estimand sensible, then the regression specification plausibly makes reasonable choices—thereby dispelling concerns like "negative weights" altogether. If the regression reveals a defensible design but not an economically motivated estimand, the researcher can conveniently construct estimates for alternative estimands using the implicit design; these alternative estimates are then comparable to the regression estimates because they have the same implicit treatment randomization model. Finally, the researcher can also assess how and whether the regression-chosen estimand overrepresents individuals with certain treatment effects.

If no defensible implicit designs are found, empirical researchers are better served either by "design-native" methods that explicitly model the process of treatment assignment (see Wager, 2024, for a review of the modern literature on estimating causal effects), or by defending the regression specification as restrictive models of potential outcomes. Interpreting the regression estimates as design-based is no longer tenable. Nevertheless, if the original regression encodes substantive knowledge of the assignment process and is only mildly misspecified, it is possible to at least nudge things in a design-based direction by *recalibrating* its implicit design. We term this do-no-harm approach *design patching* and cautiously recommend it as a simple fix.

When applied to some common specifications, our framework uncovers new theoretical insights. First, we unify and extend many results in the literature in starkly distinct settings, by simply computing potential weights and implicit designs (Angrist, 1998; Blandhol *et al.*, 2025; Goldsmith-Pinkham *et al.*, 2024; Kline, 2011; Athey and Imbens, 2022). For these specifications, our framework can supply a converse of sorts: We show that *the only* designs justifying causal interpretations are exactly those analyzed in the literature.

Second, we show that specifications that interact a binary treatment with covariates have delicate design-based interpretations (Lin, 2013; Miratrix *et al.*, 2013; Imbens and Wooldridge, 2009; Kline, 2011; Zhao *et al.*, 2025). *Different* designs are necessary for causal interpretations of different contrasts of the coefficients (Chattopadhyay and Zubizarreta, 2023). These designs are often *mutually incompatible*: Requiring all contrasts to be causal necessitates generally logically inconsistent restrictions on the design.

Lastly, we show that two-way fixed effects (TWFE) specifications have fragile and stringent design-based interpretations under staggered adoption. Adding time-varying covariates or estimating on imbalanced panels can render the equations (1.2) unsolvable; even absent these features, the only solution to (1.2) involves totally randomized treatment paths.

This paper is particularly related to Chattopadhyay and Zubizarreta (2023), who characterize and interpret the weights on observed outcomes for several leading linear regression specifications. Our perspective departs from theirs in at least two ways: First, we treat their weights as the observed counterpart to the potential weights, much like observed outcomes are potential outcomes evaluated at the observed treatment. Second, we consider necessary conditions for interpreting the regression as estimating *some* causal effect, whereas Chattopadhyay and Zubizarreta (2023) focus on sufficient conditions for interpreting regression as estimating the average treatment effect.

This paper proceeds as follows. Section 2 contains our main results. To build intuition, Section 2.1 starts with the simple setting with cross-sectional data and binary treatments. Section 2.2 contains our main results, which analyze regressions in the population. Section 2.3 discusses a few practical takeaways. Section 3 then applies our framework to a litany of regression specifications, yielding new theoretical results. Section 4 illustrates our diagnostics with Blakeslee *et al.* (2020), who study the economic impact of water loss in rural India, using plausibly as-if random variation. Section 5 concludes.

2. Potential weights and implicit designs

Consider a finite population of units $i \in [n] \equiv \{1, \dots, n\}$. Each individual receives one of $J+1$ treatments $\mathbf{w} \in \mathcal{W}$. $\mathcal{W}$ is some generic finite set with $J+1 < \infty$ elements. We denote by $\mathbf{W}_i$ the realized treatment of unit i. Each unit is associated with a vector of covariates $\mathbf{x}_i$ and $J+1$ vector-valued potential outcomes of length T, $\{\mathbf{y}_i(\mathbf{w}) \in \mathbb{R}^T : \mathbf{w} \in \mathcal{W}\}$, without restrictions.[5] Upon treatment realization, we observe a corresponding realized outcome $\mathbf{Y}_i = \mathbf{y}_i(\mathbf{W}_i)$.

To accentuate the fact that identification allegedly comes from variation in treatment assignment, we isolate this variation by thinking of $(\mathbf{x}_i, \mathbf{y}_i(\cdot))$ as fixed numbers and only considering the randomness in $\mathbf{W}_i$.[6] To that end, let Π^* denote the joint distribution of

[5] For expositional clarity, we assume that the dimension of the outcome vector is the same across individuals (i.e., balanced panels). Section 3.3 discusses imbalanced panels.

[6] This finite-population perspective is also the starting point for the literature on design-based *statistical inference*. Our focus is on justifying identification instead and is distinct from inferential issues. Both perspectives are "design-based" in the sense that the variation comes from the treatment assignment.

$(\mathbf{W}_1, \ldots, \mathbf{W}_n)$ and let $\boldsymbol{\pi}^*$ denote the associated marginal treatment assignment probabilities (i.e., propensity scores):

$$\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_n^*) \text{ where } \pi_i^*(\mathbf{w}) = \mathrm{P}_{\Pi^*}(\mathbf{W}_i = \mathbf{w}).$$

We refer to such a collection $\boldsymbol{\pi}$ of marginal treatment probabilities as a *design*. We call $\boldsymbol{\pi}^*$ the *true design* and Π a *joint design*.[7]

We think of $(\Pi^*, \boldsymbol{\pi}^*)$ as unknown and chosen by Nature; they are also unknowable since only a single draw $\mathbf{W}_i$ from $\pi_i^*(\cdot)$ is ever observed. $\boldsymbol{\pi}^*$ could in principle be in arbitrary configuration with the covariates and the potential outcomes—meaning that units select into treatment in an arbitrary fashion. However, if a researcher would like to defend a regression estimand on design-based grounds, they should then be prepared to defend restrictions on Π^* consistent with the data. We precisely uncover these restrictions.

Linear regressions are generically written as $\mathbf{Y}_{it}$ regressed on some known transformations of covariates and treatment $z_t(\mathbf{x}_i, \mathbf{W}_i) \in \mathbb{R}^K$ with coefficients $\beta \in \mathbb{R}^K$:

$$\mathbf{Y}_{it} = z_t(\mathbf{x}_i, \mathbf{W}_i)'\beta + \epsilon_{it}.$$

We are interested in interpreting certain regression contrasts $\boldsymbol{\tau} = \Lambda\beta \in \mathbb{R}^k$ as a vector of causal effects, for a known matrix $\Lambda \in \mathbb{R}^{k \times K}$, $k \leq K$. Importantly, the regression is not treated as a model for $\mathbf{y}_i(\cdot)$, but simply as an algorithm that—in the population—produces some estimand $\boldsymbol{\tau}$ given $(\Pi^*, (\mathbf{y}_i(\cdot), \mathbf{x}_i)_{i=1}^n)$.

This setup is sufficiently general to encompass cross-sectional ($T = 1$), panel ($T > 1$), binary treatment ($J + 1 = 2$), multivalued treatment ($J + 1 > 2$), scalar contrast ($k = 1$), and multiple contrasts ($k > 1$) settings.[8] To flesh out the intuition in (1.1) and (1.2), we start informally with the binary-treatment, scalar-outcome, and scalar-contrast case ($T = J = k = 1$). Our main results push this essential intuition to the more general case and formalize the exercise.

Alternatively, one could imagine drawing $(\mathbf{x}_i, \mathbf{y}_i(\cdot)) \overset{\text{i.i.d.}}{\sim} P_0$ from some population and then subsequently conditioning on the drawn values. To translate our results between the finite-population and superpopulation perspectives, it is usually sufficient to replace averaging over the finite population ($\frac{1}{n}\sum_{i=1}^n$) with the population expectation over $\mathbb{E}_{\mathbf{x}_i, \mathbf{y}_i(\cdot) \sim P_0}$. The superpopulation analogue of our results would have the conditional-on-$\mathbf{x}_i$ expectation version of (1.2) holding for P_0-almost every $\mathbf{x}$. We prefer the finite-population version for mainly cosmetic reasons, as the extra measure-theoretical details would distract from our central point.

[7] Because we focus on linear estimators, only $\boldsymbol{\pi}^*$ matters for population properties; Π^* in turn matters for the statistical properties of the regression estimator.

[8] Our results do extend to settings with infinitely-valued $\mathbf{w}$ (e.g. continuous treatments), but they become much less powerful, essentially since there are only finitely many restrictions for infinitely many objects.

2.1. **Leading example and core intuition.** We do not bold variables in this familiar setting. To further fix ideas, a popular regression in this case may be:

$$Y_i = \alpha + \tau W_i + x_i'\gamma + \epsilon_i, \tag{2.1}$$

for which $z(x_i, W_i) = [1, W_i, x_i']', \beta = [\alpha, \tau, \gamma']', \Lambda = [0, 1, 0'_{\dim(x)}]$. Since treatment is binary, for a design $\boldsymbol{\pi}$, we write $\pi_i = \pi_i(1)$ and $1 - \pi_i = \pi_i(0)$. We consider potentially more intricate specifications, keeping (2.1) as a useful example.

As an empirical motivation, consider Blakeslee *et al.* (2020), who study the impact of water loss in rural India on employment and income. Water loss is proxied by a binary W_i, indicating whether the first borewell household i drilled has failed at the time of the authors' survey. Blakeslee *et al.* (2020) appear to argue for identification on design-based grounds, writing, e.g., "a household's access to groundwater depends on highly irregular and quasi-random properties of the subsurface beneath its land" (p.206).

The true design $\boldsymbol{\pi}^*$—the natural process of borewell failure—is unknown, but the authors argue that it nevertheless does not preclude causal inference. Since borewell failure—a function of complex hydrogeological features—is difficult to predict, households with failed wells are otherwise similar to their more fortunate counterparts. To buttress this argument, the authors painstakingly and carefully describe hydrogeological features that determine borewell failure, even having inserted specialized camera equipment into failed wells.

As is common practice, however, the authors estimate treatment effects via (2.1), for many different outcome variables. This appears somewhat in tension with the rich descriptions of treatment assignment: These regressions are not connected—at least not explicitly—to those rich descriptions of $\boldsymbol{\pi}^*$ in the paper. What should a reader believe about borewell failure to interpret the authors' regression estimates as causal effects?

Consider a regression of Y_i on $z(x_i, W_i)$, which may be more complex than (2.1). The population regression coefficient is defined as an analogue of the familiar regression formula (cf. $\mathbb{E}[x_i x_i']^{-1}\mathbb{E}[x_i y_i]$)

$$\beta = \left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{W_i \sim \pi_i^*}\left[z(x_i, W_i) z(x_i, W_i)'\right] \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W_i \sim \pi_i^*}[z(x_i, W_i) y_i(W_i)] \right).$$

The two terms in β average over the finite population ($\frac{1}{n}\sum_{i=1}^n$) and integrate over the only random objects $W_1, \ldots, W_n$ in our setup.

Critically, let

$$G_n \equiv G_n(\boldsymbol{\pi}^*) \equiv \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{W_i \sim \pi_i^*}\left[z(x_i, W_i) z(x_i, W_i)'\right]$$

denote the *Gram matrix* of this regression. G_n is a population object that depends on $\boldsymbol{\pi}^*$. However, since G_n aggregates the individually unknowable π_i^*, the Gram matrix can usually be estimated well by its sample counterpart $\hat{G}_n = \frac{1}{n}\sum_{i=1}^n z(x_i, W_i)z(x_i, W_i)'$.[9] Much like in analyses of identification (Koopmans and Reiersol, 1950), we abstract away from the estimation noise and treat G_n as known. This known-G_n perspective is crucial for our results: We shall provide additional arguments for it at various points later.

With these preparations, observe that the estimand τ can be written in the form of (1.1):

$$\tau = \Lambda\beta = \frac{1}{n}\sum_{i=1}^n \pi_i^* \underbrace{\Lambda G_n^{-1} z(x_i, 1)}_{\rho_i(1)} y_i(1) + (1-\pi_i^*)\underbrace{\Lambda G_n^{-1} z(x_i, 0)}_{\rho_i(0)} y_i(0). \tag{2.2}$$

Here, the *potential weights* $\rho_i(w) = \Lambda G_n^{-1} z(x_i, w)$ are known up to G_n. Just like potential outcomes $y_i(w)$, the potential weights are indexed by the unit and the treatment assignment.

Potential weights are readily connected to the weighting in the regression *estimator*. The least-squares estimator—itself linear in Y_i—can be written as

$$\hat{\tau} = \Lambda\hat{\beta} = \Lambda\hat{G}_n^{-1}\frac{1}{n}\sum_{i=1}^n z(x_i, W_i)Y_i = \frac{1}{n}\sum_{i=1}^n \underbrace{\Lambda\hat{G}_n^{-1} z(x_i, W_i)}_{\hat{\rho}_i(W_i)} y_i(W_i).$$

The weight on Y_i in $\hat{\tau}$ corresponds exactly to the empirical counterpart of $\rho_i(W_i)$—the potential weights for i evaluated at i's observed outcome. Comparing with (2.2) and ignoring the difference $\hat{G}_n \neq G_n$, we can think of the potential weight $\rho_i(1-W_i)$ of the *counterfactual* treatment $1-W_i$ as what the regression estimator *would have done* to the potential outcome $y_i(1-W_i)$, had unit i received the opposite assignment instead. In this sense, the potential weights encode how the regression compares the two potential outcomes.

Specializing (2.2) to (2.1), the potential weights are

$$\rho_i(w) = \frac{w - \tilde{x}_i'\delta}{\frac{1}{n}\sum_{i=1}^n \tilde{x}_i'\delta(1-\tilde{x}_i'\delta)} \text{ for } \tilde{x}_i \equiv [1, x_i']', \ \delta \equiv \left(\frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \tilde{x}_i\pi_i^*\right).$$

The $\rho_i(w)$ is proportional to w minus the population projection of π_i^* onto the covariates, whose projection coefficient is δ. Analogously, we may compute that $\hat{\rho}_i(W_i)$ replaces w with W_i and δ with its sample counterpart $\hat{\delta} = \left(\frac{1}{n}\sum_{i=1}^n \tilde{x}_i\tilde{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \tilde{x}_i W_i\right)$.

The representation (2.2) simply writes τ as an aggregation of potential outcomes. Not all such aggregations deserve to be called causal effects. To interpret (2.2) as "causal," we must defend certain restrictions on $\pi_i^*\rho_i(1), \pi_i^*\rho_i(0)$. For instance, were one to assert that

[9] In particular, since the regression *estimator* replaces G_n with $\hat{G}_n$, it is implausible that the regression estimator is satisfactory (e.g., consistent) but $\hat{G}_n$ is far from G_n. Appendices A and C.2 provide formal guarantees for $\hat{G}_n$.

(2.2) is the average treatment effect (ATE)

$$\tau \overset{(2.2)}{=} \frac{1}{n}\sum_{i=1}^{n} \pi_i^* \rho_i(1) y_i(1) + (1-\pi_i^*)\rho_i(0) y_i(0) \overset{\text{Interpret as ATE}}{=} \frac{1}{n}\sum_{i=1}^{n} y_i(1) - y_i(0),$$

then one in effect asserts $\pi_i^* \rho_i(1) = 1$ and $(1-\pi_i^*)\rho_i(0) = -1$. Since $\rho_i(1), \rho_i(0)$ are known given the estimable G_n, these assertions have empirical content.

Rather than demanding that τ estimates the ATE, we consider the following strictly weaker property that all causal comparisons should satisfy. Namely, they should be invariant to transformations of potential outcomes that only shift underlying levels but not the individual treatment effects. Blandhol *et al.* (2025) call the (superpopulation analogue of) this property *level irrelevance*. We call estimands that satisfy this property causal contrasts.

Definition 2.1. For τ some function of $(\boldsymbol{\pi}^*, (y_i(\cdot), x_i)_{i=1}^n)$, we say that τ is a *causal contrast* under $\boldsymbol{\pi}^*$ if τ is always unchanged when we replace all potential outcomes $y_i(w)$ with $y_i(w) + c_i$ for arbitrary $c_i \in \mathbb{R}$.

For a τ that is linear in $\{(y_i(1), y_i(0)) : i \in [n]\}$, Definition 2.1 is equivalent to τ being a weighted average treatment effect, allowing for negative unit-level weights. This is a natural and weak requirement for causal objects.[10] Allowing for negative weights is perhaps too lenient for some treatment effects, but it accommodates for, e.g., differences in treatment effects between subgroups. Satisfying Definition 2.1 simply means that τ is a comparison of potential outcomes and does not by itself justify τ's economic relevance.

On the other hand, imposing Definition 2.1 is also strong in the sense that it rules out any modeling restrictions on $y_i(0)$, as the constants c_i's can be arbitrarily and adversarially chosen.[11] In this sense, Definition 2.1 reflects the fact that identification rests *purely* on design-based grounds. Had we imposed, say, that $y_i(0)$ is a linear function of x_i, we could reasonably decide that certain aggregations τ that do not satisfy Definition 2.1 are causal as well. That said, even when we are willing to impose outcome models, we might retain an interest in whether these outcome restrictions can be relaxed under certain designs. Thus, for our purposes, we define causal contrasts as aggregations satisfying Definition 2.1, disambiguating it from the less precise colloquial usage of "causal effects."

[10] Definition 2.1 does rule out ratio-type comparisons like $\frac{1}{n}\sum_{i=1}^n y_i(1) / \frac{1}{n}\sum_{i=1}^n y_i(0) - 1$ or $\frac{1}{n}\sum_{i=1}^n y_i(1)/y_i(0) - 1$, which may be colloquially referred to as causal. Since a regression is linear in $y_i(\cdot)$, it would generally not estimate these nonlinear comparisons in the first place. Thus restricting to aggregations of $y_i(1) - y_i(0)$ is without essential loss.

[11] It does not rule out restrictions on individual treatment effects, since adding c_i's leave initial treatment effects unchanged.

Observe that τ satisfies Definition 2.1 under $\boldsymbol{\pi}^*$ if and only if

$$\pi_i^* \rho_i(1) + (1 - \pi_i^*)\rho_i(0) = 0 \text{ for all } i = 1, \ldots, n. \tag{2.3}$$

By (2.2), the left-hand side of (2.3) represents the effect on τ of shifting unit i's potential outcomes by $c_i = n$. By Definition 2.1, this effect must be zero. Thus, the equation (2.3) must hold for $\boldsymbol{\pi}^*$ if we are to interpret τ as a causal contrast. Of course, there is no intrinsic reason why Nature obeys (2.3) when casting the proverbial die $\boldsymbol{\pi}^*$. However, were we to argue that Nature randomizes treatment in such a manner that our regression estimates a causal contrast regardless of potential outcomes, then we must defend (2.3).

To make the restrictions in (2.3) explicit, we can solve for π_i^*, at least when $\rho_i(0) \neq \rho_i(1)$. Let π_i denote the solution to (2.3) viewed as equations in π_i^*, and we call $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ an *implicit design*:

$$\pi_i = \frac{-\rho_i(0)}{\rho_i(1) - \rho_i(0)}.$$

The restriction (2.3) can then be equivalently stated as $\pi_i^* = \pi_i$—that the true design is equal to the implicit design. This makes explicit the necessary conditions of interpreting τ causally.

Furthermore, pinpointing $\boldsymbol{\pi}$ also pinpoints the corresponding estimand. To see this, if (2.3) holds, then τ is equal to a weighted average treatment effect with weights $\omega_i \equiv \pi_i \rho_i(1) = -(1 - \pi_i)\rho_i(0)$

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \pi_i \rho_i(1)(y_i(1) - y_i(0)) = \frac{1}{n} \sum_{i=1}^{n} \omega_i (y_i(1) - y_i(0)).$$

These weights are non-negative ($\omega_i \geq 0$) when $\rho_i(1) \geq 0$ and $\rho_i(0) \leq 0$. Therefore, if we are to interpret τ as a causal contrast under $\boldsymbol{\pi}^*$, then, simply by computing the solution to linear equations (2.3), we would recover both what the design $\boldsymbol{\pi}^*$ and what our target τ must be—and thus what we must defend economically and statistically. This answers (Q2) and (Q3) posed in the introduction.

For (Q1), there are at least two ways in which π_i cannot possibly equal π_i^* and are thus indefensible. The more obvious one is if $\pi_i \notin [0, 1]$ for any i or if $\rho_i(1) = \rho_i(0) \neq 0$. This occurs when both $\rho_i(1)$ and $\rho_i(0)$ are on the same side of zero. When this happens, the implicit design is not a probability distribution, and cannot equal the true design. More subtlely, π_i is also indefensible if it implies a different Gram matrix:

$$G_n(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^{n} \pi_i z(x_i, 1) z(x_i, 1)' + (1 - \pi_i) z(x_i, 0) z(x_i, 0)' \neq G_n(\boldsymbol{\pi}^*). \tag{2.4}$$

This restriction is useful when we analyze specifications theoretically under this framework. It is harder to implement when we do not know and have to estimate G_n—though one could use it as a basis for inference (Appendix A).

Returning to the specification (2.1), solving (2.3) shows that the implicit design is precisely the population projection of π_i^* onto the covariates:

$$\pi_i = \frac{-\rho_i(0)}{\rho_i(1) - \rho_i(0)} = \frac{\tilde{x}_i'\delta}{1 - \tilde{x}_i'\delta + \tilde{x}_i'\delta} = \tilde{x}_i'\delta,$$

where the corresponding estimand is a weighted average treatment effect

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \omega_i(y_i(1) - y_i(0)) \quad \omega_i = \frac{\pi_i(1-\pi_i)}{\frac{1}{n}\sum_{j=1}^{n}\pi_j(1-\pi_j)}. \tag{2.5}$$

Thus, to interpret (2.1) as estimating an economically interesting causal effect, one must defend (a) that $\pi_i^* = \pi_i = \tilde{x}_i'\delta$ and (b) that one is interested in the $\pi_i(1-\pi_i)$-weighted average treatment effect. On the other hand, a design-based interpretation is obviously untenable if we observe $\pi_i \notin [0,1]$, which happens when the linear probability model for π_i^* produces nonsensical fitted values.[12]

The specification (2.1) is analyzed in Angrist (1998) and Blandhol *et al.* (2025). Angrist (1998) essentially shows that if $\pi_i^* = \delta'\tilde{x}_i$ for some δ, then τ in (2.1) estimates the weighted treatment effect (2.5). In turn, Blandhol *et al.* (2025) show that the linearity of the propensity score is also necessary.[13] Our framework recovers these results by simply mechanically computing $\rho_i(1)$ and $\rho_i(0)$. We echo Blandhol *et al.* (2025) in that if one would like to interpret τ as *some* causal contrast (Definition 2.1), then one must be willing to believe the linearity $\pi_i^* = \delta'\tilde{x}_i$ and that one is targeting (2.5). There is no other configuration that justifies a design-based interpretation of this regression.

Before moving on to the general case with any (k, T, J), we distill the key elements from this leading example and discuss a few subtleties.

[12]For this specification, Theorem 3.1 verifies that $G_n(\boldsymbol{\pi}) = G_n(\boldsymbol{\pi}^*)$ always holds regardless of whether $\boldsymbol{\pi} = \boldsymbol{\pi}^*$, and thus there is no additional power in the Gram matrix criterion (2.4).

[13]Both Angrist (1998) and Blandhol *et al.* (2025) analyze slightly different setups from ours. Both consider a superpopulation sampling setup in contrast to our finite population setup. Angrist (1998) considers a binary x_i in his equation (9), but the argument can be easily generalized, e.g., in Borusyak and Hull (2024); Goldsmith-Pinkham *et al.* (2024).
Corollary 1 in Blandhol *et al.* (2025) shows that assuming unconfoundedness, τ is a positively weighted average treatment effect if and only if the propensity score is linear. Our result drops the (finite-population analogues of) unconfoundedness assumptions. That said, the underlying mechanics and takeaways are very similar, and so we view our result (formally in Theorem 3.1(1)) as a reinterpretation of theirs.

(i) We take the perspective that, by observing a regression specification in the population, the triplet $(\{z(x_i,\cdot)\}_{i=1}^n, \Lambda, G_n)$ is known. Here, $(z(x_i,\cdot),\Lambda)$ describes the regression specification, and G_n encodes a set of observable implications of the data.

(ii) We write the population regression estimand τ in the form (1.1) and (2.2). Because we treat G_n as known, the *potential weights* $\rho_i(w)$ are known for all units.

(iii) We observe that imposing *level irrelevance* (Definition 2.1) amounts to imposing that π_i^* satisfies linear restrictions, in which the coefficients are the potential weights. In general, there are kT restrictions for J unknowns.

(iv) We call the solutions to these linear equations *implicit designs*. Computing them also yields the corresponding estimands by (2.2). If (and only if) τ is a causal contrast under $\boldsymbol{\pi}^*$, then $\boldsymbol{\pi}^*$ must equal one of the implicit designs.

The downstream ramification of treating G_n as known in (i) could be somewhat subtle. Alternatively, we can think of (2.3) as a set of nonlinear equations in π_i^*:

$$\pi_i^*\rho_i(1;\boldsymbol{\pi}^*) + (1-\pi_i^*)\rho_i(0;\boldsymbol{\pi}^*) = 0$$

where $\rho_i(\cdot;\boldsymbol{\pi}^*)$ depends on $\boldsymbol{\pi}^*$ through G_n. These equations then define a nonlinear surface of designs $\{\boldsymbol{\pi} : \pi_i\rho_i(1;\boldsymbol{\pi}) + (1-\pi_i)\rho_i(0;\boldsymbol{\pi}) = 0 \,\forall\, i\}$. This surface is defined solely as a function of the regression $(\Lambda, z(x_i,\cdot))$ and thus does not reflect any information in the data.

The implicit designs can then be thought of as choosing a point near[14] this surface that is "most consistent" with the data by fixing $(G_n, \rho_i(\cdot))$ and solving the corresponding equations. For the regression (2.1), this surface is the span of the column space of the covariates $\left\{\boldsymbol{\pi} : \pi_i = \tilde{x}_i'\delta, \delta \in \mathbb{R}^{\dim(x_i)+1}\right\}$; the implicit design chooses the point on the surface by projecting $\boldsymbol{\pi}^*$ to the covariates, since all other points are ruled out by large datasets. When $\boldsymbol{\pi}^*$ belongs to the surface, doing this exactly recovers $\boldsymbol{\pi}^*$.

Finally, in the context of (2.1), we also give more intuition what it means to ignore the estimation noise in $\hat{G}_n$. We can numerically write the regression estimator as an inverse propensity-weighting estimator: For $\hat{\pi}_i = \frac{-\hat{\rho}_i(0)}{\hat{\rho}_i(1)-\hat{\rho}_i(0)}$ and $\hat{\omega}_i = \hat{\pi}_i\hat{\rho}_i(1) = (1-\hat{\pi}_i)\hat{\rho}_i(0)$ the sample analogues of (π_i,ω_i),

$$\begin{aligned}\hat{\tau} &= \frac{1}{n}\sum_{i=1}^n \hat{\rho}_i(W_i)y_i(W_i) = \frac{1}{n}\sum_{i=1}^n \hat{\omega}_i\left(\frac{W_i}{\hat{\pi}_i}y_i(1) + \frac{1-W_i}{1-\hat{\pi}_i}y_i(0)\right)\\ \text{when specialized to (2.1),}\ &= \frac{1}{n}\sum_{i=1}^n \frac{\tilde{x}_i'\hat{\delta}(1-\tilde{x}_i'\hat{\delta})}{\frac{1}{n}\sum_j \tilde{x}_j'\hat{\delta}(1-\tilde{x}_j'\hat{\delta})}\left(\frac{W_i}{\tilde{x}_i'\hat{\delta}}y_i(1) - \frac{1-W_i}{1-\tilde{x}_i'\hat{\delta}}y_i(0)\right).\end{aligned} \tag{2.6}$$

[14] When a implicit design generates the same Gram matrix as the true design $G_n(\boldsymbol{\pi}) = G_n(\boldsymbol{\pi}^*)$, then the implicit design is on the surface.

Representation (2.6) makes apparent that, from a design-based perspective, we can think of a regression estimator as doing two things simultaneously. It first decides on a model for the propensity score and on a target estimand—in (2.1), the model is a linear probability model and the target is the $\pi^*(1-\pi^*)$-weighted average treatment effect. Second, it then estimates the propensity score model and plugs in the estimated coefficients from that model—in (2.1), replacing δ with $\hat{\delta}$. For design-based interpretation, we are primarily interested in the first. To that end, the Gram matrix G_n exactly encodes how the regression models propensity scores and chooses the estimand. Abstracting away from the estimation noise in $\hat{G}_n$ exactly ignores the estimation noise in the propensity model. Indeed, in (2.1), knowing G_n is equivalent to knowing the population regression coefficient δ.

2.2. **Main results.** Returning to the general setup, consider a regression of $\mathbf{Y}_{it}$ on a covariate transform $z_t(\mathbf{x}_i, \mathbf{W}_i) \in \mathbb{R}^K$. The population Gram matrix is now

$$G_n(\boldsymbol{\pi}^*) \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}_{\mathbf{W}_i\sim\pi_i^*}\left[z_t(\mathbf{x}_i,\mathbf{W}_i)z_t(\mathbf{x}_i,\mathbf{W}_i)'\right],$$

whose sample counterpart $\hat{G}_n$ omits the expectation over $\mathbf{W}_i$. We proceed by formalizing (i) through (iv) and add additional commentary along the way.

Definition 2.2 (Population regression specification)**.** Let $G_n = G_n(\boldsymbol{\pi}^*)$, assumed to be invertible. Let $\mathbf{z}(\mathbf{x}_i, \cdot)$ be a $T \times K$ matrix whose rows collect $z_t(\mathbf{x}_i, \cdot)'$. A *population regression specification* collects the objects

$$(\Lambda, G_n, \mathbf{z}(\mathbf{x}_1, \cdot), \ldots, \mathbf{z}(\mathbf{x}_n, \cdot)).$$

The corresponding population regression estimand $\boldsymbol{\tau}$ is

$$\boldsymbol{\tau} \equiv \Lambda G_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}_{\mathbf{W}_i\sim\pi_i^*}[z_t(\mathbf{x}_i,\mathbf{W}_i)\mathbf{y}_{it}(\mathbf{W}_i)].$$

This definition formalizes the analogue to item (i): If an empirical researcher shows us a regression specification, what we take to mean is that the researcher reveals how the regression is implemented ($\mathbf{z}(\mathbf{x}_i, \cdot)$), which coefficients to interpret causally (Λ), and how the regression models and estimates the process of treatment assignment (G_n). By analogy with analyses of identification (Manski, 2022; Koopmans and Reiersol, 1950), we abstract away from the estimation noise in G_n, which is discussed subsequently in Appendix A. We pause for three remarks before moving on to (ii).

Remark 2.1 (Fixed effects)**.** To justify $\hat{G}_n \approx G_n$, it is important to include unit-level fixed effects via the within-transformation. That is, unit fixed effects should be parametrized by

ensuring that the covariate transform $z_t(\mathbf{x}_i, \mathbf{w})$ sums to zero: $\sum_{t=1}^{T} z_t(\mathbf{x}_i, \mathbf{w}) = 0$ for all i and $\mathbf{w} \in \mathcal{W}$. It may be tempting to instead parametrize fixed effects as unit-level dummy variables. However, the Gram matrix G_n associated with the many-dummy parametrization is in general no longer consistently estimable.[15] One can view this as a finite-population manifestation of the incidental parameter problem. ■

Remark 2.2 (Fixed Gram designs)**.** Under certain joint designs Π^*, it is possible that $\hat{G}_n = G_n(\boldsymbol{\pi}^*)$ with probability one. Under such a design, there is no estimation noise to abstract away. As an example, suppose that we consider $Y_i = \alpha + \tau W_i + \epsilon_i$ with binary W_i, but W_i is assigned in a *completely randomized experiment*.[16] Then, the sample and the population Gram matrices are equal and fixed in every realization of the treatment,

$$\hat{G}_n = G_n = \begin{bmatrix} 1 & n_1/n \\ n_1/n & n_1/n \end{bmatrix}.$$

We term these designs *fixed Gram* designs for a particular specification. Our results in the population can alternatively be read as exact results for fixed Gram designs.[17] ■

Remark 2.3 (Fixed covariates)**.** By assuming that $\mathbf{z}(\mathbf{x}_i, \cdot)$ is known, we restrict the right-hand side of the regression to only include objects unchanged by treatment. If we include mediators or lagged outcomes, then we no longer know *counterfactual* values of the covariate transform. We subsequently cannot "complete" the representation of $\boldsymbol{\tau}$ in terms of potential outcomes in (1.1). On the other hand, known functions of the treatment, e.g. lagged treatments, do not introduce any difficulty.

Because of the incompleteness, our framework does not apply immediately to instrumental variable (IV) settings for local average treatment effect-type interpretations, unlike Blandhol *et al.* (2025); Zhao *et al.* (2025); Słoczyński (2020). That said, many IV regression estimands can be written as the ratio of two regression estimands. Our analysis can apply to the numerator and the denominator separately to yield an interpretation as ratios of causal contrasts, in terms of the randomly assigned instrument. Whether this ratio is the causal effect of the endogenous treatment of any subpopulation relies on assumptions about

[15] To see this, for unit i, let $\tilde{z}_t(\mathbf{x}_i, \mathbf{W}_i)$ denote the covariate transforms that exclude the unit dummy. Assume $z_t(\mathbf{x}_i, \mathbf{W}_i)$ includes a unit dummy. Then $\sum_{t=1}^{T} \mathbb{E}[\tilde{z}_t(\mathbf{x}_i, \mathbf{W}_i)]$ is in the Gram matrix (it is the interaction between $\tilde{z}_t$ and the unit-i dummy variable). However, this quantity is not consistently estimable as unit i is only observed once.

[16] That is, there are n_1 treated units, selected uniformly at random from the set of $\binom{n}{n_1}$ subsets.

[17] Regression estimators in the design-based statistical inference literature are frequently analyzed under fixed Gram designs, as they are unbiased for the estimand (see, e.g., Rambachan and Roth, 2020; Athey and Imbens, 2022; Zhao and Ding, 2022; Neyman, 1923/1990; Lin, 2013).

how endogenous treatments respond to the instrument. Incorporating these monotonicity-type restrictions in this framework is interesting future work. ■

Returning to (ii), exactly analogous to (2.2), we can write

$$\boldsymbol{\tau} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) \cdot \underbrace{\left[\Lambda G_n^{-1}\mathbf{z}(\mathbf{x}_i,\mathbf{w})'\right]}_{\boldsymbol{\rho}_i(\mathbf{w})\in\mathbb{R}^{k\times T}} \mathbf{y}_i(\mathbf{w}) \tag{2.7}$$

which verifies the representation (1.1). Relative to the simple case (2.2), here we sum over $J+1$ values, and the potential weights are now *matrices* of dimension $k \times T$.

Definition 2.3 (Potential weights)**.** Given a population regression specification, the *potential weights* for individual i are matrices defined by $\boldsymbol{\rho}_i(\mathbf{w}) \equiv \Lambda G_n^{-1}\mathbf{z}(\mathbf{x}_i,\mathbf{w})'$.

Remark 2.4 (Invariances to parametrization)**.** Appendix C.1 shows that the potential weights for a given contrast do not depend on how the regression is parametrized. For instance, it does not matter if we write $Y_i = \alpha + \tau W_i + \epsilon_i$ instead as $Y_i = \mu_1 \mathbb{1}(W_i = 1) + \mu_0 \mathbb{1}(W_i = 0) + \epsilon_i$ and consider $\tau = \mu_1 - \mu_0$. Appendix C.1 also shows that the potential weights are suitably invariant under the Frisch–Waugh–Lovell transform. ■

Next, we turn to level irrelevance (iii). A natural generalization is to add arbitrary numbers c_{it} to each of the $J+1$ potential outcomes for unit i at time t.

Definition 2.4 (Causal contrast)**.** We say that $\boldsymbol{\tau}$ is a vector of causal contrasts under $\boldsymbol{\pi}^*$ if it satisfies level irrelevance: That is, $\boldsymbol{\tau}$ is unchanged if we replaced all potential outcomes $\mathbf{y}_{it}(\mathbf{w})$ with $\mathbf{y}_{it}(\mathbf{w}) + c_{it}$, for any $i \in [n], t \in [T]$, and $c_{it} \in \mathbb{R}$.

Again, we think of level irrelevance as a minimal requirement for causal quantities whose identification is purely design-based and outcome model-agnostic (Blandhol *et al.*, 2025). Imposing Definition 2.4 amounts to requiring that, in each entry of $\boldsymbol{\tau}$, each individual's time-t potential outcomes are aggregated by some weights that sum to zero:

$$\boldsymbol{\tau} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\omega}_i(\mathbf{w})\mathbf{y}_i(\mathbf{w}) \text{ for some } \boldsymbol{\omega}_i(\mathbf{w}) \in \mathbb{R}^{k\times T} \text{ where } 0 = \sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\omega}_i(\mathbf{w}).$$

These aggregations are thus considerably more complex than weighted average treatment effects for Definition 2.1.

Moving on, the analogue to (iv) is the following set of equations

$$\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\boldsymbol{\rho}_i(\mathbf{w}) = 0, \quad \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) = 1. \tag{2.8}$$

Absorbing the sum-to-one condition, there are kT restrictions for J unknowns. We formally define the solutions to these equations as *implicit designs*.

Definition 2.5 (Implicit design)**.** Given the potential weights, we call $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ that satisfies (2.8) for every i an *(improper) implicit design*. If each $\pi_i(\mathbf{w}) \geq 0$, then we call $\boldsymbol{\pi}$ a *proper* implicit design. If $\boldsymbol{\pi}$ generates the same Gram matrix G_n,

$$G_n = G_n(\boldsymbol{\pi}) \equiv \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) z_t(\mathbf{x}_i, \mathbf{w}_i) z_t(\mathbf{x}_i, \mathbf{w}_i)',$$

then we say that $\boldsymbol{\pi}$ is *Gram-consistent*.

We can often expect that implicit designs are unique, if they exist. This is because the number of equations is often at least the number of unknowns ($kT \geq J$). For instance, $J + 1$ treatments generate $k = J$ contrasts; panels under staggered adoption admit fewer unique treatment times $(J + 1)$ than time horizon T. Lemma C.8 formally establishes the uniqueness of implicit designs in cross-sectional settings ($T = 1$).

Relative to an implicit design $\boldsymbol{\pi}$, the regression specification estimates the $\boldsymbol{\pi}$-weighted average of the potential outcomes, which we formally define as the *implicit estimand*, indexed by weights $\boldsymbol{\omega}_i(\boldsymbol{\pi}, \mathbf{w}) = \pi_i \rho_i(\mathbf{w})$.

Definition 2.6 (Implicit estimand)**.** Let $\boldsymbol{\pi}$ denote an implicit design of a given regression specification. The corresponding *implicit estimand* is

$$\boldsymbol{\tau}(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathbf{w} \in \mathcal{W}} \boldsymbol{\omega}_i(\boldsymbol{\pi}, \mathbf{w}) \mathbf{y}_i(\mathbf{w}),$$

where $\boldsymbol{\omega}_i(\boldsymbol{\pi}, \mathbf{w}) = \pi_i(\mathbf{w}) \boldsymbol{\rho}_i(\mathbf{w}) \in \mathbb{R}^{k \times T}$. Without introducing ambiguity, we also use "implicit estimand" to refer to the weights $\{\boldsymbol{\omega}_i(\boldsymbol{\pi}, \mathbf{w}) : i \in [n], \mathbf{w} \in \mathcal{W}\}$.

Having formalized (i) through (iv), finally, we collect our observations in the following theorem, which is the main result of this paper.

Theorem 2.1. *$\boldsymbol{\tau}$ is a vector of causal contrasts under $\boldsymbol{\pi}^*$ in the sense of Definition 2.4 if and only if $\boldsymbol{\pi}^*$ is equal to some implicit design. When this happens, if, furthermore, there is a unique implicit design $\boldsymbol{\pi}$, then $\boldsymbol{\pi}$ must be proper and Gram-consistent.*

Theorem 2.1 implies our result in Section 2.1 as a corollary.

Corollary 2.2 ($k = T = J = 1$)**.** *When $k = T = J = 1$, τ is a causal contrast under the true design $\boldsymbol{\pi}^*$ if and only if $\rho_i(1)\rho_i(0) \leq 0$ for all i and $\pi_i^* = \frac{-\rho_i(0)}{\rho_i(1) - \rho_i(0)}$ for all units i with one of $\rho_i(1)$ and $\rho_i(0)$ nonzero. When this happens, the implicit estimand is uniquely defined by $\omega_i(\boldsymbol{\pi}^*, 1) = \pi_i^* \rho_i(1) = -\omega_i(\boldsymbol{\pi}^*, 0)$, and thus*

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \omega_i^* (y_i(1) - y_i(0)) \text{ for } \omega_i^* \equiv \omega_i(\boldsymbol{\pi}^*, 1).$$

The unit-level weights ω_i^* are negative if and only if $\rho_i(1) < 0 < \rho_i(0)$.

Since the set of implicit designs is known in the population and estimable in sample, Theorem 2.1 shows that the implicit designs are informative of $\boldsymbol{\pi}^*$ under the presumption that the regression admits a causal interpretation. Thus, if a researcher interprets $\boldsymbol{\tau}$ as a design-based estimand, then they must be prepared to defend why the true assignment process $\boldsymbol{\pi}^*$ is equal to one of the implicit designs $\boldsymbol{\pi}$.

In the simple case ($k = T = J = 1$), Corollary 2.2 formalizes our discussion in Section 2.1. If it exists, the implicit design is uniquely and explicitly defined, up to units with zero potential weights and do not factor into the estimand. The implicit estimand is a weighted average treatment effect. All unit-level weights ω_i in the implicit estimand are nonnegative if no unit has the "wrongly ordered" potential weights $\rho_i(1) < 0 < \rho_i(0)$.

2.3. **Discussion.** Theorem 2.1 constructively yields plausible designs and estimands that answer (Q1)–(Q3) posed in the introduction. In our view, computing the implicit design and estimand adds transparency and connects estimation to identification. When researchers point to design-based identification arguments but do not explicitly model the design or describe the estimand, our results make clear the choices a regression makes on the model for treatment and on the target estimand. When researchers treat regression as an outcome model, a reasonable implicit design and implicit estimand make the regression interpretable to those who challenge the outcome model but are willing to entertain restrictions on treatment assignment.

Since we aim to understand regressions from the *necessary* implications of their causal interpretation, our arguments do resemble trying to "reverse engineer" an interpretation for the regression estimator. Like Mogstad and Torgovitsky (2024), who critique careless reverse engineering in IV settings, we caution that the mere fact that one could obtain an implicit design does not by itself justify the statistical and economic relevance of the design or the estimand. Whether this is the case involves judgment from the researcher and the audience. To that end, we provide a list of diagnostics that assess and improve the implicit design and implicit estimand. Section 4 illustrates these in action.

To implement these diagnostics, practitioners can explicitly compute potential weights and implicit designs—most straightforwardly by replacing G_n with $\hat{G}_n$. If they are additionally willing to assume that the true design Π^* is fixed Gram (Remark 2.2), then

Theorem 2.1 is applicable as-is. In general, $\hat{G}_n$ involves noise from realized treatment assignments, and we relegate an analysis to Appendix A.[18]

Probing the implicit design. Abstracting away from sampling noise, suppose (1.2) reveals a proper implicit design $\boldsymbol{\pi}$, we may ask two broad questions:

(1) Is $\boldsymbol{\pi}$ consistent with economic intuition?

For instance, does $\boldsymbol{\pi}$ covary positively with features that predict selection into treatment? Does it explain too much variance in the observed W_i to make us question whether W_i is unpredictable?

(2) Is $\boldsymbol{\pi}$ consistent with the data?

For instance, is $\boldsymbol{\pi}$ well-calibrated?[19] With a binary treatment, are (transformations of) covariates balanced by the propensity scores π_i?

If these checks all pass, then it continues to be plausible that $\boldsymbol{\pi} = \boldsymbol{\pi}^*$. While one can never be sure that $\boldsymbol{\pi}^*$ is truly equal to $\boldsymbol{\pi}$, at least there lacks evidence to the contrary.

Probing the implicit estimand. We may also compute the implicit estimand via Definition 2.6. Usually, the estimand would involve weights that do not correspond to treatment effect contrasts from first principles. Similarly, we may check the following broad themes:

(1) Do the weights make sense economically?

For instance, are any weights negative when $\boldsymbol{\tau}$ is not meant to be a difference of treatment effects? How large is the implicit subpopulation whose treatment effect the estimand is measuring (Poirier and Słoczyński, 2024)?

(2) Do the weights matter for the bottom line in a given application?

For instance, one could use the implicit design to construct alternative estimates and check them against the regression estimate: For user-chosen weights $\boldsymbol{\omega}_i(\mathbf{w}) \in \mathbb{R}^{k\times T}$,[20]

[18] To give some intuition that may assuage concerns, given a confidence set for G_n, it is in principle possible to project this confidence set to the set of potential weights and the set of proper and Gram-consistent implicit designs. Moreover, if one has a candidate design $\boldsymbol{\pi}$ in mind such that $\boldsymbol{\pi}$ and $G_n(\boldsymbol{\pi})$ satisfy (1.2), it is also straightforward to check whether the implied $G_n(\boldsymbol{\pi})$ is statistically close enough to $\hat{G}_n$.

[19] That is, among those units with $\pi_i \approx \pi$, do approximately $100\pi\%$ of units get treated?
When we need π_i^* to be linear, Blandhol *et al.* (2025) suggest using the Ramsey RESET test against the hypothesis that $\mathbb{E}[W_i \mid x_i]$ is linear. For fitted values $\hat{\pi}_i$, the RESET test effectively tests $\beta_2 = \cdots = \beta_p = 0$ in the regression

$$W_i = \alpha_0 + \alpha_1' x_i + \beta_1(\hat{\pi}_i)^2 + \cdots + \beta_p(\hat{\pi}_i)^p + \epsilon_i.$$

If we replace $\alpha_0 + \alpha_1' x_i$ with $\hat{\pi}_i$, the RESET test is thus similar to a test of the linearity—and thus calibration—of $\mathbb{E}[W_i \mid \hat{\pi}_i]$.

[20] In the simple case ($k = T = J = 1$), a popular choice may be the trimmed ATE (Crump *et al.*, 2009):

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \omega_i \left(\frac{W_i}{\pi_i} - \frac{1-W_i}{1-\pi_i}\right) Y_i \quad \omega_i \propto \mathbb{1}(\epsilon < \pi_i < 1-\epsilon) \quad \frac{1}{n}\sum_{i=1}^{n} \omega_i = 1.$$

one can form the inverse-propensity weighting estimator

$$\hat{\boldsymbol{\tau}}(\boldsymbol{\pi};\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}\in\mathcal{W}}\frac{\mathbb{1}(\mathbf{W}_i=\mathbf{w})}{\pi_i(\mathbf{w})}\boldsymbol{\omega}_i(\mathbf{w})\mathbf{y}_i(\mathbf{w}).$$

One can also plot estimates of $\mathbb{E}[Y_i \mid W_i = w, \pi_i = \pi]$. If these functions are flat in terms of π, then it is likely that the regression-chosen weights make little difference and alternative user-chosen weights would result in similar estimates.

If these checks pass—we find a plausible implicit design under which the regression estimates a reasonable causal contrast, then we are reassured that there could be a plausible treatment assignment process under which our regression aggregates a sensible set of comparisons. This dispels concerns about "negative weights" when we are willing to defend the implicit design as reflecting reality. Since it is usually straightforward to interpret regressions as outcome models, having a valid and outcome-agnostic design-based interpretation adds an additional robustness.

Alternatives and design patching. If no proper and Gram-consistent implicit design is found or if any of the above checks fails, then defending the regression estimand on design-based grounds may no longer be logically tenable. Practitioners essentially face two options. We may maintain that causal identification rests on treatment variation, but we need to better model the design and target a better estimand—adopting what Mogstad and Torgovitsky (2024) call "forward engineering." There is a large and growing literature in causal inference that provides such methods; rarely are such methods implementable as a simple linear regression (see monographs and reviews by Wager, 2024; Imbens and Rubin, 2015; Ben-Michael *et al.*, 2021; Chernozhukov *et al.*, 2024). Alternatively, we may retreat by defending functional form restrictions on the potential outcomes and treatment effects, sacrificing some agnosticism.[21]

A frequent scenario where design-based interpretation is rejected (for $k = J = T = 1$) is when the implicit design has a small proportion of units with $\pi_i \notin [0, 1]$. This occurs, for instance, in our empirical application to Blakeslee *et al.* (2020) as well as in Blandhol *et al.* (2025)'s illustration of Dube and Harish (2020). It is possible that the implicit propensity model suffers from some misspecification, but is otherwise close to defensible. In that case, practitioners may find the initial regression a flawed but reasonable starting point. Carefully modeling treatment assignment is an alternative, but may not be worth the hassle.

[21]Even with outcome models, identification typically still requires assumptions on the process of treatment assignment. These restrictions are typically of the form that treatment assignment does not predict values of some "structural residual" defined by an outcome model (Andrews *et al.*, 2023).

In such cases, if we cast the problem as simply trying to improve a given predictor π_i, one could consider *recalibrating* the implicit design $\hat{\pi}_i$ into $\hat{\pi}_i^{\text{patched}}$.[22] Heuristically speaking, one could view recalibration as marginally improving the quality of predictions, and doing so produces weakly improved downstream estimates.[23] For user-chosen unit-level weights ω_i (such that $\omega_i = 0$ whenever $\hat{\pi}_i^{\text{patched}} \in \{0, 1\}$), we can then compute the *design-patched* estimate

$$\hat{\tau}^{\text{patched}} = \frac{1}{n} \sum_{i=1}^{n} \omega_i \left(\frac{W_i}{\hat{\pi}_i^{\text{patched}}} - \frac{1 - W_i}{1 - \hat{\pi}_i^{\text{patched}}} \right) Y_i.$$

The patched estimate $\hat{\tau}^{\text{patched}}$ behaves similarly to the regression estimator when the implicit design is close to calibrated. It has an explicit design-based interpretation when the true design is close to the patched design.

3. Theoretical applications and examples

Having established the machinery of potential weights and implicit designs, we now reap its fruits. Our results are useful both for theoretical investigation and for practical diagnostics with a given dataset. This section illustrates the former by computing the potential weights and implicit designs for a number of population regression specifications; Section 4 illustrates how our results apply to an empirical setting. To emphasize, for both purposes, our results essentially reduce the problem to *computing* the potential weights and the set of implicit designs. This makes them broadly applicable to generate a litany of results in starkly distinct settings, with relatively little specialized analysis. Reasoning through $\boldsymbol{\pi}^* = \boldsymbol{\pi}$ theoretically is sometimes bespoke, but can often be aided by numerical methods.

Throughout this section, we assume that the population Gram matrix is invertible. Some of these specifications are well-studied in the literature and have known causal interpretations under certain design assumptions (Angrist, 1998; Goldsmith-Pinkham *et al.*, 2024; Kline, 2011; Imbens and Wooldridge, 2009; Lin, 2013; Athey and Imbens, 2022). We

[22] One simple recalibration method is what Imbens and Rubin (2015) call subclassification on the estimated propensity score. We may discretize the outputs $\hat{\pi}_i$ into $B \ll n$ bins. On bin b, output $\hat{\pi}_i^{\text{patched}}$ equal to the average W_i among observations in the bin. Chapter 17 in Imbens and Rubin (2015) provides an algorithm for choosing the bins, which we apply in Section 4. This method is also called histogram-binning in the machine learning literature (Zadrozny and Elkan, 2001). Zhao *et al.* (2025) also recommend stratifying on the estimated instrument propensity score, which is equivalent to recalibrating the estimated score by binning.

[23] When π_i is linear, this approach can be heuristically understood as perturbing the predicted values in the direction that makes rejection of Ramsey's RESET test less severe.

show that the corresponding implicit designs essentially rediscover these properties and—moreover—supply a converse of sorts. Specifically, we show that the design-based interpretations analyzed in these settings are tenable only under those designs assumed in the literature; they target exactly those estimands found in the literature. There are no other designs or estimands that are consistent with a design-based interpretation.

When applied to a number of examples that are less known, our results uncover some new insights. We show that two types of specifications have design-based interpretations only under stringent conditions. The first class contains cross-sectional regressions that involve the interaction between W and x. For these regressions, if all interactions between W and x are causal contrasts, then essentially, x must be saturated discrete covariates or W must be randomly assigned independently of x. The second class contains two-way fixed effect (TWFE) regressions that either have time-varying covariates or imbalanced panels. For these regressions, if the treatment pattern covaries with the covariates or the imbalance pattern, then implicit designs frequently do not exist.

3.1. **A unified analysis of design-based interpretation in regression.**

Theorem 3.1. *We compute the implicit designs and estimands of the following regression specifications. In all of the following scenarios, the implicit design exists and is unique. The implicit design is Gram-consistent regardless of whether $\boldsymbol{\pi} = \boldsymbol{\pi}^*$ for all scenarios except for (4).*

(1) (Angrist, 1998; Blandhol et al., 2025, Section 2.1, $k = T = J = 1$) Suppose x_i includes a constant. For $Y_i = \tau W_i + x_i'\gamma + \epsilon_i$ and τ the coefficient of interest:

- *(a) $\pi_i = x_i'\delta$ for $\delta = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n \pi_i^* x_i$*
- *(b) $\pi_i^* = \pi_i$ if and only if $\pi_i^* = x_i'\delta$*
- *(c) $\omega_i \equiv \omega_i(\boldsymbol{\pi}, 1) = -\omega_i(\boldsymbol{\pi}, 0) = \frac{\pi_i(1-\pi_i)}{\frac{1}{n}\sum_{i=1}^n \pi_i(1-\pi_i)}$. When $\pi_i^* = \pi_i$, $\omega_i \geq 0$.*

(2) (Goldsmith-Pinkham et al., 2024, $k = J > 1, T = 1$) Suppose $\mathcal{W} = \{0, 1, \ldots, J\}$ for $J > 1$ and x_i includes a constant. Let $W_{ij} = \mathbb{1}(W_i = j)$. For $Y_i = \sum_{j=1}^J \tau_j W_{ij} + x_i'\gamma + \epsilon_i$ and $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_J]'$ the coefficient of interest:

- *(a) $\pi_i(j) = x_i'\delta_j$ for $\delta_j = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n \pi_i^*(j) x_i$*
- *(b) $\pi_i^* = \pi_i$ if and only if $\pi_i^*(j) = x_i'\delta_j$ for all $j \in [J]$*
- *(c) The implicit estimand is shown in (B.2). This estimand is generally contaminated (that is, $\omega_{ij}(\boldsymbol{\pi}, \ell) \neq 0$ for some $j \in [J]$ and $\ell \notin \{0, j\}$).*

(3) (Miratrix et al., 2013; Imbens and Wooldridge, 2009; Lin, 2013, $k = T = J = 1$) Suppose x_i represents saturated discrete covariates taking values in $\{0, \ldots, L\}$, in the sense that $x_i = [x_{i1}, \ldots, x_{iL}]'$ for $x_{i\ell} = \mathbb{1}(x_i = \ell)$. Let $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. For $Y_i = \alpha_0 + \gamma_1' x_i + \tau W_i + W_i(x_i - \overline{x})'\gamma_2 + \epsilon_i$ and τ the coefficient of interest:

(a) *For all units i whose covariate value is ℓ (i.e., $x_{i\ell} = 1$), the implicit design $\pi_i = \pi_\ell$, where* $\pi_\ell = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(x_i = \ell)$

(b) $\pi_i^* = \pi_i$ *if and only if* π_i^* *is the same for all units with the same* x_i*-value*

(c) *The implicit estimand is the ATE. That is,* $\omega_i = \omega_i(\boldsymbol{\pi}, 1) = -\omega_i(\boldsymbol{\pi}, 0) = 1$.

(4) *(Kline, 2011,* $k = T = J = 1$*) Let* $\overline{x}_1 = \frac{\sum_{i=1}^n \pi_i^* x_i}{\sum_{i=1}^n \pi_i^*}$. *For* $Y_i = \alpha_0 + \gamma_1' x_i + \tau W_i + W_i(x_i - \overline{x}_1)'\gamma_2 + \epsilon_i$ *and* τ *the coefficient of interest:*

(a) $\pi_i = \frac{\delta_0 + (x_i - \overline{x})'\delta_1}{1 + \delta_0 + (x_i - \overline{x})'\delta_1}$, *where* δ_0, δ_1 *are equal to the population projection coefficients of* $\pi_i^*/(1-\pi_i^*)$ *on* $x_i - \overline{x}$ *and a constant,* weighted by $1 - \pi_i^*$

(b) $\pi_i^* = \pi_i$ *if and only if* $\pi_i^*/(1-\pi_i^*) = \delta_0 + \delta_1'(x_i - \overline{x})$

(c) *When* $\boldsymbol{\pi}^* = \boldsymbol{\pi}$, *the implicit estimand is the ATT. That is,* $\omega_i = \omega_i(\boldsymbol{\pi}, 1) = -\omega_i(\boldsymbol{\pi}, 0) = \frac{\pi_i}{\frac{1}{n}\sum_{i=1}^n \pi_i}$.

(5) *(Athey and Imbens, 2022,* $T > 1$*) Let* $\mathcal{W} \subset \{0,1\}^T$ *be the set of treatment paths such that its nonzero elements are linearly independent vectors whose span excludes the constant vector* $1_T = [1, \ldots, 1]'$.[24] *For the two-way fixed effects (TWFE) specificaiton* $\mathbf{Y}_{it} = \alpha_i + \mu_t + \tau \mathbf{W}_{it} + \epsilon_{it}$ *and the coefficient of interest* τ*:*

(a) *The implicit design is constant in* i *and is unique,* $\pi_i(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \pi_i^*(\mathbf{w})$

(b) $\pi_i^* = \pi_i$ *if and only if* π_i^* *is the same for all* i

(c) *The implicit estimand is shown in* (B.4)*, which matches Theorem 1(ii) in Athey and Imbens (2022) under staggered adoption.*[25]

Theorem 3.1 computes implicit designs and estimands for specifications considered in the literature. The literature analyzes each specification under some restrictions on the design and provides the corresponding estimand. Through our machinery, in each case, the implicit design exists, is unique, and takes exactly the form studied; the implicit estimand is exactly the corresponding estimand. In this sense, Theorem 3.1 is a set of converses to the existing results—the regression estimand is a causal contrast under $\boldsymbol{\pi}^*$ *only if* $\boldsymbol{\pi}^* = \boldsymbol{\pi}$ and the desired causal contrast is the implicit estimand.

In lieu of discussing each vignette in Theorem 3.1 in detail, we highlight a few particularly novel parts and extensions. First, Theorem 3.1(4) shows that Kline (2011)'s regression

[24] Staggered adoption corresponds to the case $\mathbf{w}_t \leq \mathbf{w}_{t+1}$ for all t. If $\mathcal{W}$ satisfies staggered adoption and excludes always-treated units, then it contains linearly independent vectors whose span excludes 1_T.

[25] One might wish to impose that the post-treatment weights are nonnegative (i.e., $\boldsymbol{\omega}_t(\boldsymbol{\pi}^*, \mathbf{w}) \geq 0$ if $\mathbf{w}_t = 1$). Failure of this condition implies that post-treatment units are severely used as comparisons for newly treated units, echoing the "forbidden comparison" issue in the recent difference-in-differences literature (Roth *et al.*, 2023; Borusyak *et al.*, 2024b; De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021). Proposition C.14 shows that when $\mathcal{W}$ only has two elements and includes a never treated unit, all weights post treatment are non-negative, but such forbidden comparisons are possible in *all* other cases.

implicitly models the propensity odds $\pi_i^*/(1-\pi_i^*)$ as linear.[26] Interestingly, Theorem 3.1(4) also shows how the regression implicitly estimates this model, since there is no obvious sample analogue to $\pi_i^*/(1-\pi_i^*)$. The coefficients δ_0, δ_1 in $\pi_i/(1-\pi_i)$ are population regression coefficients of the odds $\pi_i^*/(1-\pi_i^*)$ on the covariates, *weighted by* $1-\pi_i^*$: For instance, for $\overline{x}_0 = \sum_i (1-\pi_i^*)x_i / \sum_i (1-\pi_i^*)$, δ_1 is given by

$$\delta_1 = \left(\frac{1}{n} \sum_{i=1}^{n} (1-\pi_i^*)(x_i - \overline{x}_0)(x_i - \overline{x}_0)' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} (1-\pi_i^*) \frac{\pi_i^*}{1-\pi_i^*} (x_i - \overline{x}_0).$$

Despite not observing $\pi_i^*/(1-\pi_i^*)$, the coefficients (δ_0, δ_1) can be consistently estimated, since the denominator of $\pi_i^*/(1-\pi_i^*)$ cancels with the weighting by $(1-\pi_i^*)$.

Second, Theorem 3.1(3) and (4) leave a few puzzles, which the next subsection resolves. The regressions in both can be viewed as taking a contrast in the interacted regression $Y_i = \gamma_0 + \gamma_1' x + \tau_0 W_i + \tau_1' x_i W_i + \epsilon_i$: Theorem 3.1(3) takes $\tau_0 + \tau_1' \overline{x}$ while Theorem 3.1(4) takes $\tau_0 + \tau_1' \overline{x}_1$. Curiously, Theorem 3.1(3) asks that the covariates be saturated, while Theorem 3.1(4) has no such requirement. Moreover, Theorem 3.1(4) can be applied to generate a result for the model-based average treatment effect on the untreated (ATU) estimand $\tau_0 + \tau_1' \overline{x}_0$, where $\overline{x}_0 = \frac{\sum_i (1-\pi_i^*) x_i}{\sum_i (1-\pi_i^*)}$. Doing so shows that $\tau_0 + \tau_1' \overline{x}_0$ is a causal contrast only if the *reciprocal propensity odds* $(1-\pi_i^*)/\pi_i^*$ is linear in x_i. Worryingly, despite generated by the same specification, the ATU thus seems to require a *different* design than the ATT.

Finally, Theorem 3.1(5) shows a converse to Athey and Imbens (2022): the TWFE estimand τ fails to satisfy level irrelevance *unless* the treatment timing is totally randomized—a design that Athey and Imbens (2022) study. Our subsequent results in Section 3.3 enrich Theorem 3.1(5) by showing that the design-based interpretation of TWFE, already only possible under a stringent design, is *fragile* against two common complications—time-varying covariates and imbalanced panel. Appendix C.5 further provides extensions to one-way fixed effects and event-study specifications.

[26] There are two subtleties in our presentation of this regression. First, the average treatment effect on the treated for this finite population is defined to be

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\pi_i^*}{\frac{1}{n}\sum_{i=1}^{n} \pi_i^*} (y_i(1) - y_i(0)).$$

One could view this as the finite-population analogue of the superpopulation ATT under unconfoundedness:

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid W_i = 1] = \mathbb{E}\left[\frac{\pi(X_i)}{\mathbb{E}[\pi(X_i)]} (Y_i(1) - Y_i(0)) \right].$$

Second, this regression specification has a covariate transform that depends on $\overline{x}_1$, which involves π_i^*. Thus assuming that the covariate transform is known in population also amounts to ignoring the uncertainty in estimating $\overline{x}_1$ when we estimate it using $\sum_i W_i x_i / \sum_i W_i$.

3.2. **Forbidden interactions?** Suppose $T = J = 1$. Let x_{1i}, x_{2i} be subvectors of x_i, which may be disjoint, nested, identical, or overlapping. Consider the specification

$$Y_i = \gamma_0 + \tau_0 W_i + \tau_1' W_i x_{1i} + \gamma_1' x_{2i} + \epsilon_i. \tag{3.1}$$

Read as a linear outcome model (i.e., $\mathbb{E}[Y_i(w) \mid x_i] = \gamma_0 + \gamma_1' x_{2i} + w(\tau_0 + \tau_1' x_{1i})$), this specification models the baseline potential outcome as linear in x_{2i} and the conditional average treatment effect as linear in x_{1i}, $\mathbb{E}[Y_i(1) - Y_i(0) \mid x_i] = \tau_0 + \tau_1' x_{1i}$. Under this outcome model, $\tau_0 + \tau_1' x_{1i}$ is a causal contrast for all x_{1i}. Since the covariates $(1, x_{1i}')$ are not collinear, $\boldsymbol{\tau} = (\tau_0, \tau_1')$ are themselves vectors of causal contrasts.

It is tempting to think that the causal contrast interpretation of $\boldsymbol{\tau}$ exists independently of the outcome model. Perhaps $x_1 \mapsto \tau_0 + \tau_1' x_1$ is some linear approximation of the conditional average treatment effects. Thus, one might conjecture that without the outcome model but with the right design, $\tau_0 + \tau_1' x_1$ is nevertheless some weighted average of conditional average treatment effects evaluated at values around x_1.

Surprisingly, the next proposition shows that this is in fact implausible without a randomized experiment or saturating the covariates. It turns out that the causal interpretation of $\boldsymbol{\tau}$ necessitates that *both* π_i^* and $\pi_i^* x_{1i}$ are linear in x_{2i}. When this fails, some contrast $\tau_0 + \tau_1' x_1$ does not satisfy level irrelevance.[27]

Proposition 3.2. *Consider the specification* (3.1) *and let* $\boldsymbol{\tau} = (\tau_0, \tau_1')'$ *be the coefficients of interest. Then its corresponding implicit design exists if and only if, for some conformable* Γ_0, Γ_1, *for all* i,

$$(\delta_0 + \delta_1' x_{2i}) x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$$

where δ_0, δ_1 *are population projection coefficients of* π_i^* *on* x_{2i}*:* $\pi_i^* = \delta_0 + \delta_1' x_{2i} + \eta_i$ *and* $\pi_i^* x_{1i} = \Gamma_0 + \Gamma_1 x_{2i} + \upsilon_i$*. When this happens, the unique implicit design is* $\pi_i = \delta_0 + \delta_1' x_{i2}$*. Therefore, if* $\boldsymbol{\tau}$ *is a vector of causal contrasts under* $\boldsymbol{\pi}^*$*, then* $\pi_i^* = \delta_0 + \delta_1' x_{2i}$ *and* $\pi_i^* x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$.

The necessary condition for interpreting $\boldsymbol{\tau}$ as a *vector* of causal contrasts is that *both* the propensity score π_i^* and its interaction with the covariates $\pi_i^* x_{1i}$ are linear functions of x_{2i}. When x_{1i} is included in x_{2i}, this condition is unlikely to hold in general, as $\pi_i^* x_{1i}$ would involve nonlinear transformations of x_{1i} and thus cannot be linear. Two common exceptions are available. First, this condition can hold if the propensity score is a constant function of x_{2i} (i.e. $\delta_1 = 0$). Second, if x_{1i} represents a saturated categorical variable and

[27] Concurrent work by Zhao *et al.* (2025) provides a similar result.

x_{2i} contains all other covariates interacted with x_{1i},[28] then entries of $x_{1i}x_{2i}'$ are in the linear span of x_{2i} and the condition reduces to π_i^* being linear in x_{2i}.

When x_{1i} is a scalar variable, τ_1 is often interpreted (e.g., in Blakeslee *et al.* (2020)) as some approximation of treatment effect heterogeneity comparing units with larger x_{1i} to units with smaller x_{1i}, and τ_0 is often interpreted as a baseline treatment effect for units with $x_{1i} = 0$ (Hainmueller *et al.*, 2019). Proposition 3.2 circumscribes the design-based interpretation of these treatment effect contrasts. If x_{1i} is binary, and x_{2i} includes x_{1i} and its interactions with all other covariates, then $\boldsymbol{\tau}$ is a causal contrast only under linear propensity scores. On the other hand, if x_{1i} is a continuous covariate included in x_{2i}, then design-based interpretation requires both linear propensity scores and—implausibly—that the nonlinear quantity $\delta_2' x_{2i} x_{1i}$ is in the linear span of x_{2i}.

One way to understand this difficulty is by reasoning through why $\tau_0 + \tau_1' x_{1i}$ is not necessarily a best linear approximation of the conditional average treatment effect, even when we think of the units as sampled from a superpopulation. One could think of (3.1) as two regressions, one on the treated $W = 1$ and one on the untreated $W = 0$. Both regressions are indeed best linear approximations to $\mathbb{E}[Y(1) \mid x, W = 1]$ and $\mathbb{E}[Y(0) \mid x, W = 0]$, which are equal to the mean potential outcomes $\mathbb{E}[Y(1) \mid x], \mathbb{E}[Y(0) \mid x]$ under unconfoundedness. The contrast $\tau_0 + \tau_1' x_1$ is then the difference of the fitted values of these two regressions. Importantly, however, the two regressions are only best linear approximations with respect to *different distributions* of the covariates ($x_i \mid W = 1$ vs. $x_i \mid W = 0$). Thus, their difference is not a best linear approximation to the conditional average treatment effect under any particular distribution of x. Shifting $Y(1)$ and $Y(0)$ by the same arbitrary amount therefore causes asymmetric behavior in the two regressions, leading to a failure of level irrelevance.

When $x_{1i} = x_{2i} = x_i$, this result supplements Theorem 3.1(3)–(4) by cementing the intuition that different contrasts of the interacted regression have design-based interpretations under *different* designs. These designs are often logically incompatible, unless the covariates are appropriately saturated or if the propensity score is constant. Indeed, insisting on all such contrasts being causal imposes a knife-edge condition for the design. As a manifestation of this phenomenon, Theorem 3.1(4) shows that the model-based ATT $\tau_0 + \tau_1' \overline{x}_1$ maintains its interpretation under the design whose odds is linear, but the ATU requires that the reciprocal propensity odds is instead linear. Theorem 3.1(3) evades this difficulty by requiring that the covariates are saturated.

[28] That is, x_{1i} contains mutually exclusive binary random variables, and x_{2i} contains x_{1i}, some set of other covariates x_{3i}, and all interactions $x_{3ik}x_{1i\ell}$.

One might wonder what can be said of the model-based ATE $\tau_0 + \tau_1'\overline{x}$ without saturating the covariates (cf. Theorem 3.1(3)). Theorem 1 in Chattopadhyay and Zubizarreta (2023) shows that, for this contrast to actually be the ATE under a design $\boldsymbol{\pi}^*$, $\boldsymbol{\pi}^*$ needs both $(1-\pi_i^*)/\pi_i^*$ and $\pi_i^*/(1-\pi_i^*)$ to be linear in the covariates, which is implausible unless π_i^* is constant or the covariates are saturated. If we simply require $\tau_0 + \tau_1'\overline{x}$ to be a causal contrast—but not necessarily the ATE—then the implicit design does exist and is equal to some fractional linear function $\pi_i = \frac{\theta_0+\theta_1'(x-\overline{x})}{1-\Gamma_2'(x-\overline{x})}$. The requirement that $\pi_i = \pi_i^*$ then imposes additional restrictions on $(\theta_0, \theta_1, \Gamma_2)$. These constraints are unpleasant in the sense that they depend on the values $x_{1:n}$: The same design $\pi_i^* = \frac{\theta_0+\theta_1'(x_i-\overline{x})}{1-\Gamma_2'(x_i-\overline{x})}$ may be consistent with design-based interpretation of $\tau_0 + \tau_1'\overline{x}$ for one set of covariates $x_{1:n}$ but not for another $\tilde{x}_{1:n}$, even though $\pi_i^* \in (0,1)$ under both sets of covariates.[29] These issues are formalized in Remark C.1 in Appendix C.4.

One way to read these results is that, without a randomized experiment, the interacted regression is *less* robust than the uninteracted regression (2.1), in terms of its design-based interpretation.[30] The uninteracted regression introduces variance weighting for the estimand, but maintains validity under a simple design. The interacted regression removes the weighting when the covariates are saturated, but estimates a weighted average treatment effect under a bizarre design in general.

Is there a simple regression that both removes the variance weighting and is design-based under linear propensity scores? Unfortunately, Proposition C.11 shows that the answer is *no*, at least not with right-hand-side variables that are linear in $[1, x_i, W_i, W_i x_i]$.[31] As a result, targeting the average treatment effect under the same implicit design as (2.1) necessitates moving beyond simple regression estimators.

3.3. **The fragility of design-based TWFE.** Theorem 3.1(5) shows that the standard TWFE regression has design-based interpretation only under total randomization of the treatment paths. This design is restrictive, and we may wonder whether it is possible to modify the regression so that designs are allowed to be functions of observable information. The most straightforward idea—indeed common empirical practice—is to include time-varying covariates in a TWFE regression.

Surprisingly, doing so can often *destroy* the design-based interpretation of the corresponding estimand altogether. In panel settings, usually the equations defining the implicit

[29] Note that the linear-odds parametrization of Kline (2011) is not unpleasant in this way.

[30] This is in contrast to Lin (2013)'s arguments against Freedman (2008). Lin (2013), along with Negi and Wooldridge (2021), shows that the interacted regression is in many ways superior in a randomized experiment.

[31] One could estimate the uninteracted regression and weigh by $1/(\pi_i(1-\pi_i))$ to remove the variance weighting, but this approach requires estimating π_i separately.

design are overdetermined, and so small modifications to the potential weights can make the system of equations unsolvable. With time-varying covariates, this happens precisely when the covariates matter for identification and change the estimand. On the other hand, when adding the covariates do not change the estimand, constant propensity scores remain the unique implicit design, and thus the covariates do not enrich the design either.

Proposition 3.3. *Assume $\mathcal{W} \subset \{0,1\}^T$. Consider the regression $\mathbf{Y}_{it} = \alpha_i + \gamma_t + \tau \mathbf{W}_{it} + \delta' \mathbf{x}_{it}$ where τ is the coefficient of interest. Let $\beta_{w \to x}$ be the population projection coefficient of W_{it} on x_{it} under $\boldsymbol{\pi}^*$, including individual and time fixed effects. For the set of implicit designs to be non-empty, a necessary condition is*

$$\left(\mathbf{x}_i - \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j \right) \beta_{w \to x} \in \operatorname{span}(\mathcal{W} \cup \{1_T\}) \text{ for all } i \in [n].$$

When $\beta_{w \to x} = 0$, if $\mathcal{W}$ contains linearly independent vectors whose span excludes 1_T, the implicit design is uniquely equal to $\pi_i(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \pi_i^*(\mathbf{w})$ for all i as in Theorem 3.1(5).

Proposition 3.3 states that a necessary condition for the set of implicit designs under $\boldsymbol{\pi}^*$ to be nonempty is that a particular linear combination of the demeaned covariates sits in the linear span of $\mathcal{W}$ and 1_T for every unit. This condition arises because we essentially need that the mean treatment $\sum_{\mathbf{w} \in \mathcal{W}} \pi_i^*(\mathbf{w}) \mathbf{w}$ is *exactly* described by two-way fixed effects with time-varying covariates, analogous to the intuition for (2.1). This then restricts the space of the covariate vectors, since they need to generate vectors that lie in the linear span of $\mathcal{W}$.

Under staggered adoption, $\operatorname{span}(\mathcal{W} \cup \{1_T\})$ is the set of vectors that are piecewise constant between adjacent adoption dates. This subspace is highly restrictive if there are relatively few adoption dates. If $\beta_{w \to x} \neq 0$, it is thus knife-edge that $\left(\mathbf{x}_i - \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j \right) \beta_{w \to x}$ happens to be located in that subspace, unless columns of $\mathbf{x}_i$ happens to be piecewise constant over t as well.[32] On the other hand, if $\beta_{w \to x}$ is zero under $\boldsymbol{\pi}^*$, including the covariates makes no difference to the coefficient on W_{it} in population. Therefore, Proposition 3.3 implies that TWFE with time-varying covariates does not in general have a design-based interpretation. Researchers using such a specification either believe that the covariates do not affect treatment assignment and are thus immaterial to identification from a design-based perspective, or they are implicitly embedding certain outcome modeling assumptions.

Proposition 3.3 shows that the design-based interpretation for TWFE is delicate, and seemingly perfunctory modifications to the regression can destory it altogether. In a similar

[32] This *is* plausible if the time-varying covariates are interactions of fixed covariates with the time fixed effects ($\mathbf{x}_{it}'\delta = x_i'\delta_t$). Proposition C.15 shows that for this specification, causal interpretation is possible necessarily under linear generalized propensity scores $\pi_i(\mathbf{w}) = \delta_0(\mathbf{w}) + \delta_1(\mathbf{w})' x_i$.

vein, TWFE on an imbalanced panel also has this feature. Such a regression is only design-based when the missingness pattern is uncorrelated with the treatment assignment pattern, in which case the design must again be total randomization of treatment paths. To introduce this result, suppose units are observed in some subperiod $\mathcal{T}_i \subset \{1, \dots, T\}$. We consider the TWFE specification

$$\mathbf{Y}_{it} = \alpha_i + \gamma_t + \tau \mathbf{W}_{it} + \epsilon_{it} \tag{3.2}$$

over $i = 1, \dots, n$ and $t \in \mathcal{T}_i$. The objects $\mathcal{T}_1, \dots, \mathcal{T}_n$ are considered fixed and observed.

Let $\mathcal{W}_{\mathcal{T}}$ be a multiset collecting the non-zero treatment paths restricted to some subperiod $\mathcal{T} \subset [T]$. That is, $\mathcal{W}_{\mathcal{T}}$ collects the vectors $(\mathbf{w}_t : t \in \mathcal{T})$ for all $\mathbf{w} \in \mathcal{W} \setminus \{0_T\}$, possibly with repetition. We say that $\mathcal{W}$ has *rich variation* on $\mathcal{T}$ if the vectors in $\mathcal{W}_{\mathcal{T}}$ are linearly independent and their linear span excludes the vector of all ones.[33]

Proposition 3.4. *Let the population regression specification be characterized by (3.2) under $\boldsymbol{\pi}^*$, where τ is the coefficient of interest. Suppose $\mathcal{W}$ satisfies staggered adoption and excludes always-treated units. Assume further that there is a commonly observed period $\mathcal{T} \subset \bigcap_{i=1}^n \mathcal{T}_i$ such that the treatment paths have rich variation in $\mathcal{T}$.*

Let $\overline{\pi}(\mathbf{w}) \equiv \frac{1}{n}\sum_{i=1}^n \pi_i^*(\mathbf{w})$. Let $Q_i(\mathbf{w}) = \frac{\sum_{t \in \mathcal{T}_i} \mathbf{w}_t}{|\mathcal{T}_i|}$ be the proportion of treated periods within unit i's observed period for treatment path $\mathbf{w}$. Then an implicit design exists if and only if $\boldsymbol{\pi}^*$ is uncorrelated with the missingness pattern in the sense that for all $t \in [T]$:

$$\underbrace{\sum_{\mathbf{w} \in \mathcal{W}} \overline{\pi}(\mathbf{w}) \sum_{i: t \in \mathcal{T}_i} (\mathbf{w}_t - Q_i(\mathbf{w}))}_{\sum_{i: t \in \mathcal{T}_i} \mathbb{E}_{\mathbf{w} \sim \overline{\pi}}[(\mathbf{w}_t - Q_i(\mathbf{w}))]} = \underbrace{\sum_{\mathbf{w} \in \mathcal{W}} \sum_{i: t \in \mathcal{T}_i} \pi_i^*(\mathbf{w})(\mathbf{w}_t - Q_i(\mathbf{w}))}_{\sum_{i: t \in \mathcal{T}_i} \mathbb{E}_{\mathbf{w} \sim \pi_i^*}[(\mathbf{w}_t - Q_i(\mathbf{w}))]}. \tag{3.3}$$

When this happens, the implicit design is unique and satisfies $\pi_i(\mathbf{w}) = \overline{\pi}(\mathbf{w})$ for all i.

Proposition 3.4 gives a necessary and sufficient condition for an implicit design to exist, and characterizes the implicit design when it does. In short, an implicit design exists if and only if the missingness patterns are uncorrelated with the treatment timing in a particular sense (3.3). When it exists, the implicit design is the same for all units. Therefore, if $\boldsymbol{\pi}^*$ justifies τ as a design-based estimand in the sense of Definition 2.4, then $\boldsymbol{\pi}^*$ must equal to the implicit design and thus must describe random assignment.

The condition (3.3) describes a situation where for any given time period t, among units for whom t is observed, the average expected de-meaned treatment path under $\boldsymbol{\pi}^*$,

[33] As an example, note that if $\mathcal{W}$ satisfies staggered adoption and excludes an always treated unit, then the period that spans all adoption dates, $\mathcal{T} = \{t_{\min} - 1, \dots, t_{\max}\}$—where $t_{\min}$ is the first adoption date and $t_{\max}$ is the last adoption date, is a subperiod on which $\mathcal{W}$ has rich variation.

$\mathbb{E}_{\mathbf{w}\sim\pi_i^*}[\mathbf{w}_t - Q_i(\mathbf{w})]$, is equal to its analogue under the random assignment $\overline{\pi}$. This condition is difficult to satisfy unless the design is constant $\pi_i^*(\mathbf{w}) = \overline{\pi}(\mathbf{w})$ in the first place. As a result, Proposition 3.4 shows that TWFE is fragile to imbalanced panels.

4. Empirical application

We illustrate several empirical recommendations (Section 2.3) by reanalyzing Blakeslee *et al.* (2020). Blakeslee *et al.* (2020) study the economic impact of water loss in rurual India, using borewell failure (W_i) as a quasi-experimentally assigned treatment. They consider a range of income and employment outcomes and conclude that (i) well failure causes a decline in agricultural income and employment, but reallocation to off-farm offsets the lost income, and (ii) those living in high economic development areas adapt more easily.

The main empirical analyses of Blakeslee *et al.* (2020) contain estimates of the following regression (their equation (1) reproduced in our notation): For a household i, let $v(i)$ denote its village and $t(i)$ denote the year when it drilled its first borewell. Consider the coefficient τ in

$$Y_i = \tau W_i + \underbrace{\alpha_0 + x_i'\gamma + \eta_{1v(i)} + \eta_{2t(i)}}_{\check{x}_i'\mu} + \epsilon_i. \tag{4.1}$$

where x_i collects a vector of household characteristics and includes "the age, caste, and literacy of the household head, as well as the total land inherited by the household" (p.208). By Theorem 3.1(1) (Blandhol *et al.*, 2025; Angrist, 1998), τ is a causal contrast only if π_i^* is linear in $\check{x}_i$. When that happens, the estimand is a weighted average treatment effect with weights proportional to $\pi_i^*(1-\pi_i^*)$.

In their heterogeneity analysis, Blakeslee *et al.* (2020) consider the regression that interacts (4.1) with an indicator h_i of whether the village $v(i)$ has high economic development

$$Y_i = \tau_0 W_i + \tau_1 W_i h_i + \underbrace{(\alpha_0 + x_i'\gamma_0 + \eta_{1v(i)} + \eta_{2t(i)} + \alpha_1 h_i + h_i x_i'\gamma_1 + \kappa_{2t(i)} h_i)}_{\tilde{x}_i'\mu} + \epsilon_i, \tag{4.2}$$

where τ_0 is interpreted as a treatment effect for those with $h_i = 0$ and τ_1 is interpreted as a difference of treatment effects among $h_i = 1$ versus $h_i = 0$. Our result in Proposition 3.2 show that τ_0 and τ_1 in (4.2) are both causal contrasts only if π_i^* is linear in $\tilde{x}_i$ and $\pi_i^* h_i$ is also linear in $\tilde{x}_i$. Because h_i is in the linear span of the village fixed effects, it is easy to check that $\pi_i^* h_i$ is linear if π_i^* is linear—and thus the implicit design exists. In this case, we may also compute and verify that the corresponding estimand τ_1 takes the form of a

difference of two variance-weighted estimands:

$$\tau_1 = \frac{1}{n_1} \sum_{i:h_i=1} \omega_{1i}(y_i(1) - y_i(0)) - \frac{1}{n_0} \sum_{i:h_i=0} \omega_{0i}(y_i(1) - y_i(0))$$

such that $\frac{1}{n_1} \sum_{i:h_i=1} \omega_{1i} = \frac{1}{n_0} \sum_{i:h_i=0} \omega_{0i} = 1$, and $\omega_{1i}, \omega_{0i} \propto \pi_i^*(1 - \pi_i^*)$.

Interestingly, the implicit designs underlying (4.1) and (4.2) are different, as (4.2) involves additional covariates. Since the specifications are nested, one can test $H_0 : \pi_i^*$ is linear in $\check{x}_i$ against $H_1 : \pi_i^*$ is linear in $\tilde{x}_i$. A standard Wald test does reject H_0 in favor of H_1, providing evidence against the implicit design for (4.1). In the remainder of the section, we focus on the specification (4.2), as well as the version of (4.1) that includes $\tilde{x}_i$ instead of $\check{x}_i$ as the covariates:

$$Y_i = \tau W_i + \tilde{x}_i' \mu + \epsilon_i. \tag{4.3}$$

Specifications (4.2) and (4.3) in turn have the same implicit design.

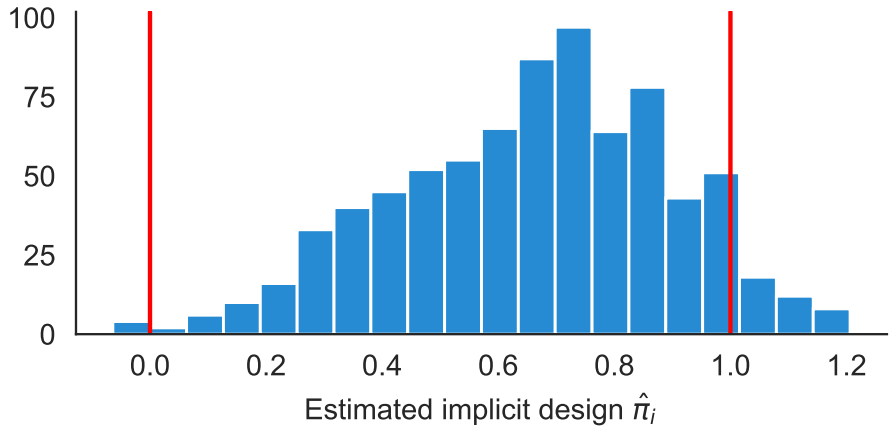


FIGURE 1. The distribution of the estimated implicit design for specification (4.2). 56 out of 786 observations (or 7% of the observations) are outside $[0, 1]$.

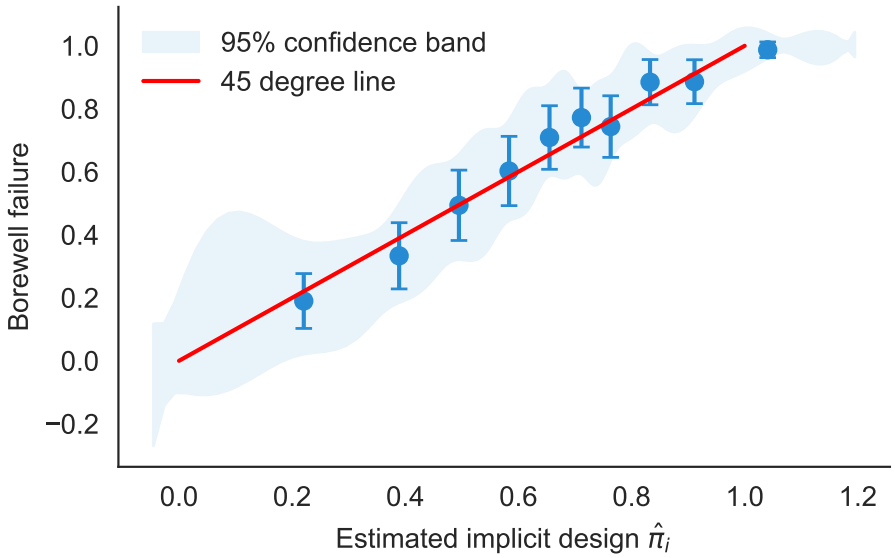
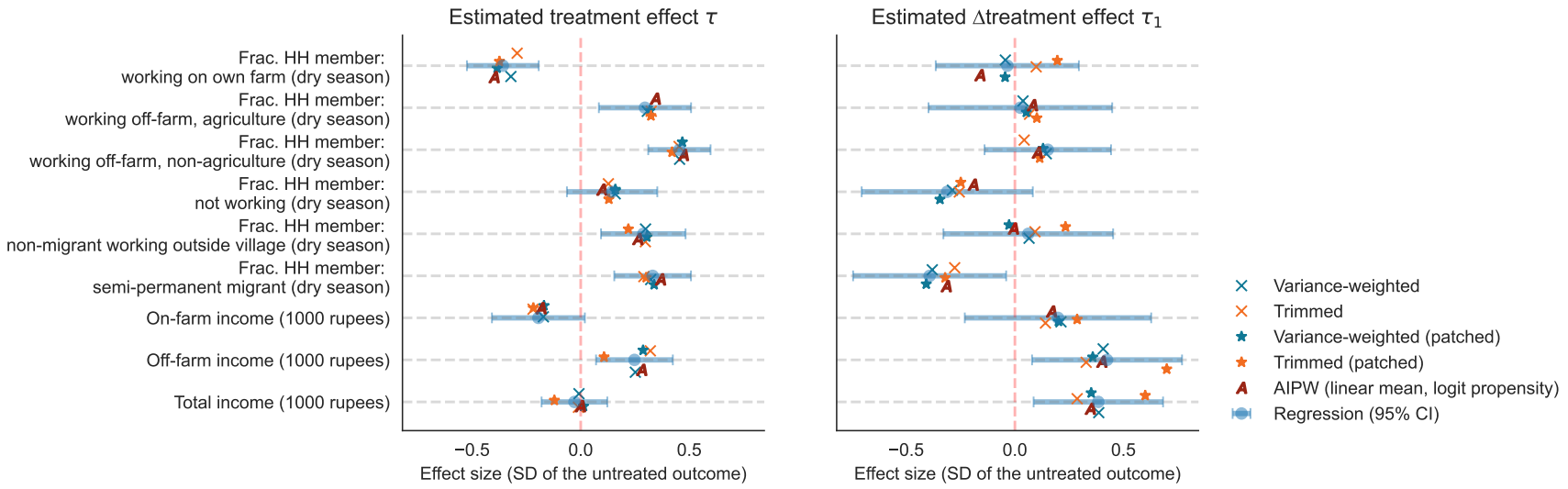


FIGURE 2. Calibration of the implicit design. This is a binned scatterplot of W_i on $\hat{\pi}_i$, with associated pointwise confidence intervals and uniform confidence bands (Cattaneo *et al.*, 2024).



Notes. "Variance-weighted" refers to weighting with weights proportional to $\pi_i(1-\pi_i)$. "Trimmed" refers to weighting with weights proportional to $\mathbb{1}(\pi_i \in [0.02, 0.98])$ (Crump *et al.*, 2009). In either weighting scheme, the weights for $\pi_i \notin [0,1]$ are set to zero. "Patched" uses the procedure outlined in Section 2.3, where we subclassify on the implicit design so as to recalibrate the implicit propensity scores (Chapter 17, Imbens and Rubin, 2015). The binning in the subclassification uses the data-driven procedure in Imbens and Rubin (2015), which recursively partitions the estimated propensity scores until either bins are too small or the mean propensity score is similar among treated and untreated units within a bin. "AIPW" refers to augmented inverse propensity weighting; it uses a logit specification for the propensity score and a linear specification for the outcome means. Both specifications use the same set of covariates as (4.3). Finally, "regression" estimates (4.3) and (4.2) with ordinary least-squares, clustering standard errors by village following Blakeslee *et al.* (2020) (The estimates in Blakeslee *et al.* (2020) use weighted least squares with unit-level sampling weights; since their footnote 7 reports that their results are insensitive to these weights, we ignore them and report ordinary least-squares estimates). □

FIGURE 3. Alternative coefficient estimates for τ in (4.3) and τ_1 in (4.2)

Figure 1 plots the estimated implicit design. A small but substantial portion of units have estimated implicit designs that are outside of $[0, 1]$, which immediately raises questions about the design-based interpretation of (4.2). Calibration performance of the implicit design is reasonable, as shown in Figure 2; however, Ramsey's RESET test against the linearity in $\tilde{x}_i$ does decisively reject (p-value: 0.00). In terms of predictiveness, the implicit design accounts for about 25% of variation in W_i, indicating that observable characteristics of households do predict treatment. Consistent with Blakeslee *et al.* (2020)'s explanation, most of the predictive power comes from the village and drill-time fixed effects (the within-R^2 is only 0.8%). Blakeslee *et al.* (2020) (p.220) worry about selection on unobserved confounders, most plausibly "wealthier and more skilled farmers being less likely to experience borewell failure." The estimated implicit design from their specification does appear to rule out this type of selection—among those with various agricultural equipment pre-drilling versus those without, the mean estimated implicit designs are remarkably similar.[34]

These diagnostics on the implicit design undermine a design-based interpretation of the specification (4.2), answering (Q1) in the negative. However, at least as an illustration, it may be interesting to understand the regression specification further. Even ignoring the various deficiencies of $\hat{\boldsymbol{\pi}}$, we may still be concerned with various choices in the implicit estimand. For instance, the specification (4.3) includes variance weighting, and we may ask—holding the propensity model fixed—how much the variance weighting alters the estimate. The estimand τ_1 in (4.1) is a difference of two variance-weighted estimands. A priori, we cannot rule out that the difference is driven by the weighting scheme compared to the difference in average treatment effects. The remainder of this section investigates these questions.

We can simply compute a battery of alternative estimates, using the implicit design. These are shown in Figure 3 for both the treatment effect τ in (4.3) and the treatment effect difference τ_1 in (4.2). First, the simplest assessment is whether the units with out-of-bounds implicit design feature prominently in the regression estimate. To that end, the variance-weighted estimates $\times$ uses the same implicit design and weighting scheme as the implicit estimand, but removes the out-of-bounds units. For both τ_0 and τ_1, these $\times$ estimates are almost identical to the regression estimates, indicating that the regression estimates put

[34] We assess this by regressing $\hat{\pi}_i$ on indicators for whether a household owns a $\{\text{tractor}, \text{seed drill}, \text{thresher}, \text{motorcycle}\}$ before they drilled their first borewell. None of these covariates, jointly or separately, is statistically significant at the conventional level. The largest $|t|$-statistic among these is 1.04. These covariates are not included in the specification (4.2).

relatively little weight on out-of-bounds units. Removing out-of-bounds units by design-patching the implicit design $\star$ also does not meaningfully alter the estimate.

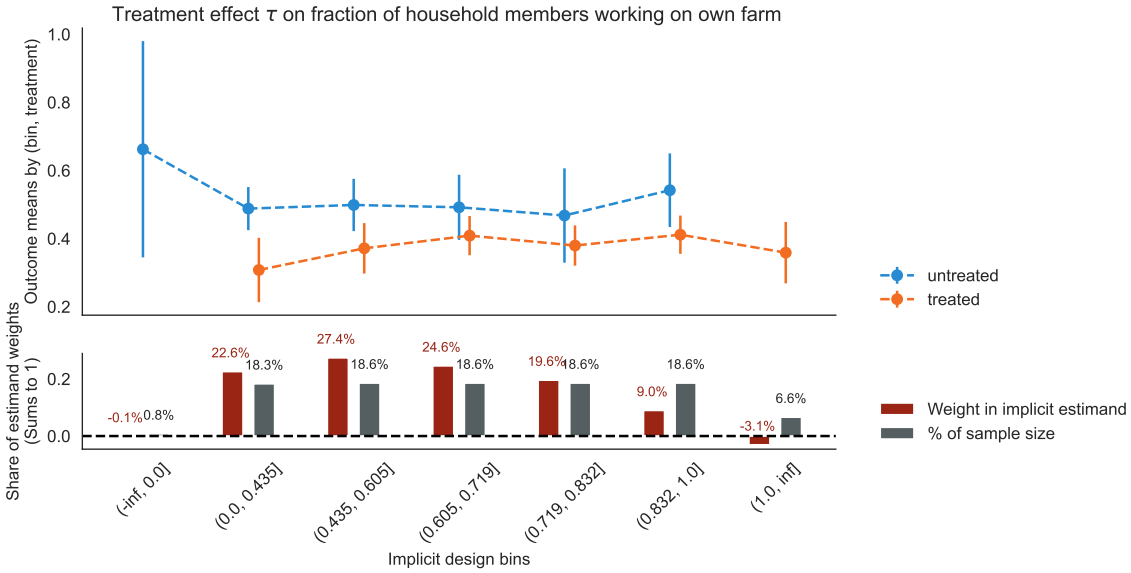
Second, we may assess whether the variance-weighting in the implicit estimand matters by considering weighting schemes that treat units more equally. Since the propensity score estimates are often close or equal to zero and one, overlap violations make estimating the average treatment effect infeasible. Thus, we trim the propensity scores to $[0.02, 0.98]$ and construct corresponding estimators for the trimmed average treatment effect (Crump *et al.*, 2009). These estimates—especially the patched estimates $\star$[35]—are more different from the regression estimates, though not substantially so compared to sampling noise. On average, the difference between $\star$ and the regression estimates is about 0.5–0.7 times the standard error of the regression estimate.

Finally, moving entirely away from the implicit designs in the regression, we also compute estimates by augmented inverse propensity weighting (*A*IPW) by using a simple logit model for the propensity scores and a linear model for the outcome means. These alternative estimates are again similar to the regression estimates, indicating that the outcomes in this application are not so adversarially configured: The implicit design, while clearly rejected, nevertheless produces estimates that are similar to alternative estimates.

We can also unpack why the alternative weighting schemes do not appear to make a difference by examining the correlation between the outcomes and the implicit design. Figure 4(a) partitions the implicit designs into 7 bins by quantile, so that each bin within $[0, 1]$ contains the same number of units. On each bin, it displays the mean treated and mean untreated outcomes as well as the weight placed on each bin by the implicit estimand. As the bin size become small, computing the difference between the blue and orange curves, weighted by the weights, approximates the regression estimate. As expected, the regression upweights units with implicit design closer to 0.5. In particular, the out-of-bounds units receive a total of about -3% in weights, explaining their apparent irrelevance to the bottom line. Remarkably, for agricultural employment, the treatment effects hardly vary by the implicit design. Households with low and high propensity to lose water access appear to be affected similarly, with about 10% fewer members engaged in agricultural work, suggesting borewell failure is orthogonal to vulnerability to water loss.

[35]The patched implicit design has no $\hat{\pi}_i^{\text{patched}}$ within $(0, 0.02) \cup (0.98, 1)$; thus the trimming only trims out those with overlap failure.

(a) Treatment effect τ in (4.3)



(b) Heterogeneous treatment effect τ_1 in (4.2)



FIGURE 4. The correlations of outcomes and implicit designs

Figure 4(b) considers the analogous decomposition for τ_1.[36] The weighting schemes for $h_i = 1$ versus $h_i = 0$ are indeed different in τ_1: High employment area ($h_i = 1$) puts larger weight for households more likely to lose water access—peaking at the bin $(0.605, 0.719]$

[36] In Figure 4(b), if everyone is treated in a bin, we treat the mean control outcome as zero, and vice versa. Thus the "treatment effects" for $\hat{\pi}_i \leq 0$ represent negative mean control outcomes, and the "treatment effects" for $\hat{\pi}_i > 1$ represent mean treated outcomes.

as opposed to at $(0.435, 0.605]$. Thus the apparent treatment effect heterogeneity reflects in part the difference in weighting. But since the differences in conditional average treatment effects are effectively constant and zero, the different weighting again makes little difference to the bottom line estimate.

Our exercise demonstrates a few ways to use implicit designs and estimands. One can examine the distribution and calibration properties of the estimated implicit design $\hat{\boldsymbol{\pi}}$ (Figures 1 and 2). One can use $\hat{\boldsymbol{\pi}}$ to construct implement alternative estimands that maintain the implicit model of treatment assignment (Figure 3). One can also examine how outcomes covary with observed selection (Figure 4). These exercises add transparency and evaluate robustness of the regression estimates, especially if practitioners, like Blakeslee *et al.* (2020), stake a claim on design-based causal identification.

5. Conclusion

Linear regressions are ubiquitous. Interpreting their results as causal, thanks to random assignment, is similarly commonplace. This paper studies the necessary implications of so interpreting. We do so by studying the comparisons that regression estimands make, in terms of potential outcomes, under random assignment. For these comparisons to admit a causal interpretation—in the absence of assumptions on the outcomes—the true process of treatment assignment must satisfy certain restrictions. Given the Gram matrix of a regression, these restrictions are linear in the design and often exactly pin down the distribution of treatment. The resulting set of designs that satisfy these restrictions can be thought of as models of treatment assignment that the regression implicitly estimates. Each design also pinpoints a corresponding estimand that the regression implicitly chooses.

Understanding design-based interpretation of regressions in this way essentially reduces to mechanical computations that can be scaled and automated. These computations can aid in examining new theoretical properties of particular specifications, itself the subject of a highly influential recent literature. In several theoretical vignettes, these computations unify and strengthen disparate strands of the literature. Additionally, we find that regressions with interactions and with two-way fixed effects have fragile design-based interpretations. This calls for caution and nuance when using them and presenting their results.

Directly computing implicit designs and estimands in practice provides a set of simple diagnostics for practitioners who wish to understand the design-based properties of a given regression. Doing so makes transparent the statistical and economic choices masked by a regression specification—especially when the regression is read simply as an algorithm for comparisons. Having opened up the black box, we can examine each of its components:

e.g., evaluating whether the implicit design is plausible, improving the implicit design if it has obvious deficiencies, assessing whether the regression targets an economically interesting estimand, and constructing estimates for alternative estimands. Additionally, making these implicit choices transparent may nudge design-minded practitioners to wrest the wheel from familiar methods and choose methods that are more natively design-based.

We conclude by speculating on design-based statistical inference following a regression. If the regression has a unique implicit design and is interpreted under random treatment assignment, it may be attractive to consider design-based inference—especially in settings where superpopulation sampling is difficult to conceptualize (e.g. spatial settings). Design-based inference allows one to quantify uncertainty without appealing to sampling, and sometimes without asymptotic arguments. Design-based inference usually assumes a known assignment process, which is often difficult to articulate and specify in observational studies. For instance, Borusyak and Hull (2023) specify counterfactual assignment of Chinese high-speed rail lines by permuting the ordering of line completion in time. Pollmann (2020) proposes to find counterfactual locations for a spatial treatment using convolutional neural networks. Our approach potentially provides an additional option in observational settings to elicit the assignment process by tethering it to a regression specification that one contemplates using. Since the implicit design is often unique, when a regression-using researcher is internally consistent, then they are implicitly specifying the design.

In practice, we could consider design-based inference taking the estimated implicit design $\hat{\pi}_i$ as the true design, ignoring the fact that $\hat{\pi}_i$ is estimated from data. We anticipate doing so is justifiable when the Π^* is a fixed Gram design. When it is not, we nevertheless suspect doing so to be justifiable in large samples when $\hat{\boldsymbol{\pi}}$ is consistent, but we leave formal justification to future work.

References

ABADIE, A. and IMBENS, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, **29** (1), 1–11. 2

ANDREWS, I., BARAHONA, N., GENTZKOW, M., RAMBACHAN, A. and SHAPIRO, J. M. (2023). *Structural Estimation Under Misspecification: Theory and Implications for Practice*. Tech. rep., National Bureau of Economic Research. 21

ANGRIST, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, **66** (2), 249–288. 3, 4, 6, 13, 22, 23, 31, 51

— (2001). Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *Journal of business & economic statistics*, **19** (1), 2–28. 5

— and PISCHKE, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press. 5

— and — (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, **24** (2), 3–30. 2

ARKHANGELSKY, D. and IMBENS, G. W. (2023). Fixed effects and the generalized mundlak estimator. *Review of Economic Studies*, p. rdad089. 3, 77

—, —, LEI, L. and LUO, X. (2021). Double-robust two-way-fixed-effects regression for panel data. *arXiv preprint arXiv:2107.13737*, **12**. 3

ARONOW, P. M. and SAMII, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, **60** (1), 250–267. 3

ATHEY, S. and IMBENS, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, **226** (1), 62–79. 3, 6, 16, 22, 24, 25, 51, 52, 57

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61** (4), 962–973. 2

BEN-MICHAEL, E., FELLER, A., HIRSHBERG, D. A. and ZUBIZARRETA, J. R. (2021). The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*. 2, 21

BLAKESLEE, D., FISHMAN, R. and SRINIVASAN, V. (2020). Way down in the hole: Adaptation to long-term water loss in rural india. *American Economic Review*, **110** (1), 200–224. 6, 7, 9, 21, 27, 31, 34, 35, 38

BLANDHOL, C., BONNEY, J., MOGSTAD, M. and TORGOVITSKY, A. (2025). *When is TSLS actually late?* Tech. rep., National Bureau of Economic Research. 3, 4, 5, 6, 11, 13, 16, 17, 20, 21, 23, 31, 51

BORUSYAK, K. and HULL, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica*, **91** (6), 2155–2185. 39

— and — (2024). *Negative weights are no concern in design-based specifications*. Tech. rep., National Bureau of Economic Research. 2, 3, 5, 13

—, — and JARAVEL, X. (2024a). Design-based identification with formula instruments: A review. *The Econometrics Journal*, p. utae003. 2

—, JARAVEL, X. and SPIESS, J. (2024b). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies*, p. rdae007. 24, 52

BUGNI, F. A., CANAY, I. A. and MCBRIDE, S. (2023). Decomposition and interpretation of treatment effects in settings with delayed outcomes. *arXiv preprint arXiv:2302.11505*. 3

CATTANEO, M. D., CRUMP, R. K., FARRELL, M. H. and FENG, Y. (2024). On binscatter. *American Economic Review*, **114** (5), 1488–1514. 33

CHATTOPADHYAY, A. and ZUBIZARRETA, J. R. (2023). On the implied weights of linear regression for causal inference. *Biometrika*, **110** (3), 615–629. 6, 7, 28

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*. 21

CHETVERIKOV, D., HAHN, J., LIAO, Z. and SHENG, S. (2023). Logit-based alternatives to two-stage least squares. *arXiv preprint arXiv:2312.10333*. 3

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96** (1), 187–199. 20, 34, 36

CURRIE, J., KLEVEN, H. and ZWIERS, E. (2020). Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 110, pp. 42–48. 2

DE CHAISEMARTIN, C. and D'HAULTFOEUILLE, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review*, **110** (9), 2964–2996. 24, 52

DUBE, O. and HARISH, S. (2020). Queens. *Journal of Political Economy*, **128** (7), 2579–2652. 21

DURRETT, R. (2019). *Probability: theory and examples*, vol. 49. Cambridge university press. 45, 66

FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, **40** (2), 180–193. 2, 5, 28

GOLDSMITH-PINKHAM, P. (2024). *Tracking the Credibility Revolution across Fields*. Tech. rep. 2

—, HULL, P. and KOLESÁR, M. (2024). Contamination bias in linear regressions. *American Economic Review*, **114** (12), 4015–4051. 3, 5, 6, 13, 22, 23, 51, 58, 59

GOODMAN-BACON, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of econometrics*, **225** (2), 254–277. 24, 52

HAILE, P. (2018). Theory and measurement: A perspective on empirical economics (lecture slides). 5

HAINMUELLER, J., MUMMOLO, J. and XU, Y. (2019). How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis*, **27** (2), 163–192. 27

HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35** (4), 1491–1523. 45, 67

HECKMAN, J. J. and URZUA, S. (2010). Comparing iv with structural models: What simple iv can and cannot identify. *Journal of Econometrics*, **156** (1), 27–37. 5

HWANG, S.-G. (2004). Cauchy's interlace theorem for eigenvalues of hermitian matrices. *The American mathematical monthly*, **111** (2), 157–159. 69

IMBENS, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic literature*, **48** (2), 399–423. 5

— (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, **50** (2), 373–419. 2, 5

— and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** (2), 467–475. 3, 5

— and RUBIN, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press. 21, 22, 34

— and WOOLDRIDGE, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, **47** (1), 5–86. 6, 22, 23, 51

KLINE, P. (2011). Oaxaca-blinder as a reweighting estimator. *American Economic Review*, **101** (3), 532–537. 3, 5, 6, 22, 24, 28, 51

KOLESÁR, M. (2024). *Econ 539B Lecture Notes: Recent Progress on Regression*. Tech. rep. 2

KOLESÁR, M. and PLAGBORG-MØLLER, M. (2024). Dynamic causal effects in a nonlinear world: the good, the bad, and the ugly. *arXiv preprint arXiv:2411.10415*. 3

KOOPMANS, T. C. and REIERSOL, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics*, **21** (2), 165–181. 10, 15

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, **113** (521), 390–400. 2

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. 3, 5, 6, 16, 22, 23, 28, 51

MANSKI, C. and TAMER, E. (2019). The et interview: Professor charles manski. *Econometric Theory*, **35** (2), 233–294. 5

MANSKI, C. F. (2022). Identification and statistical decision theory. *Econometric Theory*, pp. 1–17. 15

MIRATRIX, L. W., SEKHON, J. S. and YU, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **75** (2), 369–396. 6, 23, 51

MOGSTAD, M. and TORGOVITSKY, A. (2024). *Instrumental Variables with Unobserved Heterogeneity in Treatment Effects*. Tech. rep., National Bureau of Economic Research. 3, 5, 19, 21

NEGI, A. and WOOLDRIDGE, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, **40** (5), 504–534. 28

NEYMAN, J. (1923/1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472. 16

POIRIER, A. and SŁOCZYŃSKI, T. (2024). Quantifying the internal validity of weighted estimands. *arXiv preprint arXiv:2404.14603*. 20

POLLMANN, M. (2020). Causal inference for spatial treatments. *arXiv preprint arXiv:2011.00373*. 39

RAMBACHAN, A. and ROTH, J. (2020). Design-based uncertainty for quasi-experiments. *arXiv preprint arXiv:2008.00602*. 16, 45, 67

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. 24, 52

SŁOCZYŃSKI, T. (2020). When should we (not) interpret linear iv estimands as late? *arXiv preprint arXiv:2011.06695*. 3, 16

SŁOCZYŃSKI, T. (2022). Interpreting ols estimands when treatment effects are heterogeneous: smaller groups get larger weights. *Review of Economics and Statistics*, **104** (3), 501–509. 3

WAGER, S. (2024). Causal inference: A statistical learning approach. 6, 21

YITZHAKI, S. (1996). On using linear regressions in welfare economics. *Journal of Business & Economic Statistics*, **14** (4), 478–486. 5

ZADROZNY, B. and ELKAN, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, vol. 1, pp. 609–616. 22

ZHAO, A. and DING, P. (2022). Regression-based causal inference with factorial experiments: estimands, model specifications and design-based properties. *Biometrika*, **109** (3), 799–815. 16

—, — and LI, F. (2025). Interacted two-stage least squares with treatment effect heterogeneity. *arXiv preprint arXiv:2502.00251*. 3, 6, 16, 22, 26

ZUBIZARRETA, J. R., PAREDES, R. D. and ROSENBAUM, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. 2

Appendix A. Estimating potential weights

This section ties up the loose ends in the sampling uncertainty for G_n. Naturally, we estimate potential weights and implicit designs by plugging in the sample Gram matrix $\hat{G}_n$. When we fail to find valid estimated implicit designs, we may wonder whether that is due to sampling noise in $\hat{G}_n$ rather than the fault of the regression specification. When we do find a valid estimated implicit design, we may also ask whether $\hat{\boldsymbol{\pi}}$ is close to $\boldsymbol{\pi}$. To answer these questions, we provide a set of simple estimation and inference guarantees.[37]

In terms of inference, conceptually, for $\hat{\mathcal{G}}_n$ a $(1-\alpha)$-confidence set for G_n, we may form corresponding confidence sets $\hat{\mathcal{R}}_i$ for $\boldsymbol{\rho}_i$ and $\hat{\mathcal{P}}$ for $\boldsymbol{\pi}^*$ by projection:

$$\hat{\mathcal{R}}_i = \left\{ \Lambda G_n^{-1} \mathbf{z}(\mathbf{x}_i, \mathbf{w})' : G_n \in \hat{\mathcal{G}}_n \right\}$$

$$\hat{\mathcal{P}} = \Bigg\{ \boldsymbol{\pi} : G \equiv \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) \mathbf{z}(\mathbf{x}_i, \mathbf{w}) \mathbf{z}(\mathbf{x}_i, \mathbf{w})' \in \hat{\mathcal{G}}_n \qquad \text{(Gram-consistent)}$$

$$\forall i,\ \pi_i \in \Delta(\mathcal{W}) \qquad \text{(Proper)}$$

$$\forall i,\ \sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) \Lambda G^{-1} \mathbf{z}(\mathbf{x}_i, \mathbf{w})' = 0 \Bigg\}. \qquad \text{(Level irrelevance)}$$

It is easy to see that, when $\boldsymbol{\pi}^*$ satisfies (1.2), $\mathrm{P}[\boldsymbol{\pi}^* \in \hat{\mathcal{P}}] = \mathrm{P}[G_n(\boldsymbol{\pi}^*) \in \hat{\mathcal{G}}_n] \geq 1 - \alpha$. $\hat{\mathcal{G}}_n$ is readily available from, e.g., a simultaneous confidence set for all entries of G_n. These confidence sets are usually available upon verification of high-probability bounds for $\|\hat{G}_n - G_n\|$ or central limit theorems for $\sqrt{n}(\hat{G}_{n,jk} - G_{n,jk})$. Unfortunately, fully computing $\hat{\mathcal{P}}$ is a difficult nonlinear optimization problem. Nonetheless, checking whether a candidate $\boldsymbol{\pi} \in \hat{\mathcal{P}}$ is straightforward. Thus, if the researcher has a particular design in mind—perhaps in response to the red flags raised by estimated implicit designs—they can check whether such a design is a plausible justification for the specification in question.

The rest of this subsection states consistency guarantees for the estimated potential weights and implicit designs. Under mild assumptions, it is often possible to show that $\hat{G}_n - G_n \xrightarrow{p} 0$. This, coupled with an assumption that G_n is suitably invertible, allows us to conclude that:

(1) The estimated potential weights $\hat{\boldsymbol{\rho}}_i(\mathbf{w})$ are consistent for $\boldsymbol{\rho}_i(\mathbf{w})$.

(2) In many cases, the consistency of potential weights implies consistency of estimated implied designs and estimands, though the precise statement depends on further regularity conditions. A key obstruction to a general result is that small perturbations to coefficients in linear equations may result in large perturbations to their solutions.

[37] Since we view the estimated implicit designs and potential weights as diagnostics and sanity checks, we do not provide particularly strong or optimal guarantees and leave those to future work. In our view, if a practitioner fails to find valid estimated implicit designs, even if that is due to noise, it should nevertheless raise alarms for the interpretation of the regression estimator.

To discuss asymptotics, we embed our finite population in a sequence of populations. Formally, let $\mathbf{y}_1(\cdot), \mathbf{x}_1, \mathbf{y}_2(\cdot), \mathbf{x}_2, \ldots$ denote a sequence of potential outcomes and covariates. Let Π_n^* denote the joint distribution of treatment assignments for the first n units. We shall consider the behavior as $n \to \infty$, relative to these sequences. We impose a high-level condition on the sequence of populations such that the estimated Gram matrix is consistent for the population Gram matrix.[38]

Assumption A.1. $\hat{G}_n - G_n \xrightarrow{p} 0$.

Next, we impose the following assumption on the population Gram matrices to ensure that $\hat{G}_n^{-1} - G_n^{-1} \xrightarrow{p} 0$ as well.[39]

Assumption A.2. *The sequence of population Gram matrices G_n is such that their minimum eigenvalues are bounded below: For some $\epsilon > 0$, for all n, $\lambda_{\min}(G_n) \geq \epsilon > 0$.*

Assumption A.2 makes sure that the population regression specification is strongly identified, in the sense that the Gram matrix is bounded away from singularity. Under Assumption A.1 and Assumption A.2, we can show that the estimated potential weights are consistent for their population counterparts.

Proposition A.1. *Under Assumptions A.1 and A.2, the estimated potential weights are consistent: For every i, as $i \leq n \to \infty$, $\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w}) \xrightarrow{p} 0$. If $\|z_t(\mathbf{x}_i, \mathbf{w})\|_\infty$ is bounded uniformly in $i \in [n]$ and $\mathbf{w} \in \mathcal{W}$, then the consistency is also uniform: As $n \to \infty$*

$$\max_{i \in [n], \mathbf{w} \in \mathcal{W}} |\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty \xrightarrow{p} 0. \tag{A.1}$$

where $|\cdot|_\infty$ takes the entrywise maximum absolute value.

Naturally, we define the set of estimated implicit designs as the set of solutions to the sample counterpart of (2.8): $\hat{\boldsymbol{\pi}}$ is an estimated implicit design if for every i,

$$\sum_{\mathbf{w} \in \mathcal{W}} \hat{\pi}_i(\mathbf{w}) \hat{\boldsymbol{\rho}}_i(\mathbf{w}) = 0, \quad \sum_{\mathbf{w} \in \mathcal{W}} \hat{\pi}_i(\mathbf{w}) = 1. \tag{A.2}$$

Unfortunately, in general, the estimated implicit designs are not always well-behaved. This is because they are the solutions to a system of overdetermined linear equations with noisy coefficients, and small perturbations to the coefficients can result in large changes in the solutions—or even to their existence in the first place.[40] Thus, it is possible that a population implicit design exists—meaning that (2.8) is under or exactly determined—but no estimated implicit designs do. Failing

[38] Assumption A.1 holds whenever the underlying populations are such that a law of large numbers hold. If the treatments are independently assigned, for instance, then Assumption A.1 is true under standard laws of large number (e.g., Theorem 2.2.6 in Durrett (2019)), which we state in Lemma C.4. Appendix C.2.1 verifies Assumption A.1 for designs that involve sampling without replacement (Hájek, 1964; Rambachan and Roth, 2020).

[39] Under Assumption A.1 and Assumption A.2, one can also show that $\hat{G}_n$ is invertible with probability tending to one, and thus we may write $\hat{G}_n^{-1}$ without essential loss of generality.

[40] For instance, it is possible for noise in the coefficients of a underdetermined system to turn the system into an overdetermined one.

a general result, we present weaker consistency results that apply in at least in the cross-section ($T = 1$).[41]

In particular, if the population implicit design is unique and equal to $\boldsymbol{\pi}^*$, the estimated implicit designs are consistent for $\boldsymbol{\pi}^*$. We illustrate with the following result in the cross-sectional case ($T = 1$) where the system of equations (2.8) is exactly determined ($k = J$). To state this result, let $R_i = [\boldsymbol{\rho}_i(0), \ldots, \boldsymbol{\rho}_i(J)] \in \mathbb{R}^{k\times(J+1)}$ be the matrix whose columns are potential weights at each treatment level. Let $R_i = \sum_{r=1}^{J} \sigma_{ir} u_{ir} v_{ir}'$ be its singular value decomposition, for $\sigma_{1i} \geq \ldots \geq \sigma_{iJ} \geq 0$, $u_{i1}, \ldots, u_{iJ}$ orthonormal vectors in $\mathbb{R}^J$ and $v_{i1}, \ldots, v_{iJ}$ orthonormal vectors in $\mathbb{R}^{J+1}$.

Proposition A.2. *Assume that $k = J$. Suppose that:*

(i) The estimand $\boldsymbol{\tau}$ is a vector of causal contrasts in the sense of Definition 2.4 under $\boldsymbol{\pi}^*$.

(ii) Given $\rho_i(\cdot)$, the smallest singular value of R_i is uniformly bounded away from zero: For some $\epsilon > 0$, $\liminf_{n\to\infty} \min_{i\in[n]} \sigma_{iJ} > \epsilon > 0$.

(iii) The estimated potential weights are consistent in the sense of (A.1).

Then, for some $C > 0$, a unique estimated implicit design $\hat{\boldsymbol{\pi}}$ that is bounded by C (that is, $\max_i \max_{\mathbf{w}\in\mathcal{W}} |\hat{\pi}_i(\mathbf{w})| \leq C$) exists with probability tending to one. Let $\hat{\boldsymbol{\pi}}$ be an estimated implicit design if it exists, and otherwise let $\hat{\boldsymbol{\pi}}$ be an arbitrary probability vector. Then $\hat{\boldsymbol{\pi}}$ is consistent for $\boldsymbol{\pi}^*$:

$$\max_{i\in[n],\mathbf{w}\in\mathcal{W}} |\hat{\pi}_i(\mathbf{w}) - \pi_i^*(\mathbf{w})| \xrightarrow{p} 0.$$

The key condition in Proposition A.2 is (ii), which ensures that the solution to the population level-irrelevance restrictions (2.8) is unique and robust to small perturbations of the potential weights. In the binary treatment, cross-sectional setting, (ii) is equivalent to that the potential weights are bounded away from zero: $\rho_i(1)^2 + \rho_i(0)^2 > \epsilon^2$. Lemma C.8 further provides sufficient conditions for (ii) to hold in the $T = 1$ case.

Appendix B. Proofs

Theorem 2.1. *$\boldsymbol{\tau}$ is a vector of causal contrasts under $\boldsymbol{\pi}^*$ in the sense of Definition 2.4 if and only if $\boldsymbol{\pi}^*$ is equal to some implicit design. When this happens, if, furthermore, there is a unique implicit design $\boldsymbol{\pi}$, then $\boldsymbol{\pi}$ must be proper and Gram-consistent.*

Proof. We first show that Definition 2.4 is equivalent to the design satisfying (2.8) for a population regression estimand. By (2.7), $\boldsymbol{\tau}$ satisfies Definition 2.4 if and only if for all i, t

$$\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\boldsymbol{\rho}_{it}(\mathbf{w})c_{it} = 0 \iff \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\boldsymbol{\rho}_{it}(\mathbf{w}) = 0.$$

The sum-to-one condition is automatically satisfied by π_i^*. This is then equivalent to (2.8).

[41] Proposition C.5 shows that if an estimated implicit design $\hat{\boldsymbol{\pi}}$ exists with bounded entries, then it is consistent in the sense that it approximately solves the population level irrelevance equations: $\sum_{\mathbf{w}} \hat{\pi}_i(\mathbf{w})\boldsymbol{\rho}_i(\mathbf{w})$ tends to zero.

For the first statement and the $\implies$ direction, if $\boldsymbol{\tau}$ is a vector of causal contrasts for the regression specification, then $\pi_i^*(\cdot)$ solves (2.8), and thus $\boldsymbol{\pi}^*$ is an implicit design. For the $\impliedby$ direction, if $\boldsymbol{\pi}^*$ is an implicit design, then $\boldsymbol{\pi}^*$ solves (2.8), which means $\boldsymbol{\tau}$ satisfies level irrelevance.

When $\boldsymbol{\pi}^*$ is an implicit design, if there is only one implicit design, then $\boldsymbol{\pi}^*$ must be that design. As a result, that design is proper and Gram-consistent. □

Corollary 2.2 ($k = T = J = 1$)**.** *When $k = T = J = 1$, τ is a causal contrast under the true design $\boldsymbol{\pi}^*$ if and only if $\rho_i(1)\rho_i(0) \leq 0$ for all i and $\pi_i^* = \frac{-\rho_i(0)}{\rho_i(1)-\rho_i(0)}$ for all units i with one of $\rho_i(1)$ and $\rho_i(0)$ nonzero. When this happens, the implicit estimand is uniquely defined by $\omega_i(\boldsymbol{\pi}^*, 1) = \pi_i^*\rho_i(1) = -\omega_i(\boldsymbol{\pi}^*, 0)$, and thus*

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \omega_i^*(y_i(1) - y_i(0)) \text{ for } \omega_i^* \equiv \omega_i(\boldsymbol{\pi}^*, 1).$$

The unit-level weights ω_i^* are negative if and only if $\rho_i(1) < 0 < \rho_i(0)$.

Proof. For the $\implies$ direction, if τ is a causal contrast, then $\boldsymbol{\pi}^*$ is an implicit design. Hence

$$\pi_i^*\rho_i(1) + (1 - \pi_i^*)\rho_i(0) = 0.$$

Since $\pi_i^* \in [0, 1]$, this implies that $\rho_i(1), \rho_i(0)$ must weakly be on opposite sides of zero. If one of them is nonzero, then $\pi_i^* = \frac{-\rho_i(0)}{\rho_i(1)-\rho_i(0)}$. On the other hand, for the $\impliedby$ direction, suppose $\rho_i(1)\rho_i(0) \leq 0$ and $\pi_i^* = \frac{-\rho_i(0)}{\rho_i(1)-\rho_i(0)} \in [0, 1]$ for all i where one of $\rho_i(1), \rho_i(0)$ is nonzero. Then

$$\pi_i^*\rho_i(1) + (1 - \pi_i^*)\rho_i(0) = 0$$

for all i where one of $\rho_i(1), \rho_i(0)$ is nonzero by construction. For the other units (i.e., those with $\rho_i(1) = \rho_i(0) = 0$), the condition holds trivially, irrespective of π_i^*, since both potential weights are zero. Thus τ satisfies level irrelevance under $\boldsymbol{\pi}^*$.

Note that, when this happens, $\omega_i(\boldsymbol{\pi}^*, 0) = (1 - \pi_i^*)\rho_i(0) = -\pi_i^*\rho_i(1)$. This proves the representation of τ by (2.7). Moreover,

$$\pi_i^*\rho_i(1) < 0 \iff \rho_i(1) < 0 \text{ and } \rho_i(0) \neq 0 \iff \rho_i(1) < 0 < \rho_i(0)$$

since $\rho_i(1)\rho_i(0) \leq 0$. □

Proposition A.1. *Under Assumptions A.1 and A.2, the estimated potential weights are consistent: For every i, as $i \leq n \to \infty$, $\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w}) \xrightarrow{p} 0$. If $\|z_t(\mathbf{x}_i, \mathbf{w})\|_\infty$ is bounded uniformly in $i \in [n]$ and $\mathbf{w} \in \mathcal{W}$, then the consistency is also uniform: As $n \to \infty$*

$$\max_{i\in[n],\mathbf{w}\in\mathcal{W}} |\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty \xrightarrow{p} 0. \tag{A.1}$$

where $|\cdot|_\infty$ takes the entrywise maximum absolute value.

Proof. Take $\eta > 0$. For the result with fixed i, it suffices to show that $\mathrm{P}[\|\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})\|_F > \eta] \to 0$. Note that

$$\begin{aligned}\mathrm{P}\left(\|\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})\|_F > \eta\right) \leq{}& \mathrm{P}[\lambda_{\min}(\hat{G}_n) \leq \epsilon/2] \\ &+ \mathrm{P}\left[\|\Lambda(\hat{G}_n^{-1} - G_n^{-1})\mathbf{z}(\mathbf{x}_i, \mathbf{w})'\|_F > \eta, \lambda_{\min}(\hat{G}_n) > \epsilon/2\right].\end{aligned}$$

The first term converges to zero by Assumptions A.1 and A.2 and the Hoffman–Wielandt inequality. Note that, by the submultiplicativity of the Frobenius norm, when $\lambda_{\min}(\hat{G}_n) > \epsilon/2$,

$$\|\hat{G}_n^{-1} - G_n^{-1}\|_F \leq \|G_n^{-1}\|_F \|G_n - \hat{G}_n\|_F \|\hat{G}_n^{-1}\|_F \leq \frac{2}{\epsilon^2}\|G_n - \hat{G}_n\|_F \xrightarrow{p} 0$$

Thus, for some C dependent on ϵ, Λ, and $\|z_t(\mathbf{x}_i, \mathbf{w})\|_\infty$,

$$\mathrm{P}\left[\|\Lambda(\hat{G}_n^{-1} - G_n^{-1})\mathbf{z}(\mathbf{x}_i, \mathbf{w})'\|_F > \eta, \lambda_{\min}(\hat{G}_n) > \epsilon/2\right] \leq \mathrm{P}\left[C\|G_n - \hat{G}_n\|_F > \eta\right] \to 0.$$

Therefore $\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w}) \xrightarrow{p} 0$ for every fixed i.

To prove convergence that is uniform in i, note that, if $\lambda_{\min}(\hat{G}_n) > \epsilon/2$, up to constants,

$$\begin{aligned}\max_i |\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty &\lesssim \max_i \|\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})\|_F \\ &\lesssim \max_i \|\mathbf{z}(\mathbf{x}_i, \mathbf{w})\|_F \cdot \|\hat{G}_n^{-1} - G_n^{-1}\|_F \\ &\lesssim \|\hat{G}_n^{-1} - G_n^{-1}\|_F \cdot \max_i \max_t \|z_t(\mathbf{x}_i, \mathbf{w})\|_\infty \\ &\lesssim \|\hat{G}_n^{-1} - G_n^{-1}\|_F \\ &\lesssim \|G_n - \hat{G}_n\|_F.\end{aligned}$$

Therefore,

$$\mathrm{P}\left(\max_i |\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty > \eta\right) \leq \mathrm{P}[\lambda_{\min}(\hat{G}_n) \leq \epsilon/2] + \mathrm{P}\left[\|G_n - \hat{G}_n\|_F \gtrsim \eta\right] \to 0.$$

□

Proposition A.2. *Assume that $k = J$. Suppose that:*

(i) The estimand $\boldsymbol{\tau}$ is a vector of causal contrasts in the sense of Definition 2.4 under $\boldsymbol{\pi}^*$.

(ii) Given $\rho_i(\cdot)$, the smallest singular value of R_i is uniformly bounded away from zero: For some $\epsilon > 0$, $\liminf_{n\to\infty} \min_{i\in[n]} \sigma_{iJ} > \epsilon > 0$.

(iii) The estimated potential weights are consistent in the sense of (A.1).

Then, for some $C > 0$, a unique estimated implicit design $\hat{\boldsymbol{\pi}}$ that is bounded by C (that is, $\max_i \max_{\mathbf{w}\in\mathcal{W}} |\hat{\pi}_i(\mathbf{w})| \leq C$) exists with probability tending to one. Let $\hat{\boldsymbol{\pi}}$ be an estimated implicit design if it exists, and otherwise let $\hat{\boldsymbol{\pi}}$ be an arbitrary probability vector. Then $\hat{\boldsymbol{\pi}}$ is consistent for $\boldsymbol{\pi}^*$:

$$\max_{i\in[n],\mathbf{w}\in\mathcal{W}} |\hat{\pi}_i(\mathbf{w}) - \pi_i^*(\mathbf{w})| \xrightarrow{p} 0.$$

Proof. Let $\hat{R}_i = [\hat{\boldsymbol{\rho}}_i(0), \ldots, \boldsymbol{\rho}_i(J)]$ be the sample analogue of R_i. Both $\hat{R}_i$ and R_i may depend on n. Assume n is sufficiently large such that $\min_{i\in[n]} \sigma_{iJ} > \epsilon/2 > 0$. Thus, each R_i is a $J \times (J+1)$ matrix with rank J such that $R_i\pi_i^* = 0$. As a minor abuse of notation, let π_i, π_i^* be represented as vectors in $\mathbb{R}^{J+1}$. Let $u_i = \pi_i^*/\|\pi_i^*\|$ be the unit vector in the (one-dimensional) null space of R_i. Note that $1/(J+1) \le \|\pi_i^*\|_\infty \le \|\pi_i^*\| \le \|\pi_i^*\|_1 = 1$, and thus $1'u_i = 1/\|\pi_i^*\| \in [1, J+1]$ is bounded above and below.

Fix some C to be chosen. Let E be complement to the event that there exists a unique bounded estimated implicit design. Note that $E = \bigcup_{i\in[n]} E_i$, where each E_i is the complement to the event that $\hat{R}_i\hat{\pi}_i = 0$ has a unique solution $\hat{\pi}_i$ with $1'\hat{\pi}_i = 1$ where $\|\hat{\pi}_i\|_\infty \le C$. Let $\hat{\sigma}_{i1} \ge \cdots \ge \hat{\sigma}_{iJ} \ge 0$ be the singular values of $\hat{R}_i$. Suppose $\hat{\sigma}_{iJ} > 0$, then $\hat{R}_i$ has full rank and the equation $\hat{R}_i\hat{\pi}_i = 0$ has a unique one-dimensional space of solutions. If that space exists, let $\hat{u}_i$ be a unit vector in that space, unique up to sign. If $1'\hat{u}_i \neq 0$, then we choose the sign of $\hat{u}_i$ so that $1'\hat{u}_i > 0$. Note that if $|\hat{u}_i'1| > 1/C$, then $\hat{\pi}_i = \hat{u}_i/\hat{u}_i'1$ is a bounded estimated implicit design with $\|\hat{\pi}_i\|_\infty \le C$.

Note that

$$E_i \subset \{\hat{\sigma}_{iJ} \le \epsilon/4\} \cup \left\{\hat{\sigma}_{iJ} > \epsilon/4, |\hat{u}_i'1| < 1/C\right\}.$$

We will show that $E_i \subset \left\{\|\hat{R}_i - R_i\|_F > \kappa\right\}$ for some κ. For the event $\{\hat{\sigma}_{iJ} \le \epsilon/4\}$, the Hoffman–Wielandt inequality implies that

$$|\hat{\sigma}_{iJ} - \sigma_{iJ}|^2 \le \sum_{k=1}^{J} |\hat{\sigma}_{ik} - \sigma_{ik}|^2 \le \|R_i - \hat{R}_i\|_F^2,$$

and thus $\hat{\sigma}_{iJ} \ge \epsilon/2 - \|R_i - \hat{R}_i\|_F$, which further implies that $\{\hat{\sigma}_{iJ} < \epsilon/4\} \subset \left\{\|R_i - \hat{R}_i\|_F > \epsilon/4\right\}$.

Let $\hat{u}_i = \hat{c}_i u_i + \sqrt{1-\hat{c}_i^2}\hat{u}_{\perp,i}$ where $\hat{c}_i, \hat{u}_{\perp,i}$ are uniquely chosen so that $\hat{u}_{\perp,i}$ is a unit vector orthogonal to u_i. Then

$$|\hat{u}_i'1| \ge |\hat{c}_i 1'u_i| - \sqrt{1-\hat{c}_i^2}|1'\hat{u}_{\perp,i}| \ge |\hat{c}_i| - \sqrt{1-\hat{c}_i^2}\sqrt{J+1}.$$
$$(1'u_i \ge 1, |1'\hat{u}_i| \le \|1\|\|\hat{u}_i\| = \sqrt{J+1})$$

Now, note that

$$\|R_i\hat{u}_i\| = \|(R_i - \hat{R}_i)\hat{u}_i\| \le \|R_i - \hat{R}_i\|_F.$$

On the other hand,

$$\|R_i\hat{u}_i\| = \sqrt{1-\hat{c}_i^2}\|R_i\hat{u}_{\perp,i}\| \ge \sqrt{1-\hat{c}_i^2}\sigma_{iJ} \ge \sqrt{1-\hat{c}_i^2}\frac{\epsilon}{2},$$

and thus

$$1 - \hat{c}_i^2 \le \frac{4}{\epsilon^2}\|\hat{R}_i - R_i\|_F^2 \iff |\hat{c}_i| \ge \sqrt{1 - \frac{4}{\epsilon^2}\|\hat{R}_i - R_i\|_F^2}. \tag{B.1}$$

Hence,

$$|\hat{u}_i'1| \ge \sqrt{1 - \frac{4}{\epsilon^2}\|\hat{R}_i - R_i\|_F^2} - \frac{2\sqrt{J+1}}{\epsilon}\|\hat{R}_i - R_i\|_F.$$

Suppose $\|R_i - \hat{R}_i\|_F < \epsilon/(4\sqrt{J+1})$, then

$$|\hat{u}_i' 1| \geq \sqrt{1 - \frac{1}{J+1}} - \frac{1}{2} \geq \sqrt{1/2} - 1/2 \geq 0.207.$$

Thus, if we pick $C > 1/0.207$, then

$$E_i \subset \{\hat{\sigma}_{iJ} < \epsilon/4\} \cup \left\{\hat{\sigma}_{iJ} > \epsilon/4, |\hat{u}_i' 1| < 1/C\right\} \subset \left\{\|R_i - \hat{R}_i\|_F > \frac{\epsilon}{4\sqrt{J+1}}\right\}.$$

As a result,

$$\mathrm{P}\left[\bigcup_i E_i\right] \leq \mathrm{P}\left[\max_{i\in[n]} \|R_i - \hat{R}_i\|_F > \frac{\epsilon}{4\sqrt{J+1}}\right] \to 0$$

by assumption. This proves the first part regarding the existence of a unique bounded implicit design.

For the second part, note that when $\hat{\sigma}_{iJ} > 0$,

$$\hat{\pi}_i - \pi_i = \frac{\hat{u}_i}{1'\hat{u}_i} - \frac{u_i}{1'u_i} = \frac{1'}{1'\hat{u}_i}(u_i - \hat{u}_i)\frac{\hat{u}_i}{1'u_i} + \frac{1}{1'u_i}(\hat{u}_i - u_i).$$

Thus

$$\|\hat{\pi}_i - \pi_i\|_2 \leq \frac{1}{0.207}\sqrt{J+1}\|u_i - \hat{u}_i\|_2 + \|\hat{u}_i - u_i\|_2$$

if $\|R_i - \hat{R}_i\|_F \leq \epsilon/(4\sqrt{J+1})$.

To bound $\|\hat{u}_i - u_i\|$, we first show that $\hat{c}_i > 0$ when R_i is close to $\hat{R}_i$. When $\|R_i - \hat{R}_i\|_F \leq \frac{\epsilon}{4\sqrt{J+1}}$, we note that by (B.1),

$$|\hat{c}_i| \geq \sqrt{1 - \frac{1}{4(J+1)}} \geq \sqrt{7/8}.$$

When $\|R_i - \hat{R}_i\|_F \leq \frac{\epsilon}{4\sqrt{J+1}}$, we also have that, because we pick $\hat{u}_i$ to have $1'\hat{u}_i \geq 0$,

$$0.207 \leq 1'\hat{u}_i = \hat{c}_i + \sqrt{1-\hat{c}_i^2}\, 1'\hat{u}_{\perp,i} \leq \hat{c}_i + \sqrt{1-\hat{c}_i^2}\sqrt{J+1} \leq \hat{c}_i + \frac{2}{\epsilon}\frac{\epsilon}{4\sqrt{J+1}}\sqrt{J+1} = \hat{c}_i + \frac{1}{2}.$$

Thus $\hat{c}_i > 0.207 - 0.5$. Since $|\hat{c}_i| > \sqrt{7/8}$, we conclude that $\hat{c}_i > \sqrt{7/8} > 0$. Thus, when $\|R_i - \hat{R}_i\|_F \leq \frac{\epsilon}{4\sqrt{J+1}}$, (B.1) implies

$$c_i \geq \sqrt{1 - \frac{4}{\epsilon^2}\|R_i - \hat{R}_i\|_F^2} \geq 1 - \frac{4}{\epsilon^2}\|R_i - \hat{R}_i\|_F^2.$$

Now, when $\|R_i - \hat{R}_i\|_F \leq \frac{\epsilon}{4\sqrt{J+1}}$,

$$\begin{aligned}\|\hat{u}_i - u_i\|_2^2 &\leq (1-\hat{c}_i)^2 + (1-\hat{c}_i^2) \\ &= 2 - 2\hat{c}_i \leq \frac{8}{\epsilon^2}\|R_i - \hat{R}_i\|_F^2\end{aligned}$$

Therefore, for some constant $C(J,\epsilon)$,

$$\|\hat{\pi}_i - \pi_i\|_2 \leq C(J,\epsilon)\|R_i - \hat{R}_i\|_F.$$

Hence, for $\eta > 0$,

$$\left\{ \max_{i\in[n]} \|\hat{\pi}_i - \pi_i\|_2 > \eta \right\} \subset \left\{ \max_{i\in[n]} \|R_i - \hat{R}_i\|_F > \max\left(\frac{\epsilon}{4\sqrt{J+1}}, \frac{1}{C(J,\epsilon)} \right) \right\}$$

The probability on the right-hand side converges to zero by assumption. This proves the convergence of $\hat{\pi}_i$ to π_i uniformly in i in $\|\cdot\|_2$. Since $\|\cdot\|_\infty \leq \|\cdot\|_2$ in $\mathbb{R}^{J+1}$, this concludes the proof. □

Theorem 3.1. *We compute the implicit designs and estimands of the following regression specifications. In all of the following scenarios, the implicit design exists and is unique. The implicit design is Gram-consistent regardless of whether $\boldsymbol{\pi} = \boldsymbol{\pi}^*$ for all scenarios except for (4).*

(1) (Angrist, 1998; Blandhol et al., 2025, Section 2.1, $k = T = J = 1$) Suppose x_i includes a constant. For $Y_i = \tau W_i + x_i'\gamma + \epsilon_i$ and τ the coefficient of interest:

- *(a) $\pi_i = x_i'\delta$ for $\delta = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n \pi_i^* x_i$*
- *(b) $\pi_i^* = \pi_i$ if and only if $\pi_i^* = x_i'\delta$*
- *(c) $\omega_i \equiv \omega_i(\boldsymbol{\pi}, 1) = -\omega_i(\boldsymbol{\pi}, 0) = \frac{\pi_i(1-\pi_i)}{\frac{1}{n}\sum_{i=1}^n \pi_i(1-\pi_i)}$. When $\pi_i^* = \pi_i$, $\omega_i \geq 0$.*

(2) (Goldsmith-Pinkham et al., 2024, $k = J > 1, T = 1$) Suppose $\mathcal{W} = \{0, 1, \ldots, J\}$ for $J > 1$ and x_i includes a constant. Let $W_{ij} = \mathbb{1}(W_i = j)$. For $Y_i = \sum_{j=1}^J \tau_j W_{ij} + x_i'\gamma + \epsilon_i$ and $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_J]'$ the coefficient of interest:

- *(a) $\pi_i(j) = x_i'\delta_j$ for $\delta_j = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n \pi_i^*(j) x_i$*
- *(b) $\pi_i^* = \pi_i$ if and only if $\pi_i^*(j) = x_i'\delta_j$ for all $j \in [J]$*
- *(c) The implicit estimand is shown in (B.2). This estimand is generally contaminated (that is, $\omega_{ij}(\boldsymbol{\pi}, \ell) \neq 0$ for some $j \in [J]$ and $\ell \notin \{0, j\}$).*

(3) (Miratrix et al., 2013; Imbens and Wooldridge, 2009; Lin, 2013, $k = T = J = 1$) Suppose x_i represents saturated discrete covariates taking values in $\{0, \ldots, L\}$, in the sense that $x_i = [x_{i1}, \ldots, x_{iL}]'$ for $x_{i\ell} = \mathbb{1}(x_i = \ell)$. Let $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. For $Y_i = \alpha_0 + \gamma_1' x_i + \tau W_i + W_i(x_i - \overline{x})'\gamma_2 + \epsilon_i$ and τ the coefficient of interest:

- *(a) For all units i whose covariate value is ℓ (i.e., $x_{i\ell} = 1$), the implicit design $\pi_i = \pi_\ell$, where $\pi_\ell = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(x_i = \ell)$*
- *(b) $\pi_i^* = \pi_i$ if and only if π_i^* is the same for all units with the same x_i-value*
- *(c) The implicit estimand is the ATE. That is, $\omega_i = \omega_i(\boldsymbol{\pi}, 1) = -\omega_i(\boldsymbol{\pi}, 0) = 1$.*

(4) (Kline, 2011, $k = T = J = 1$) Let $\overline{x}_1 = \frac{\sum_{i=1}^n \pi_i^* x_i}{\sum_{i=1}^n \pi_i^*}$. For $Y_i = \alpha_0 + \gamma_1' x_i + \tau W_i + W_i(x_i - \overline{x}_1)'\gamma_2 + \epsilon_i$ and τ the coefficient of interest:

- *(a) $\pi_i = \frac{\delta_0 + (x_i - \overline{x})'\delta_1}{1 + \delta_0 + (x_i - \overline{x})'\delta_1}$, where δ_0, δ_1 are equal to the population projection coefficients of $\pi_i^*/(1-\pi_i^*)$ on $x_i - \overline{x}$ and a constant,* weighted by $1 - \pi_i^*$
- *(b) $\pi_i^* = \pi_i$ if and only if $\pi_i^*/(1-\pi_i^*) = \delta_0 + \delta_1'(x_i - \overline{x})$*
- *(c) When $\boldsymbol{\pi}^* = \boldsymbol{\pi}$, the implicit estimand is the ATT. That is, $\omega_i = \omega_i(\boldsymbol{\pi}, 1) = -\omega_i(\boldsymbol{\pi}, 0) = \frac{\pi_i}{\frac{1}{n}\sum_{i=1}^n \pi_i}$.*

(5) (Athey and Imbens, 2022, $T > 1$) Let $\mathcal{W} \subset \{0,1\}^T$ be the set of treatment paths such that its nonzero elements are linearly independent vectors whose span excludes the constant vector

$1_T = [1, \ldots, 1]'$.[42] *For the two-way fixed effects (TWFE) specificaiton* $\mathbf{Y}_{it} = \alpha_i + \mu_t + \tau \mathbf{W}_{it} + \epsilon_{it}$ *and the coefficient of interest* τ*:*

(a) The implicit design is constant in i *and is unique,* $\pi_i(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \pi_i^*(\mathbf{w})$

(b) $\pi_i^* = \pi_i$ *if and only if* π_i^* *is the same for all* i

(c) The implicit estimand is shown in (B.4)*, which matches Theorem 1(ii) in Athey and Imbens (2022) under staggered adoption.*[43]

Proof. We prove the following:

(1) By Theorem C.1, the potential weights of this regression are the same as those for the Frisch–Waugh–Lovell-transformed regression of Y_i on $W - \delta' x_i$. For the FWL-transformed regression, its Gram matrix is a positive scalar

$$\tilde{G}_n = \frac{1}{n}\sum_{i=1}^n \pi_i^*(1 - \delta' x_i)^2 + (1 - \pi_i^*)(\delta' x_i)^2 > 0$$

Thus the potential weights are

$$\rho_i(w) = \tilde{G}_n^{-1}(w - x_i'\delta).$$

Since $\tilde{G}_n > 0$, $\rho_i(1) \neq \rho_i(0)$ for all i.

(a) By Corollary 2.2, the implicit design is equal to

$$\pi_i = -\frac{\rho_i(0)}{\rho_i(1) - \rho_i(0)} = x_i'\delta.$$

To show Gram-consistency of π_i, note entries of the Gram matrix of the original regression take the form

$$\frac{1}{n}\sum_{i=1}^n \pi_i z_k(x_i, 1) z_\ell(x_i, 1) + (1 - \pi_i) z_k(x_i, 0) z_\ell(x_i, 0)$$

where $z_k(x_i, w) z_\ell(x_i, w)$ take one of the following forms: $1, w, w x_{i\ell}, x_{ik}, x_{i\ell} x_{ik}$. Thus, it suffices to check that

$$\frac{1}{n}\sum_{i=1}^n \pi_i^* = \frac{1}{n}\sum_{i=1}^n \pi_i$$
$$\frac{1}{n}\sum_{i=1}^n \pi_i^* x_i = \frac{1}{n}\sum_{i=1}^n \pi_i x_i.$$

[42] Staggered adoption corresponds to the case $\mathbf{w}_t \leq \mathbf{w}_{t+1}$ for all t. If $\mathcal{W}$ satisfies staggered adoption and excludes always-treated units, then it contains linearly independent vectors whose span excludes 1_T.

[43] One might wish to impose that the post-treatment weights are nonnegative (i.e., $\boldsymbol{\omega}_t(\boldsymbol{\pi}^*, \mathbf{w}) \geq 0$ if $\mathbf{w}_t = 1$). Failure of this condition implies that post-treatment units are severely used as comparisons for newly treated units, echoing the "forbidden comparison" issue in the recent difference-in-differences literature (Roth *et al.*, 2023; Borusyak *et al.*, 2024b; De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021). Proposition C.14 shows that when $\mathcal{W}$ only has two elements and includes a never treated unit, all weights post treatment are non-negative, but such forbidden comparisons are possible in *all* other cases.

Both are true since π_i is the projection of π_i^* onto x_i, which contains a constant.

(b) Immediately, $\pi_i^* = \pi_i$ implies $\pi_i^* = x_i'\delta$. On the other hand, if $\pi_i^* = x_i'\delta$, then the corresponding implicit design is equal to $x_i'\delta$ by (a).

(c) We may compute

$$\omega_i = \pi_i \rho_i(1) = \tilde{G}_n^{-1} x_i'\delta(1 - x_i'\delta).$$

and that

$$\begin{aligned}
\tilde{G}_n &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W_i \sim \pi_i^*}\left[W_i^2 - 2W_i\pi_i + \pi_i^2\right] \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^n \pi_i^*}_{= \frac{1}{n}\sum_{i=1}^n \pi_i \text{ since } x_i \text{ includes a constant}} - 2\frac{1}{n}\sum_{i=1}^n \pi_i^*\pi_i + \frac{1}{n}\sum_{i=1}^n \pi_i^2 \\
&= \frac{1}{n}\sum_{i=1}^n \pi_i - 2\frac{1}{n}\sum_{i=1}^n (\underbrace{\pi_i^* - \pi_i}_{\text{orthogonal to } \pi_i} + \pi_i)\pi_i + \frac{1}{n}\sum_{i=1}^n \pi_i^2 \\
&= \frac{1}{n}\sum_{i=1}^n \pi_i(1 - \pi_i).
\end{aligned}$$

Lastly, when $\pi_i = \pi_i^* \in (0,1)$, $\omega_i > 0$ since $\tilde{G}_n > 0$.

(2) By Theorem C.1, the potential weights of this regression are the same as those for the FWL-transformed regression

$$Y_i = \sum_j \tau_j(W_{ij} - \delta_j' x_i) + \epsilon_i$$

The Gram matrix for this regression is a positive definite matrix whose $k\ell^{\text{th}}$ entry is:

$$\tilde{G}_{n,k\ell} = \begin{cases} -\frac{1}{n}\sum_{i=1}^n \pi_i^*(k)(1 - \delta_k' x_i)\delta_\ell' x_i + \pi_i^*(\ell)\delta_k' x_i(1 - \delta_\ell x_i) + (1 - \pi_k^* - \pi_\ell^*)(\delta_k' x_i)(\delta_\ell' x_i), & \text{if } k \neq \ell \\ \frac{1}{n}\sum_{i=1}^n \pi_i^*(k)(1 - \delta_k' x_i)^2 + (1 - \pi_i^*(k))(\delta_k' x_i)^2 & \text{if } k = \ell. \end{cases}$$

As a result, the potential weights are of the form

$$\rho_i(j) = \tilde{G}_n^{-1}\left(e_j - \begin{bmatrix} \delta_1' x_i \\ \vdots \\ \delta_J' x_i \end{bmatrix}\right)$$

where e_j is the j^{th} standard basis vector and $e_0 = 0_J$ is the vector of zeros.

(a) Set $\pi_i(0) = 1 - \sum_{j=1}^J \pi_i(j)$. The equations (2.8) are then equivalent to

$$\tilde{G}_n^{-1}\sum_{j=1}^J \pi_i(j)\left(e_j - \begin{bmatrix} \delta_1' x_i \\ \vdots \\ \delta_J' x_i \end{bmatrix}\right) = 0.$$

Since $\tilde{G}_n$ is positive definite, we must then have

$$\sum_{j=1}^{J} \pi_i(j) \left(e_j - \begin{bmatrix} \delta_1' x_i \\ \vdots \\ \delta_J' x_i \end{bmatrix} \right) = 0$$

The j^{th} coordinate of the above equations is of the form

$$0 = \pi_i(j) - \pi_i(j)\delta_j' x_i - \delta_j' x_i \sum_{\ell \in \{0,\ldots,J\} \setminus \{j\}} \pi_i(\ell) = \pi_i(j) - \delta_j' x_i.$$

The implicit design is thus $\pi_i(j) = \delta_j' x_i$.

The proof for Gram consistency is analogous to the case in (1)(a).

(b) Analogous to (1)(b)

(c) Inspecting $\omega_i(\boldsymbol{\pi}, \ell)$, we compute that

$$\omega_i(\boldsymbol{\pi}, \ell) = \tilde{G}_n^{-1} \pi_i(\ell) \left(e_\ell - \begin{bmatrix} \pi_i(1) \\ \vdots \\ \pi_i(J) \end{bmatrix} \right)$$

Thus the implicit estimand is

$$\begin{bmatrix} \tau_1 \\ \vdots \\ \tau_J \end{bmatrix} = \frac{1}{n} \sum_i \sum_{\ell=0}^{J} \omega_i(\boldsymbol{\pi}, \ell) y_i(\ell). \tag{B.2}$$

For the contamination statement, it suffices to show an example. The setup in Example B.1 features a contaminated estimand.

(3) The specification is equivalent to the following specification:

$$Y_i = \sum_{\ell=0}^{L} \alpha_\ell + \tau_\ell W_i \tag{B.3}$$

where $\tau = \left(1 - \sum_{\ell=1}^{L} \overline{x}_\ell\right) \tau_0 + \sum_{\ell=1}^{L} \overline{x}_\ell \tau_\ell \equiv \sum_{\ell=0}^{L} \overline{x}_\ell \tau_\ell$.

For this specification (B.3), let us order the covariates to be

$$[x_{i0}, x_{i0}W_i, x_{i1}, x_{i1}W_i, \ldots, x_{iL}, x_{iL}W_i].$$

Note that the population Gram matrix is of the form

$$G_n = \begin{bmatrix} G_0 & & & \\ & G_1 & & \\ & & \ddots & \\ & & & G_L \end{bmatrix}$$

where

$$G_\ell = \begin{bmatrix} \overline{x}_\ell & \overline{x}_\ell \pi(\ell) \\ \overline{x}_\ell \pi(\ell) & \overline{x}_\ell \pi(\ell) \end{bmatrix}$$

where $\pi(\ell)$ is the mean of π_i^* among those with $c_i = \ell$. Since G_n is block-diagonal, its inverse is similarly block-diagonal with

$$G_\ell^{-1} = \frac{1}{\overline{x}_\ell \pi(\ell)(1-\pi(\ell))} \begin{bmatrix} \pi(\ell) & -\pi(\ell) \\ -\pi(\ell) & 1 \end{bmatrix}$$

on the diagonal. Note that $\Lambda = [0, \overline{x}_0, 0, \overline{x}_1, \ldots, 0, \overline{x}_L]$. Thus

$$\Lambda G_n^{-1} = \left[\frac{-\pi(0)}{\pi(0)(1-\pi(0))}, \frac{1}{\pi(0)(1-\pi(0))}, \ldots, \frac{-\pi(L)}{\pi(L)(1-\pi(L))}, \frac{1}{\pi(L)(1-\pi(L))} \right].$$

For someone with $x_i = \ell$, the covariate transform is of the form

$$z(x_i, w) = [0, \ldots, 0, 1, w, 0, \ldots, 0]'$$

where it is nonzero at the ℓ^{th} pair. Thus

$$\rho_i(w) = \Lambda G_n^{-1} z(x_i, w) = \frac{w - \pi(\ell)}{\pi(\ell)(1-\pi(\ell))}.$$

(a) By Corollary 2.2, the only implicit design sets $\pi_i = \pi(\ell)$ for $c_i = \ell$. This design is Gram-consistent by our explicit calculation of the Gram matrix.

(b) Analogous to (1)(b).

(c) Note that the implicit estimand is described by $\omega_i = \pi(\ell)\rho_i(1) = 1$. Therefore the implicit estimand is the average treatment effect.

(4) The proof for this result is rather long and is delegated to self-contained claims. In the notation setup by Proposition C.9, the potential weights are

$$\rho_i(w) = \tilde{G}_n^{-1}(w - \alpha_0 - \alpha_1'(x_i - \overline{x}) - \Gamma_{21}'(w(x_i - \overline{x}_1) - \Gamma_{01} - \Gamma_{11}(x_i - \overline{x}_1)))$$

where

$$\tilde{G}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W_i \sim \pi_i^*} \left[\left\{ W_i - \alpha_0 - \alpha_1'(x_i - \overline{x}) - \Gamma_{21}'(W_i(x_i - \overline{x}_1) - \Gamma_{01} - \Gamma_{11}(x_i - \overline{x}_1)) \right\}^2 \right].$$

(a) The implicit design is shown in Proposition C.10. Proposition C.10 shows that the odds for the implicit design is equal to

$$\pi_i/(1-\pi_i) = a_0 + a_1'(x - \overline{x}_0)$$

where a_0, a_1 are population projection coefficients of $\pi_i^*/(1-\pi_i^*)$ on $x - \overline{x}_0$, weighted by $1 - \pi_i^*$. Since it does not matter where we center x_i for this regression, as it includes an intercept, we can reparametrize to show that $\pi_i/(1-\pi_i) = \delta_0 + \delta_1'(x - \overline{x})$ as described.

(b) Immediately, $\pi_i^* = \pi_i$ implies that $\pi_i^*/(1-\pi_i^*) = \delta_0 + \delta_1'(x - \overline{x})$. On the other hand, if $\pi_i^*/(1-\pi_i^*) = \delta_0 + \delta_1'(x-\overline{x})$, then the corresponding implicit design has odds equal to $\pi_i/(1-\pi_i) = \delta_0 + \delta_1'(x-\overline{x}) = \pi_i^*/(1-\pi_i^*)$.

(c) We calculate, in the notation of Proposition C.9

$$\tilde{G}_n \rho_i(1)\pi_i = \left\{(1 - \Gamma_{21}'(x_i - \overline{x}_1)) - \theta_{01} - \theta_{11}'(x_i - \overline{x})\right\}\pi_i.$$

By the proof of Proposition C.10,

$$(1 - \Gamma_{21}'(x_i - \overline{x}_1)) - \theta_{01} - \theta_{11}'(x_i - \overline{x})$$

does not depend on i, since $\Gamma_{21} = -\theta_{11}$. As a result, $\rho_i(1)\pi_i$ is proportional to π_i.

It suffices to show that $\frac{1}{n}\sum_i \rho_i(1)\pi_i = 1$. Showing this is rather tedious by computing $\tilde{G}_n$. Alternatively, we note that $\omega_i = \rho_i(1)\pi_i$ do not depend on individual outcomes. When $\pi_i^* = \pi_i$, the regression estimates $\frac{1}{n}\sum_{i=1}^n \omega_i(y_i(1) - y_i(0))$ by Corollary 2.2. If the outcomes are such that the model is well-specified: $y_i(w) = vw$, then the regression estimates $\tau = v$ exactly for all $v \in \mathbb{R}$. But for this to happen, we must have

$$\tau = \frac{1}{n}\sum_{i=1}^n \omega_i(y_i(1) - y_i(0)) = v\frac{1}{n}\sum_{i=1}^n \omega_i = v.$$

As a result $\frac{1}{n}\sum_{i=1}^n \omega_i = 1$.

(5) Let $\pi(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \pi_i^*(\mathbf{w})$. By Theorem C.1, the potential weights of this regression are equal to those of the regression

$$\mathbf{Y}_{it} = \ddot{\mathbf{W}}_{it} + \epsilon_{it}$$

where

$$\ddot{\mathbf{w}} = \mathbf{w} - 1_T\frac{1_T'}{T}\mathbf{w} - \sum_{\mathbf{w}_1 \in \mathcal{W}} \pi(\mathbf{w}_1)\mathbf{w}_1 + 1_T\frac{1_T'}{T}\sum_{\mathbf{w}_1 \in \mathcal{W}} \pi(\mathbf{w}_1)\mathbf{w}_1.$$

For the FWL-transformed regression, the Gram matrix $\tilde{G}_n$ is a positive scalar by Theorem C.1.

(a) Any implicit design $\tilde{\pi}_i$ then satisfies

$$0 = \tilde{G}_n^{-1} \cdot \sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}_i(\mathbf{w})\ddot{\mathbf{w}} = \tilde{G}_n^{-1}\left\{\sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}_i(\mathbf{w})\mathbf{w} - 1_T\sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}_i(\mathbf{w})\frac{1_T'\mathbf{w}}{T} - \sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w})\mathbf{w} + 1_T\sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w})\frac{1_T'}{T}\mathbf{w}\right\}$$

Note that $\tilde{\pi}_i(\mathbf{w}) = \pi(\mathbf{w})$ solves this equation, and so $\pi(\mathbf{w})$ is an implicit design. We now show its uniqueness.

The preceding display implies that the vector

$$\sum_{\mathbf{w}\in\mathcal{W}} \left(\tilde{\pi}_i(\mathbf{w}) - \pi(\mathbf{w})\right)\mathbf{w} \in \text{span}(1_T)$$

However, the only vector in $\text{span}(1_T)$ and $\text{span}(\mathcal{W})$ is the zero vector by assumption. Thus $\sum_{\mathbf{w}\in\mathcal{W}} \left(\tilde{\pi}_i(\mathbf{w}) - \pi(\mathbf{w})\right)\mathbf{w} = \sum_{\mathbf{w}\in\mathcal{W},\mathbf{w}\neq 0} \left(\tilde{\pi}_i(\mathbf{w}) - \pi(\mathbf{w})\right)\mathbf{w} = 0$. Since $\mathcal{W} \setminus \{0\}$ is a linearly independent collection of vectors, we conclude that $\tilde{\pi}_i(\mathbf{w}) = \pi(\mathbf{w})$ for all $\mathbf{w} \neq 0$. Since

both probability vectors sum to one, we conclude that $\tilde{\pi}_i(\mathbf{w}) = \pi(\mathbf{w})$. Therefore, $\pi(\mathbf{w})$ is the unique implicit design.

For the Gram consistency of $\pi(\cdot)$, observe that the Gram matrix of the original regression is equal to

$$G_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\mathbf{z}(\mathbf{w})'\mathbf{z}(\mathbf{w}) = \sum_{\mathbf{w}\in\mathcal{W}} \left\{\frac{1}{n}\sum_{i=1}^{n}\pi_i^*(\mathbf{w})\right\}\mathbf{z}(\mathbf{w})'\mathbf{z}(\mathbf{w})$$

and thus depends on $\boldsymbol{\pi}_i^*$ only through $\pi(\mathbf{w})$. Therefore $G_n(\boldsymbol{\pi}) = G_n$.

(b) Immediately, if π_i^* is equal to the implicit design, then it must be constant for all i. Conversely, if π_i^* is constant for all i, then the implicit design in (a) is equal to it.

(c) For $\pi_i(\mathbf{w}) = \pi(\mathbf{w})$, define

$$\mathbf{w}_\pi = \sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w})\mathbf{w} \quad g_t(\mathbf{w}) = \ddot{\mathbf{w}}_t = \mathbf{w}_t - \mathbf{w}_{\pi,t} - \frac{1}{T}(1_T'\mathbf{w} - 1_T'\mathbf{w}_\pi).$$

Note that

$$\tilde{G}_n = \sum_{t=1}^{T}\sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w})g_t(\mathbf{w})^2.$$

Then by definition,

$$\boldsymbol{\omega}_t(\boldsymbol{\pi}, \mathbf{w}) = \pi(\mathbf{w})\boldsymbol{\rho}_{it}(\mathbf{w}) = \frac{\pi(\mathbf{w})g_t(\mathbf{w})}{\sum_{t=1}^{T}\sum_{\mathbf{w}\in\mathcal{W}}\pi(\mathbf{w})g_t(\mathbf{w})^2}. \tag{B.4}$$

Inspecting Theorem 1(ii) in Athey and Imbens (2022), in their expression, for $\mathbf{w}$ having adoption time $a \in \{2, \ldots, T, \infty\}$, the weight put on the potential outcome $y_t(\mathbf{w})$ is equal to

$$\gamma_{t,a} = \frac{\pi(\mathbf{w})g(t,a)}{\sum_{t=1}^{T}\sum_{\mathbf{w}\in\mathcal{W}}\pi(\mathbf{w})g(t,a(\mathbf{w}))^2}$$

for the expression in their (3.1)

$$\begin{aligned} g(t, a(\mathbf{w})) = &\left(\mathbb{1}(a(\mathbf{w}) \le t) - \sum_{\mathbf{w}\in\mathcal{W}:a(\mathbf{w})\le t}\pi(\mathbf{w})\right) \\ &+ \frac{1}{T}\left(a(\mathbf{w})\mathbb{1}(a(\mathbf{w}) \le T) - \sum_{\mathbf{w}\in\mathcal{W}\setminus\{0_T\}} a(\mathbf{w})\pi(\mathbf{w})\right) + \frac{T+1}{T}(\mathbb{1}(a = \infty) - \pi(0_T)). \end{aligned}$$

Here, we let $a(\mathbf{w})$ denote the adoption time of a treatment path $\mathbf{w}$.

Now, observe that

$$\begin{aligned} \mathbb{1}(a(\mathbf{w}) \le t) &= \mathbf{w}_t \\ \sum_{\mathbf{w}\in\mathcal{W}:a(\mathbf{w})\le t}\pi(\mathbf{w}) &= \sum_{\mathbf{w}\in\mathcal{W}}\mathbf{w}_t\pi(\mathbf{w}) = \mathbf{w}_{\pi,t} \end{aligned}$$

$$a(\mathbf{w})\mathbb{1}(a(\mathbf{w}) \leq T) = \begin{cases} 0 & \mathbf{w} = 0_T \\ \text{number of untreated periods } +1 = T + 1 - 1_T'\mathbf{w} & \text{otherwise.} \end{cases}$$

$$\sum_{\mathbf{w} \in \mathcal{W} \setminus \{0_T\}} a(\mathbf{w})\pi(\mathbf{w}) = T + 1 - 1_T'\mathbf{w}_\pi - (T+1)\pi(0_T)$$

$$\text{Therefore, } \left(a(\mathbf{w})\mathbb{1}(a(\mathbf{w}) \leq T) - \sum_{\mathbf{w} \in \mathcal{W} \setminus \{0_T\}} a(\mathbf{w})\pi(\mathbf{w}) \right)$$
$$= 1_T'\mathbf{w}_\pi + \begin{cases} (T+1)(\pi(0_T) - 1) & \mathbf{w} = 0_T \\ -1_T'\mathbf{w} + (T+1)\pi(0_T) & \text{otherwise} \end{cases}$$

Therefore,

$$g(t,a) = \mathbf{w}_t - \mathbf{w}_{\pi,t} - \frac{1}{T}\left(1_T'\mathbf{w} - 1_T'\mathbf{w}_\pi\right) = g_t(\mathbf{w})$$

and these expressions match. □

Example B.1. It is simple to analyze Theorem 3.1(2) fully numerically, at least given a concrete data-generating process. To do so, we follow the numerical example in Section 2.2 of Goldsmith-Pinkham *et al.* (2024). Consider the specification

$$Y_i = \alpha_0 + \tau_1 W_{i1} + \tau_2 W_{i2} + \gamma' x_i + \epsilon_i$$

with binary x_i and $J = 2$. Suppose exactly half the units have $x_i = 1$. For the units with $x_i = 0$, $\pi_i^*(0) = 0.5$, $\pi_i^*(1) = 0.05$, and $\pi_i^*(2) = 0.45$. For the units with $x_i = 1$, $\pi_i^*(0) = 0.1$, $\pi_i^*(1) = 0.45$, and $\pi_i^*(2) = 0.45$. The coefficients of interest are $\boldsymbol{\tau} = [\tau_1, \tau_2]'$.

Numerically, we can verify that the assignment probabilities $\boldsymbol{\pi}^*$ are the only solution to (2.8), and thus $\boldsymbol{\pi}^*$ is the only implicit design for this regression specification. The corresponding implicit estimand for τ_j is then

$$\tau_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{J} \underbrace{\pi_i(k)\rho_{ij}(k)}_{\omega_{ij}(\boldsymbol{\pi},k)} y_i(k),$$

which is the sum of potential outcomes weighted by $\omega_{ij}(\boldsymbol{\pi}, k)$.

We can calculate the implicit estimand as well, and inspecting the implicit estimand allows us to recover the contamination bias in Goldsmith-Pinkham *et al.* (2024). For each x_i value and each estimand, we tabulate $\omega_{ij}(\cdot)$ for individuals with those x_i:

	τ_1			τ_2		
	$\omega_{i1}(0)$	$\omega_{i1}(1)$	$\omega_{i1}(2)$	$\omega_{i2}(0)$	$\omega_{i2}(1)$	$\omega_{i2}(2)$
$x_i = 0$	$-140/106$	$41/106$	$99/106$	$-160/106$	$9/106$	$151/106$
$x_i = 1$	$-72/106$	$171/106$	$-99/106$	$-52/106$	$-9/106$	$61/106$

We find that the implicit estimand for τ_j does not solely involve $y_i(0)$ and $y_i(j)$, contrary to intuition; moreover, τ_j does not measure the same estimand for individuals with $x_i = 0$ and $x_i = 1$. This echoes the result by Goldsmith-Pinkham *et al.* (2024). In fact, under $\boldsymbol{\pi}^*$, the estimand τ_1 is

$$\tau_1 = \frac{1}{2}\left[\frac{41}{106}\overline{\tau}_{1|x=0} + \frac{99}{106}\overline{\tau}_{2|x=0}\right] + \frac{1}{2}\left[\frac{171}{106}\overline{\tau}_{1|x=1} - \frac{99}{106}\overline{\tau}_{2|x=1}\right]$$

where $\overline{\tau}_{k|x=j}$ is the mean of $y_i(k) - y_i(0)$ among those with $x_i = j$. This decomposition exactly matches the decomposition (7) in Goldsmith-Pinkham *et al.* (2024). ■

Proposition 3.2. *Consider the specification (3.1) and let $\boldsymbol{\tau} = (\tau_0, \tau_1')'$ be the coefficients of interest. Then its corresponding implicit design exists if and only if, for some conformable Γ_0, Γ_1, for all i,*

$$(\delta_0 + \delta_1' x_{2i})x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$$

where δ_0, δ_1 are population projection coefficients of π_i^* on x_{2i}: $\pi_i^* = \delta_0 + \delta_1' x_{2i} + \eta_i$ and $\pi_i^* x_{1i} = \Gamma_0 + \Gamma_1 x_{2i} + \upsilon_i$. When this happens, the unique implicit design is $\pi_i = \delta_0 + \delta_1' x_{i2}$. Therefore, if $\boldsymbol{\tau}$ is a vector of causal contrasts under $\boldsymbol{\pi}^*$, then $\pi_i^* = \delta_0 + \delta_1' x_{2i}$ and $\pi_i^* x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$.

Proof. By Theorem C.1, we can analyze the regression

$$Y_i = \tau_0(W_i - \delta_0 - \delta_1' x_{2i}) + \tau_1'\left(W_i x_{1i} - \Gamma_0 - \Gamma_1 x_{2i}\right) + \epsilon_i$$

instead. Here, δ_0, δ_1 are population projection coefficients of π_i^* on $1, x_{2i}$ and Γ_0, Γ_1 are population projection coefficients of $\pi_i^* x_{1i}$ on $1, x_{2i}$. The potential weights for this FWL-transformed regression is

$$\rho_i(w) = \tilde{G}_n^{-1}\begin{bmatrix} w - \delta_0 - \delta_1' x_{i2} \\ w x_{1i} - \Gamma_0 - \Gamma_1 x_{2i} \end{bmatrix},$$

where $\tilde{G}_n$ is a positive definite matrix by Theorem C.1. Now, the implicit design solves (2.8)

$$\tilde{G}_n^{-1}\left(\pi_i \begin{bmatrix} 1 - \delta_0 - \delta_1' x_{i2} \\ x_{1i} - \Gamma_0 - \Gamma_1 x_{2i} \end{bmatrix} + (1 - \pi_i)\begin{bmatrix} -\delta_0 - \delta_1' x_{i2} \\ -\Gamma_0 - \Gamma_1 x_{2i} \end{bmatrix}\right) = 0.$$

This equation admits a solution if and only if

$$\pi_i \begin{bmatrix} 1 - \delta_0 - \delta_1' x_{i2} \\ x_{1i} - \Gamma_0 - \Gamma_1 x_{2i} \end{bmatrix} + (1 - \pi_i)\begin{bmatrix} -\delta_0 - \delta_1' x_{i2} \\ -\Gamma_0 - \Gamma_1 x_{2i} \end{bmatrix} = 0,$$

which immediately rearranges into

$$\pi_i = \delta_0 + \delta_1' x_{i2} \quad \pi_i x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}.$$

If the implicit design exists, then π_i satisfies the above display, meaning that $(\delta_0 + \delta_1' x_{i2})x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$ for all i. On the other hand, if $(\delta_0 + \delta_1' x_{i2})x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$, then setting $\pi_i = \delta_0 + \delta_1' x_{i2}$ would solve these equations. Hence an implicit design exists iff $(\delta_0 + \delta_1' x_{i2})x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$.

When this happens, $\pi_i = \delta_0 + \delta_1' x_{i2}$. Applying Theorem 2.1, if $\boldsymbol{\tau}$ is a vector of causal contrasts, then $\pi_i^* = \pi_i = \delta_0 + \delta_1' x_{i2}$ and $(\delta_0 + \delta_1' x_{i2}) x_{1i} = \pi_i^* x_{1i} = \Gamma_0 + \Gamma_1 x_{2i}$, since an implicit design exists. □

Proposition 3.3. *Assume $\mathcal{W} \subset \{0,1\}^T$. Consider the regression $\mathbf{Y}_{it} = \alpha_i + \gamma_t + \tau \mathbf{W}_{it} + \delta' \mathbf{x}_{it}$ where τ is the coefficient of interest. Let $\beta_{w \to x}$ be the population projection coefficient of W_{it} on x_{it} under $\boldsymbol{\pi}^*$, including individual and time fixed effects. For the set of implicit designs to be non-empty, a necessary condition is*

$$\left(\mathbf{x}_i - \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \right) \beta_{w \to x} \in \operatorname{span}(\mathcal{W} \cup \{1_T\}) \text{ for all } i \in [n].$$

When $\beta_{w \to x} = 0$, if $\mathcal{W}$ contains linearly independent vectors whose span excludes 1_T, the implicit design is uniquely equal to $\pi_i(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \pi_i^*(\mathbf{w})$ for all i as in Theorem 3.1(5).

Proof. By Theorem C.1, we can focus on the population regression of $\mathbf{y}_i$ on $\ddot{\mathbf{W}}_i - \ddot{\mathbf{x}}_i \beta_{w \to x}$. Here,

$$\ddot{\mathbf{w}}_t = \mathbf{w}_t - \frac{1'\mathbf{w}}{T} - \frac{1}{n} \sum_{i=1}^n \sum_{\tilde{\mathbf{w}} \in \mathcal{W}} \pi_i^*(\tilde{\mathbf{w}}) \tilde{\mathbf{w}}_t + \frac{1}{n} \sum_{i=1}^n \sum_{\tilde{\mathbf{w}} \in \mathcal{W}} \pi_i^*(\tilde{\mathbf{w}}) \frac{1'\tilde{\mathbf{w}}_t}{T} \qquad \text{(Scalar)}$$

and

$$\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \frac{1'\mathbf{x}_i}{T} - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{it} + \frac{1}{n} \sum_{i=1}^n \frac{1'\mathbf{x}_i}{T}. \qquad \text{(Dimension } 1 \times K\text{)}$$

The potential weight vector is then

$$\boldsymbol{\rho}_i(\mathbf{w}) = \frac{\ddot{\mathbf{w}} - \ddot{\mathbf{x}}_i \beta_{w \to x}}{\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbf{W}_i \sim \pi_i^*(\cdot)} \left[(\ddot{\mathbf{W}}_i - \ddot{\mathbf{x}}_i \beta_{w \to x})' (\ddot{\mathbf{W}}_i - \ddot{\mathbf{x}}_i \beta_{w \to x}) \right]}.$$

where the denominator is strictly positive by Theorem C.1.

The condition (2.8) is then

$$0 = \sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) \rho_i(\mathbf{w}) \implies \ddot{\mathbf{x}}_i \beta_{w \to x} = \sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) \ddot{\mathbf{w}} \qquad \text{(B.5)}$$

The right-hand side is a linear combination of the columns of $\mathcal{W}$ and 1_T. Thus, a necessary condition for the existence of implicit designs is that $\ddot{\mathbf{x}}_i \beta_{w \to x} \in \operatorname{span}(\mathcal{W} \cup \{1_T\})$. The left-hand side is further more

$$\left(\mathbf{x}_i - \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \right) \beta_{w \to x} + c 1_T$$

for some scalar c. Thus, a necessary condition is furthermore

$$\left(\mathbf{x}_i - \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \right) \beta_{w \to x} \in \operatorname{span}(\mathcal{W} \cup \{1_T\}). \qquad \square$$

When we know that $\beta_{w\to x} = 0$, then the condition (2.8) is

$$\sum_{\mathbf{w}\in\mathcal{W}} \pi_i(\mathbf{w})\ddot{\mathbf{w}} = 0.$$

By the same argument as for Theorem 3.1(5)(a), the unique implicit design is $\pi_i(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \pi_i^*(\mathbf{w})$.

Proposition 3.4. *Let the population regression specification be characterized by (3.2) under $\boldsymbol{\pi}^*$, where τ is the coefficient of interest. Suppose $\mathcal{W}$ satisfies staggered adoption and excludes always-treated units. Assume further that there is a commonly observed period $\mathcal{T} \subset \bigcap_{i=1}^n \mathcal{T}_i$ such that the treatment paths have rich variation in $\mathcal{T}$.*

Let $\overline{\pi}(\mathbf{w}) \equiv \frac{1}{n}\sum_{i=1}^n \pi_i^*(\mathbf{w})$. Let $Q_i(\mathbf{w}) = \frac{\sum_{t\in\mathcal{T}_i}\mathbf{w}_t}{|\mathcal{T}_i|}$ be the proportion of treated periods within unit i's observed period for treatment path $\mathbf{w}$. Then an implicit design exists if and only if $\boldsymbol{\pi}^*$ is uncorrelated with the missingness pattern in the sense that for all $t \in [T]$:

$$\underbrace{\sum_{\mathbf{w}\in\mathcal{W}} \overline{\pi}(\mathbf{w}) \sum_{i:t\in\mathcal{T}_i} (\mathbf{w}_t - Q_i(\mathbf{w}))}_{\sum_{i:t\in\mathcal{T}_i}\mathbb{E}_{\mathbf{w}\sim\overline{\pi}}[(\mathbf{w}_t - Q_i(\mathbf{w}))]} = \underbrace{\sum_{\mathbf{w}\in\mathcal{W}} \sum_{i:t\in\mathcal{T}_i} \pi_i^*(\mathbf{w})(\mathbf{w}_t - Q_i(\mathbf{w}))}_{\sum_{i:t\in\mathcal{T}_i}\mathbb{E}_{\mathbf{w}\sim\pi_i^*}[(\mathbf{w}_t - Q_i(\mathbf{w}))]}. \tag{3.3}$$

When this happens, the implicit design is unique and satisfies $\pi_i(\mathbf{w}) = \overline{\pi}(\mathbf{w})$ for all i.

Proof. For a given unit i and $t \in \mathcal{T}_i$, let $L_i = |\mathcal{T}_i|$ and let $R_i(\mathbf{w}) = \sum_{s\in\mathcal{T}_i}\mathbf{w}_s$ so that $Q_i(\mathbf{w}) = L_i^{-1}R_i(\mathbf{w})$. The covariate transform—where we specify the unit fixed effect via within transformation—is

$$z_{it}(\mathbf{w})' = [\dot{\mathbb{1}}_{it1}, \ldots, \dot{\mathbb{1}}_{itT}, \dot{\mathbf{w}}_t]$$

where

$$\dot{\mathbb{1}}_{its} = \begin{cases} 0 & s \notin \mathcal{T}_i \\ 1 - \frac{1}{L_i} & s = t \\ -\frac{1}{L_i} & s \in \mathcal{T}_i \setminus \{t\} \end{cases}$$

is the demeaned time dummy and

$$\dot{\mathbf{w}}_t = \mathbf{w}_t - Q_i(\mathbf{w}).$$

By Theorem C.1, it suffices to analyze the potential weights of regressing Y_{it} on $\ddot{W}_{it}$, where

$$\ddot{W}_{it} = \dot{W}_{it} - \sum_{s=1}^T \delta_s \dot{\mathbb{1}}_{its} = \dot{W}_{it} - \sum_{s\in\mathcal{T}_i} \delta_s \dot{\mathbb{1}}_{its}.$$

and δ_s are the population projection coefficients of $\dot{W}_{it}$ on $\dot{\mathbb{1}}_{its}$. Let $V \equiv \frac{1}{n}\sum_{i=1}^n \sum_{t\in\mathcal{T}_i}\mathbb{E}[\ddot{W}_{it}^2]$, which is strictly positive by Theorem C.1. The potential weights are then

$$\boldsymbol{\rho}_{it}(\mathbf{w}) = V^{-1}\left(\dot{\mathbf{w}}_t - \sum_{s\in\mathcal{T}_i}\delta_s\dot{\mathbb{1}}_{its}\right) = V^{-1}\left(\dot{\mathbf{w}}_t - \delta_t + \frac{1}{L_i}\sum_{s\in\mathcal{T}_i}\delta_s\right).$$

An implicit design $\boldsymbol{\pi}$ exists if and only if there is some $\pi_i(\cdot)$ and $\sum_{\mathbf{w}} \pi_i(\mathbf{w}) = 1$ such that

$$\sum_{\mathbf{w}\in\mathcal{W}} \pi_i(\mathbf{w})\boldsymbol{\rho}_{it}(\mathbf{w}) = 0 \text{ for all } i \text{ and } t \in \mathcal{T}_i.$$

This is further equivalent to that for all i, $t \in \mathcal{T}_i$,

$$\sum_{\mathbf{w}\in\mathcal{W}} \pi_i(\mathbf{w}) \left(\mathbf{w}_t - \frac{R_i(\mathbf{w})}{L_i} \right) = \delta_t - \frac{1}{L_i} \sum_{s\in\mathcal{T}_i} \delta_s. \tag{B.6}$$

Suppose first that such a $\boldsymbol{\pi}$ exists. Fix $t \in \mathcal{T}$ and consider two units i, j, where $\mathcal{T}_i \cap \mathcal{T}_j \supset \mathcal{T}$ by assumption. Then, by (B.6), we have that

$$\delta_t = \sum_{\mathbf{w}\in\mathcal{W}} \pi_i(\mathbf{w})\mathbf{w}_t - \sum_{\mathbf{w}\in\mathcal{W}} \pi_i(\mathbf{w})\frac{R_i}{L_i} + \frac{1}{L_i}\sum_{s\in\mathcal{T}_i}\delta_s = \sum_{\mathbf{w}\in\mathcal{W}} \pi_j(\mathbf{w})\mathbf{w}_t - \sum_{\mathbf{w}\in\mathcal{W}} \pi_j(\mathbf{w})\frac{R_j}{L_j} + \frac{1}{L_j}\sum_{s\in\mathcal{T}_j}\delta_s.$$

Thus, rearranging, we have for all $t \in \mathcal{T}$,

$$\sum_{\mathbf{w}\in\mathcal{W},\mathbf{w}\neq 0} (\pi_i(\mathbf{w}) - \pi_j(\mathbf{w}))\mathbf{w}_t = M(i,j)$$

where $M(i,j)$ does not depend on t. Viewed as linear combinations for vectors in $\mathcal{W}_{\mathcal{T}}$, the left-hand side collecting over all $t \in \mathcal{T}$ lies in the span of $\mathcal{W}_{\mathcal{T}}$. The right-hand side lies in the span of the constant vector on $|\mathcal{T}|$ dimensions. By assumption, these two spans contain only the zero vector in common. Thus, both sides are equal to zero:

$$\sum_{\mathbf{w}\in\mathcal{W},\mathbf{w}\neq 0} (\pi_i(\mathbf{w}) - \pi_j(\mathbf{w}))\mathbf{w}_t = 0 \text{ for all } t \in \mathcal{T}.$$

By linear independence of vectors in $\mathcal{W}_{\mathcal{T}}$, we conclude that $\pi_i(\mathbf{w}) = \pi_j(\mathbf{w})$. Therefore, the implicit design must be constant across units:

$$\pi_i(\mathbf{w}) = \pi(\mathbf{w})$$

for some $\pi(\cdot)$.

Next, we show that if $\boldsymbol{\pi}$ exists, then it must equal to $\overline{\pi}(\cdot)$. The fact that δ_t are projection coefficients is equivalent to the following orthogonality conditions holding for all $t \in [T]$: For all $s \in \{1, \ldots, T\}$,

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\ddot{\mathbf{w}}_{it}\dot{\mathbb{1}}_{its} \\
&= \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\sum_{t\in\mathcal{T}_i}\left(\mathbf{w}_t - R_i(\mathbf{w})/L_i - \sum_{\ell\in\mathcal{T}_i}\delta_\ell\dot{\mathbb{1}}_{it\ell}\right)\dot{\mathbb{1}}_{its} \\
&= \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\sum_{t\in\mathcal{T}_i}\left(\mathbf{w}_t - R_i(\mathbf{w})/L_i\right)\dot{\mathbb{1}}_{its} - \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{t\in\mathcal{T}_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell\dot{\mathbb{1}}_{it\ell}\dot{\mathbb{1}}_{its}
\end{aligned}$$

$$= \frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) \sum_{t\in\mathcal{T}_i} (\mathbf{w}_t - R_i(\mathbf{w})/L_i)\, \mathbb{1}(t=s) - \frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \sum_{t\in\mathcal{T}_i} \sum_{\ell\in\mathcal{T}_i} \delta_\ell \mathbb{1}(t=\ell) \dot{\mathbb{1}}_{its}$$

($\dot{\mathbf{w}}$ and $\dot{\mathbb{1}}_i$ sum to zero over $t \in \mathcal{T}_i$)

$$= \frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) \left(\mathbf{w}_s - R_i(\mathbf{w})/L_i\right) - \frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \sum_{\ell\in\mathcal{T}_i} \delta_\ell \dot{\mathbb{1}}_{i\ell s}$$

$$= \frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) \left(\mathbf{w}_s - R_i(\mathbf{w})/L_i\right) - \frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \sum_{\ell\in\mathcal{T}_i} \delta_\ell (\mathbb{1}(\ell=s) - 1/L_i)$$

Rearranging, we have that the orthogonality condition is equivalent to

$$\frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) \left(\mathbf{w}_s - Q_i(\mathbf{w})\right) = \frac{1}{n} \sum_{i:s\in\mathcal{T}_i} \left(\delta_s - \frac{1}{L_i} \sum_{\ell\in\mathcal{T}_i} \delta_\ell\right) \tag{B.7}$$

for all $s \in [T]$.

Pick $t \in \mathcal{T}$ and consider the corresponding (B.7). Note that $t \in \mathcal{T}_i$ for all i, and hence

$$\delta_t = \frac{1}{n} \sum_{i=1}^n \frac{1}{L_i} \sum_{\ell\in\mathcal{T}_i} \delta_\ell + \sum_{\mathbf{w}\in\mathcal{W}} \underbrace{\left(\frac{1}{n} \sum_{i=1}^n \pi_i^*(\mathbf{w})\right)}_{\overline{\pi}(\mathbf{w})} \mathbf{w}_t - \frac{1}{n} \sum_{i=1}^n \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) Q_i(\mathbf{w}).$$

Given (B.6), we can plug in

$$\delta_t = \sum_{\mathbf{w}\in\mathcal{W}} \pi_i(\mathbf{w}) \left(\mathbf{w}_t - Q_i(\mathbf{w})\right) + \frac{1}{L_i} \sum_{s\in\mathcal{T}_i} \delta_s = \sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w}) \left(\mathbf{w}_t - Q_i(\mathbf{w})\right) + \frac{1}{L_i} \sum_{s\in\mathcal{T}_i} \delta_s.$$

Rearranging, we obtain that

$$\sum_{\mathbf{w}\in\mathcal{W}, \mathbf{w}\neq 0} (\pi(\mathbf{w}) - \overline{\pi}(\mathbf{w})) \mathbf{w}_t = M'(i)$$

for some $M'(i)$ that does not depend $t \in \mathcal{T}$. Again, the left-hand side, over $t \in \mathcal{T}$, is in the span of $\mathcal{W}_\mathcal{T}$, whereas the right-hand side is in the span on $1_\mathcal{T}$. As a result, both sides are equal to zero by assumption. Since vectors in $\mathcal{W}_\mathcal{T}$ are linearly independent, we have that

$$\pi(\mathbf{w}) = \overline{\pi}(\mathbf{w}) \text{ for all } \mathbf{w} \neq 0 \implies \pi(\mathbf{w}) = \overline{\pi}(\mathbf{w}) \text{ for all } \mathbf{w} \in \mathcal{W}.$$

Now, we plug $\pi_i(\mathbf{w}) = \overline{\pi}(\mathbf{w})$ into (B.6), and then plug the expression into (B.7). We obtain that for all $t \in [T]$,

$$\sum_{i:t\in\mathcal{T}_i} \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) (\mathbf{w}_t - Q_i(\mathbf{w})) = \sum_{i:t\in\mathcal{T}_i} \sum_{\mathbf{w}\in\mathcal{W}} \overline{\pi}(\mathbf{w}) (\mathbf{w}_t - Q_i(\mathbf{w})).$$

This is (3.3). Thus, we proved that if an implicit design exists, then the above condition holds and it must be uniquely equal to $\overline{\pi}(\cdot)$.

For the reverse direction, let us assume that (3.3) holds. For a given t, set

$$\delta_t = \sum_{\mathbf{w}\in\mathcal{W}} \overline{\pi}(\mathbf{w})\mathbf{w}_t, \tag{B.8}$$

and thus

$$\delta_t - \frac{1}{L_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell = \sum_{\mathbf{w}\in\mathcal{W}} \overline{\pi}(\mathbf{w})\left(\mathbf{w}_t - Q_i(\mathbf{w})\right).$$

This means that $\pi_i(\mathbf{w}) = \overline{\pi}(\mathbf{w})$ satisfies (B.6). Thus $\overline{\pi}(\cdot)$ is an implicit design upon checking that (B.8) defines the projection coefficients.

Summing over i where $t \in \mathcal{T}_i$:

$$\frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\left\{\delta_s - \frac{1}{L_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell\right\} = \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\overline{\pi}(\mathbf{w})\left(\mathbf{w}_t - Q_i(\mathbf{w})\right) = \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})(\mathbf{w}_t - Q_i(\mathbf{w})),$$

where the last equality follows from (3.3). This verifies the orthogonality condition (B.7) for t. Therefore, our choice of (B.8) does indeed equal the projection coefficients. This concludes the proof. □

Appendix C. Additional results

C.1. Invariances.

Theorem C.1 (Frisch–Waugh–Lovell, in population)**.** *Consider a population regression specification in the sense of Definition 2.2. Partition $z_t(\mathbf{x}_i, \mathbf{w})$ into $z_{t1}(\cdot)$ and $z_{t2}(\cdot)$. Suppose $\Lambda = [\Lambda_1, 0]$ loads solely on entries in z_{t1}: $\Lambda z_t = \Lambda_1 z_{t1}$. Let*

$$\Gamma = \left(\frac{1}{n}\sum_{i,t}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})z_{t1}(\mathbf{x}_i,\mathbf{w})z_{t2}(\mathbf{x}_i,\mathbf{w})'\right)\left(\frac{1}{n}\sum_{i,t}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})z_{t2}(\mathbf{x}_i,\mathbf{w})z_{t2}(\mathbf{x}_i,\mathbf{w})'\right)^{-1}$$

be the population projection matrix of z_{t1} onto z_{t2}. Define $\tilde{z}_t(\mathbf{x}_i,\mathbf{w}) = z_{t1}(\mathbf{x}_i,\mathbf{w}) - \Gamma z_{t2}(\mathbf{x}_i,\mathbf{w})$. Consider the regression specification defined by $\Lambda_1, \tilde{z}_t(\cdot)$, and

$$\tilde{G}_n = \frac{1}{n}\sum_{i,t}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\tilde{z}_t(\mathbf{x}_i,\mathbf{w})\tilde{z}_t(\mathbf{x}_i,\mathbf{w})'.$$

Then:

(i) $\Gamma, \tilde{G}_n$ are functions of the original Gram matrix G_n. Since G_n is assumed to be invertible, Γ is well-defined and $\tilde{G}_n$ is positive definite.

(ii) The potential weights associated with the two population regression specifications are the same.

Proof. Note that we can partition

$$G_n = \frac{1}{n}\sum_{i,t}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})z_t(\mathbf{x}_i,\mathbf{w})z_t(\mathbf{x}_i,\mathbf{w})' = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22}\end{bmatrix}$$

conformably where G_{11} is the Gram matrix associated with z_{t1} and G_{22} is the Gram matrix associated with z_{t2}. Since G_n is assumed to be invertible and thus positive definite, G_{11}, G_{22} are invertible and positive definite as well. This implies that $\Gamma = G_{12}G_{22}^{-1}$ and $\tilde{G}_n = G_{11} - \Gamma G_{21}$ are well-defined. This also implies that the Schur complement $\tilde{G}_n = G_{11} - \Gamma G_{21}$ is invertible (and positive definite). This proves (i).

By the block matrix inversion formula,

$$G_n^{-1} = \begin{bmatrix} (G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1} & -(G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1}G_{12}G_{22}^{-1} \\ -G_{22}^{-1}G_{21}(G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1} & G_{22}^{-1} + G_{22}^{-1}G_{21}(G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1}G_{12}G_{22}^{-1} \end{bmatrix}.$$

Hence

$$\Lambda G_n^{-1} = \begin{bmatrix} \Lambda_1(G_{11} - \Gamma G_{21})^{-1} & -\Lambda_1(G_{11} - \Gamma G_{21})^{-1}\Gamma \end{bmatrix}$$

Now, the potential weights for the original regression specification are

$$\begin{aligned} \boldsymbol{\rho}_{it}(\mathbf{w}) &= \Lambda_1(G_{11} - \Gamma G_{21})^{-1} z_{t1}(\mathbf{x}_i, \mathbf{w}) - \Lambda_1(G_{11} - \Gamma G_{21})^{-1}\Gamma z_{t2}(\mathbf{x}_i, \mathbf{w}) \\ &= \Lambda_1(G_{11} - \Gamma G_{21})^{-1} \tilde{z}_t(\mathbf{x}_i, \mathbf{w}) \\ &= \Lambda_1 \tilde{G}_n^{-1} \tilde{z}_t(\mathbf{x}_i, \mathbf{w}). \end{aligned}$$

This proves (ii). □

Theorem C.2 (Frisch–Waugh–Lovell, in sample)**.** *Consider a population regression specification in the sense of Definition 2.2. Partition $z_t(\mathbf{x}_i, \mathbf{w})$ into $z_{t1}(\cdot)$ and $z_{t2}(\cdot)$. Suppose $\Lambda = [\Lambda_1, 0]$ loads solely in entries in z_{t1}: $\Lambda z_t = \Lambda_1 z_{t1}$. Suppose the sample Gram matrix is invertible. Let*

$$\hat{\Gamma} = \left(\frac{1}{n} \sum_{i,t} z_{t1}(\mathbf{x}_i, \mathbf{W}_i) z_{t2}(\mathbf{x}_i, \mathbf{W}_i)' \right) \left(\frac{1}{n} \sum_{i,t} z_{t2}(\mathbf{x}_i, \mathbf{W}_i) z_{t2}(\mathbf{x}_i, \mathbf{W}_i)' \right)^{-1}$$

be the sample projection matrix of z_{t1} onto z_{t2}. Define $\tilde{z}_t(\mathbf{x}_i, \mathbf{w}) = z_{t1}(\mathbf{x}_i, \mathbf{w}) - \hat{\Gamma} z_{t2}(\mathbf{x}_i, \mathbf{w})$. Consider the regression specification defined by $\Lambda_1, \tilde{z}_t(\cdot)$. Then the estimated potential weights associated with the two regression specifications are the same.

Proof. The proof entirely follows from the proof of Theorem C.1 by setting $\pi_i^*(\mathbf{w}) = \mathbb{1}(\mathbf{w} = \mathbf{W}_i)$. □

Theorem C.3 (Invariance under reparameterization)**.** *Consider a population regression specification in the sense of Definition 2.2. Consider an invertible matrix M and another population specification defined by $\tilde{z}_t(\mathbf{x}_i, \mathbf{w}) = M z_t(\mathbf{x}_i, \mathbf{w})$ and $\tilde{\Lambda} = \Lambda M'$, so that the two vectors of coefficients represent the same underlying contrasts. Then the potential weights associated with the two regression specifications are the same. The estimated potential weights are also the same.*

Proof. We can easily see that the second regression specification has Gram matrix $\tilde{G}_n = MG_nM'$ (and $\widehat{\tilde{G}_n} = M\hat{G}_nM'$). Thus the potential weights for the second specification are

$$\tilde{\boldsymbol{\rho}}_{it}(\mathbf{w}) = \tilde{\Lambda}(M')^{-1}G_n^{-1}M^{-1}Mz_t(\mathbf{x}_i, \mathbf{w}) = \boldsymbol{\rho}_{it}(\mathbf{w}).$$

The corresponding equalities similarly hold for the estimated potential weights. □

C.2. **Additional consistency results.** Recall the sequence of populations setup in Appendix A. Additionally, let $\pi^*_{1,n}, \ldots, \pi^*_{n,n}$ denote a triangular array of marginal treatment assignment probabilities, and let Π^*_n denote the joint distribution of the treatment assignments $\mathbf{W}_1, \ldots, \mathbf{W}_n$ under the n^{th} population with marginals equal to $\pi^*_{1,n}, \ldots, \pi^*_{n,n}$.

Lemma C.4. *Suppose that in each Π^*_n, treatments are independently assigned according to $\pi^*_i = \pi^*_{i,n}$. Let $G_i = \sum_{t=1}^T z_t(\mathbf{x}_i, \mathbf{W}_i)z_t(\mathbf{x}_i, \mathbf{W}_i)'$. Assume that for all $1 \le j \le K, 1 \le \ell \le K$, the average second moment of $G_{i,j\ell}$ grows slower than n: as $n \to \infty$, $\frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\pi^*_{i,n}}[G^2_{i,j\ell}]\right) \to 0$. Then Assumption A.1 holds.*

Proof. Fix a coordinate (j,k). Let $S_n = \sum_{i=1}^n G_{i,jk}$ and note that $\hat{G}_{n,jk} = \frac{1}{n}S_n$. Define $\mu_n = \mathbb{E}[S_n]$ and $\sigma_n^2 = \mathrm{Var}(S_n)$. Theorem 2.2.6 in Durrett (2019) states that if $\sigma_n^2/n^2 \to 0$ then

$$\frac{S_n - \mu_n}{n} = \hat{G}_{n,jk} - G_{n,jk} \xrightarrow{p} 0.$$

Note that

$$\sigma_n^2 \le \sum_{i=1}^n \mathbb{E}[G^2_{i,jk}].$$

Thus the condition that $\frac{1}{n^2}\sum_i \mathbb{E}[G^2_{i,jk}] \to 0$ is sufficient for $\sigma_n^2/n^2 \to 0$. Thus Theorem 2.2.6 applies and $\hat{G}_{n,jk} - G_{n,jk} \xrightarrow{p} 0$ for every entry. Since there are finitely many entries, $\hat{G}_n - G_n \xrightarrow{p} 0$. □

Proposition C.5. *Suppose that, for some $0 < C < \infty$, with probability tending to one, an estimated implicit design $\hat{\boldsymbol{\pi}}$ exists and is bounded in the sense that:*

$$\max_i \max_{\mathbf{w}\in\mathcal{W}} |\hat{\pi}_i(\mathbf{w})| < C.$$

Let $\hat{\boldsymbol{\pi}}$ be a bounded estimated implicit design if it exists, and let it be an arbitrary vector of probability distributions otherwise. If $|\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty \xrightarrow{p} 0$, then $\sum_{\mathbf{w}\in\mathcal{W}} \rho_i(\mathbf{w})\hat{\pi}_i(\mathbf{w}) \xrightarrow{p} 0$. If further (A.1) holds, then the convergence is uniform:

$$\max_{i\in[n],\mathbf{w}\in\mathcal{W}} \left|\sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\rho}_i(\mathbf{w})\hat{\pi}_i(\mathbf{w})\right|_\infty \xrightarrow{p} 0.$$

Proof. Fix $\eta > 0$. Note that

$$\mathrm{P}\left[\left|\sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\rho}_i(\mathbf{w})\hat{\pi}_i(\mathbf{w})\right|_\infty > \eta\right] \le \mathrm{P}[\text{a bounded implicit design does not exist}]$$

$$+ \mathrm{P}\left[\left|\sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\rho}_i(\mathbf{w})\hat{\pi}_i(\mathbf{w})\right|_\infty > \eta, \text{a bounded implicit design exists}\right]$$

Note that if $\hat{\pi}_i$ is a bounded implicit design, then

$$\left|\sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\rho}_i(\mathbf{w})\hat{\pi}_i(\mathbf{w})\right|_\infty = \left|\sum_{\mathbf{w}\in\mathcal{W}} (\boldsymbol{\rho}_i(\mathbf{w}) - \hat{\boldsymbol{\rho}}_i(\mathbf{w}))\hat{\pi}_i(\mathbf{w})\right|_\infty \leq C(J+1)\max_{\mathbf{w}\in\mathcal{W}} |\boldsymbol{\rho}_i(\mathbf{w}) - \hat{\boldsymbol{\rho}}_i(\mathbf{w})|_\infty.$$

Thus,

$$\mathrm{P}\left[\left|\sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\rho}_i(\mathbf{w})\hat{\pi}_i(\mathbf{w})\right|_\infty > \eta\right] \leq \mathrm{P}[\text{a bounded implicit design does not exist}]$$
$$+ \mathrm{P}\left[C(J+1)\max_{\mathbf{w}\in\mathcal{W}} |\boldsymbol{\rho}_i(\mathbf{w}) - \hat{\boldsymbol{\rho}}_i(\mathbf{w})|_\infty > \eta\right] \to 0.$$

by assumption. The claim of uniform-in-i convergence follows by essentially the same argument, where we take maximum over i. □

C.2.1. *Consistency of $\hat{G}_n$ to G_n under rejective sampling.* This subsection considers a law of large numbers under a version of sampling with replacement (what Hájek, 1964, calls "rejective sampling"), and largely follows Rambachan and Roth (2020). In particular, we assume that the treatment is binary, and there is a sequence of unconditional probabilities $p_1, p_2, \ldots$ as well as a sequence of sample sizes N_n such that Π_n^* describes the joint distribution of $W_1, \ldots, W_n$ conditional on the event $\sum_{i=1}^n W_i = N_n$, where $W_i \sim \mathrm{Bern}(p_i)$ independently unconditionally. Correspondingly, let $\pi_{i,n}$ be the probability that $W_i = 1$ under Π_n^*.

Lemma C.6. *Write $\pi_i(w) = \pi_{i,n}(w)$ and omit the n subscript. Consider $\overline{y}_w = \frac{1}{n}\sum_i \mathbb{1}(W_i = w)y_i$ where $\mathbb{E}[\overline{y}_w] = \frac{1}{n}\sum_i \pi_i(w)y_i(w)$. Suppose $C_n \equiv \sum_{i=1}^n \pi_i(w)(1-\pi_i(w)) \to \infty$ as $n \to \infty$. Then, as $n \to \infty$*

$$\mathrm{Var}(\overline{y}_w) = \frac{(1+o(1))}{n}\frac{C_n}{n}\sum_{i=1}^n \frac{\pi_i(w)(1-\pi_i(w))}{C_n}\left(y_i(w) - \frac{\sum_{j=1}^n \pi_j(w)(1-\pi_j(w))y_j(w)}{C_n}\right)^2$$

Proof. This is a restatement of Theorem 6.1 in Hájek (1964). The notation y_i in the theorem corresponds to $\pi_i(w)y_i(w)$ in our notation. □

A sufficient condition for the variance to tend to zero is the following:

Assumption C.1. *For all entries k, $z_k(x_i, w)$ is uniformly bounded by $0 < M < \infty$ and $C_n(w) \equiv \sum_{i=1}^n \pi_i(w)(1-\pi_i(w)) \to \infty$ as $n \to \infty$.*

The boundedness condition for z_k is stronger than needed. In particular, what is needed is that the $\pi_i(w)(1-\pi_i(w))$-weighted variance of $y_i(w)$ is $O(1)$.

Lemma C.7. *Under rejective sampling, suppose Assumption C.1 holds, then $\hat{G}_n \xrightarrow{p} G_n$ in probability.*

Proof. We prove this claim by showing all entries converge in probability, since there are finitely many entries. Fix some j, k. Note that

$$\hat{G}_{n,jk} = \sum_{w \in \{0,1\}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(W_i = w) z_j(x_i, w) z_k(x_i, w)}_{\overline{z}_n(w)} .$$

By Lemma C.6 and Assumption C.1 applied to $\overline{z}_n$ where $y_i(w)$ is taken to be $z_j(x_i, w)z_k(x_i, w)$, we have that $\mathrm{Var}(\overline{z}_n(w)) \to 0$ as $n \to \infty$, since $C_n(w)/n \leq 1$. Since $\mathbb{E}[\overline{z}_n(w)] = \frac{1}{n}\sum_i \pi_i(w) z_j(x_i, w) z_k(x_i, w)$, we have that

$$\hat{G}_{n,jk} - G_{n,jk} = \sum_{w \in \{0,1\}} \left\{ \overline{z}_n(w) - \frac{1}{n} \sum_i \pi_i(w) z_j(x_i, w) z_k(x_i, w) \right\} \xrightarrow{p} 0$$

by Chebyshev's inequality. □

C.3. **Uniqueness of implicit designs in cross-sections.** Consider a cross-sectional setup with $J+1$ treatments where the potential weights are

$$\rho_i(w) = \tilde{\Lambda} \tilde{G}_n^{-1} \tilde{z}(x_i, w)$$

and $\tilde{\Lambda}$ is a $J \times K$ matrix with rank J. It is possible to reparametrize the regressors (i.e. choose an invertible M such that $z(\cdot) = M\tilde{z}(\cdot)$) such that

$$\rho_i(w) = \Lambda G_n^{-1} z(x_i, w)$$

and $\Lambda = [I_J, 0]$. Without essential loss of generality, let us assume the sequence of reparametrized specifications satisfy Assumption A.2.

Partition z into z_1 and z_2 where z_1 is J-dimensional. By Theorem C.1, the potential weights are further equivalent to

$$\rho_i(w) = G_{n,1|2}^{-1} z_{1|2}(x_i, w)$$

where $z_{1|2}(x_i, w) = z_1(x_i, w) - \Gamma' z_2(x_i, w)$ for population projection coefficients Γ and

$$G_{n,1|2} = \frac{1}{n} \sum_{i,w} \pi_i^*(w) z_{1|2}(x_i, w) z_{1|2}(x_i, w)'.$$

Assumption C.2. *The residualized covariate transform is non-singular in the sense that the $J \times (J+1)$ matrix*

$$Z_i \equiv [z_{1|2}(x_i, 0), \ldots, z_{1|2}(x_i, J)]$$

whose columns are covariate transforms evaluated at a treatment level has minimum singular value (that is, the J^{th} singular value) greater than some $\eta > 0$, uniformly for all i.

Assumption C.3. *The maximum eigenvalue of G_n is bounded above by $M < \infty$ for all n.*

Lemma C.8. *Under Assumptions A.2, C.2, and C.3, condition (ii) in Proposition A.2 is satisfied for some lower bound $\epsilon > 0$ on the minimum (i.e. J^{th}) singular value.*

Note that since the potential weight matrix $G_{n,1|2}^{-1} Z_i$ is $J \times (J+1)$ and has J positive singular values bounded below, if there exists an implicit design, then it must be unique.

Proof. We first show that Assumption A.2 implies that the minimum eigenvalue of $G_{n,1|2}$ is bounded away from zero and maximum bounded by M. Note that $G_{n,1|2}$ is a Schur complement of a submatrix of G_n, and so $G_{n,1|2}^{-1}$ is a principal submatrix of G_n^{-1}. The eigenvalues of $G_{n,1|2}^{-1}$ interlace the eigenvalues of G_n^{-1} by Cauchy's interlace theorem (Hwang, 2004). In particular, the spectrum of $G_{n,1|2}^{-1}$ is included in the range of the spectrum of G_n^{-1}. Hence $\lambda_{\min}(G_{n,1|2}) > \epsilon$ by Assumption A.2 and $\lambda_{\max}(G_{n,1|2}) < M$ by Assumption C.3.

Next, it suffices to show that $G_{n,1|2}^{-1} Z_i$ has minimum singular value bounded below. Write the compact SVD as

$$Z_i = U \Sigma V'$$

for diagonal $J \times J$ matrix Σ with $UU' = U'U = V'V = I_J$. Similarly, write

$$G_{n,1|2}^{-1} = Q D Q'$$

for an orthogonal matrix Q and diagonal D. Then

$$G_{n,1|2}^{-1} Z_i Z_i' G_{n,1|2}^{-1} = Q D Q' U \Sigma^2 U' Q D Q'.$$

is a real symmetric matrix. The spectrum of this matrix is the same as the spectrum of

$$D \qquad \underbrace{Q' U \Sigma^2 U' Q}_{\text{positive definite matrix with spectrum } \Sigma^2} \qquad D.$$

Since $\lambda_{\min}(AB) \geq \lambda_{\min}(A)\lambda_{\min}(B)$ for two positive definite matrices A, B,[44] we have that the minimum eigenvalue of $G_{n,1|2}^{-1} Z_i Z_i' G_{n,1|2}^{-1}$ is bounded below by $\frac{\eta^2}{\lambda_{\max}(G_{n,1|2})^2}$. Thus, the minimum singular value of $G_{n,1|2}^{-1} Z_i$ is bounded below by η / M. □

C.4. **Additional results for interacted regression.** Proposition C.9 provides a unified analyses of the regression (C.1). It implies our results in Theorem 3.1(4) when $\overline{x}_t = \overline{x}_1$ or $\overline{x}_0$, but allows for evaluating on other points in the line connecting $\overline{x}_1$ and $\overline{x}_0$. The implicit design for the model-based ATE is analyzed when we set $t = \alpha_0 = \frac{1}{n} \sum_{i=1}^n \pi_i^*$.

Inspecting the resulting implicit designs (C.2), we find that they are in general fractional linear, encompassing the linear-odds result of $t = 1$ as a special case. Moreover, a necessary implication of $\pi_i^* = \pi_i$ is the condition (C.3), which is very difficult to satisfy unless $t = 0$ or $t = 1$. In this sense, the model-based ATU and the model-based ATT are the only estimands whose implicit design is reasonable. See Remark C.1 for an example where the implicit model for the ATE depends on the support of x_i.

[44] To see this, note that the operator norm (largest eigenvalue of a positive semidefinite matrix) is submultiplicative $\|AB\|_{op} \leq \|A\|_{op} \|B\|_{op}$. Apply this inequality to $A^{-1} B^{-1}$.

Proposition C.9. *Consider the specification indexed by $t \in \mathbb{R}$,*

$$Y_i = \gamma_0 + W_i\tau_t + \gamma_0 + \gamma_1' x_i + W_i(x_i - \overline{x}_t)'\gamma_2 + \epsilon_i \tag{C.1}$$

where $\overline{x}_t = t\overline{x}_1 + (1-t)\overline{x}_0$ for $\overline{x}_1 = \frac{\frac{1}{n}\sum_{i=1}^n \pi_i^* x_i}{\frac{1}{n}\sum_{i=1}^n \pi_i^*}$ and $\overline{x}_0 = \frac{\frac{1}{n}\sum_{i=1}^n (1-\pi_i^*) x_i}{\frac{1}{n}\sum_{i=1}^n (1-\pi_i^*)}$.

Let

$$A = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(x_i - \overline{x})'$$

$$V_t = \frac{1}{n}\sum_{i=1}^n \pi_i^*(x_i - \overline{x}_t)(x_i - \overline{x}_t)' \quad U_t = \frac{1}{n}\sum_{i=1}^n (1-\pi_i^*)(x_i - \overline{x}_t)(x_i - \overline{x}_t)'$$

$$\alpha_0 = \frac{1}{n}\sum_{i=1}^n \pi_i^*$$

$$\Gamma_{0t} = \frac{1}{n}\sum_{i=1}^n \pi_i^*(x_i - \overline{x}_t)$$

$$\alpha_1 = A^{-1}\frac{1}{n}\sum_{i=1}^n \pi_i^*(x_i - \overline{x}) = A^{-1}(\Gamma_{0t} + \alpha_0(\overline{x}_t - \overline{x}))$$

$$\Gamma_{1t} = \frac{1}{n}\sum_{i=1}^n \pi_i^*(x_i - \overline{x}_t)(x_i - \overline{x})' = \left(V_t + \Gamma_{0t}(\overline{x}_t - \overline{x})'\right) A^{-1}$$

$$\begin{aligned}\Gamma_{2t} = &\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[W_i(x - \overline{x}_t)\left(W_i(x - \overline{x}_t) - \Gamma_{0t} - \Gamma_{1t}(x_i - \overline{x})\right)'\right]\right)^{-1} \\ &\times \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(W_i - \alpha_0 - \alpha_1'(x_i - \overline{x}))\left(W_i(x_i - \overline{x}_t) - \Gamma_{0t} - \Gamma_{1t}(x - \overline{x})\right)\right] \\ = &\left(V_t - \Gamma_{0t}\Gamma_{0t}' - V_t\Gamma_{1t}' - \Gamma_{0t}(\overline{x}_t - \overline{x})'\Gamma_{1t}'\right)^{-1}\left(\Gamma_{0t} - \alpha_0\Gamma_{0t} - V_t\alpha_1 - \Gamma_{0t}(\overline{x}_t - \overline{x})'\alpha_1\right)\end{aligned}$$

Then, the potential weights satisfy

$$\rho_i(w) = \tilde{G}_n^{-1}\left(w(1 - \Gamma_{2t}'(x_i - \overline{x}_t)) - \underbrace{(\alpha_0 - \Gamma_{2t}'\Gamma_{0t})}_{\theta_{0t}} - \underbrace{(\alpha_1 - \Gamma_{1t}'\Gamma_{2t})'}_{\theta_{1t}'}(x_i - \overline{x})\right),$$

for some scalar $\tilde{G}_n > 0$. The implicit design π_i satisfies

$$\left(1 - \Gamma_{2t}'(x_i - \overline{x}_t)\right)\pi_i = \theta_{0t} + \theta_{1t}'(x_i - \overline{x}). \tag{C.2}$$

Moreover, if $\pi_i^* = \pi_i$, then a necessary condition is that

$$tV_1(\theta_{1t} + \Gamma_{2t}) = (1-t)U_0\theta_{1t}. \tag{C.3}$$

Proof. Observe that by the Frisch–Waugh–Lovell transform, the coefficient τ_t is the same as in the regression

$$Y_i = \tau_t \left(W_i - \alpha_0 - \alpha_1'(x_i - \overline{x}) - \Gamma_2' \left(W_i(x_i - \overline{x}_t) - \Gamma_0 - \Gamma_1(x_i - \overline{x})\right)\right) + \epsilon_i. \tag{C.4}$$

The quantities $\alpha_0, \alpha_1, \Gamma_0, \Gamma_1, \Gamma_2$ are defined exactly as these auxiliary regression coefficients. Thus, by Theorem C.1, the potential weights are equivalent to the potential weights of (C.4), which rearrange exactly to the potential weights displayed. The constraint on the implicit design corresponds exactly to

$$\rho_i(1)\pi_i + (1 - \pi_i)\rho_i(0) = 0.$$

For the "moreover" part, observe that we can rewrite Γ_{2t} as satisfying

$$V_t\Gamma_{2t} - \Gamma_{0t}(\overline{x}_t - \overline{x})'\Gamma_{1t}'\Gamma_{2t} - \Gamma_{0t}\Gamma_{0t}'\Gamma_{2t} - V_t\Gamma_{1t}'\Gamma_{2t} = \Gamma_{0t} - \alpha_0\Gamma_{0t} - V_t\alpha_1 - \Gamma_{0t}(\overline{x}_t - \overline{x})'\alpha_1.$$

Thus,

$$\begin{aligned}
0 &= \left(\alpha_0\Gamma_{0t} - \Gamma_{0t}\Gamma_{0t}'\Gamma_{2t}\right) + \left(V_t\alpha_1 - V_t\Gamma_{1t}'\Gamma_{2t}\right) \\
&\quad + \left(\Gamma_{0t}(\overline{x}_t - \overline{x})'\alpha_1 - \Gamma_{0t}(\overline{x}_t - \overline{x})'\Gamma_{1t}'\Gamma_{2t}\right) + V_t\Gamma_{2t} - \Gamma_{0t} \\
&= \Gamma_{0t}\theta_{0t} + \underbrace{V_t\theta_{1t} + \Gamma_{0t}(\overline{x}_t - \overline{x})'\theta_{1t}}_{\Gamma_{1t}A\theta_{1t}} + V_t\Gamma_{2t} - \Gamma_{0t}
\end{aligned}$$

Observe that

$$A\theta_{1t} = \Gamma_{0t} + \alpha_0(\overline{x}_t - \overline{x}) - V_t\Gamma_{2t} - (\overline{x}_t - \overline{x})\Gamma_{0t}'\Gamma_{2t}$$

and hence

$$V_t\Gamma_{2t} - \Gamma_{0t} = -A\theta_{1t} + \alpha_0(\overline{x}_t - \overline{x}) - (\overline{x}_t - \overline{x})\Gamma_{0t}'\Gamma_{2t} = -A\theta_{1t} + \theta_{0t}(\overline{x}_t - \overline{x})$$

Therefore

$$\begin{aligned}
0 &= \left(\Gamma_{0t} + (\overline{x}_t - \overline{x})\right)\theta_{0t} + (V_t - A + \Gamma_{0t}(\overline{x}_t - \overline{x})')\theta_{1t} \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(\pi_i^*(x_i - \overline{x}_t) - (x_i - \overline{x}_t))\theta_{0t} \\
&\quad + \left(\pi_i^*(x_i - \overline{x}_t)(x_i - \overline{x}_t)' - (x_i - \overline{x})(x_i - \overline{x})' + \pi_i^*(x_i - \overline{x}_t)(\overline{x}_t - \overline{x})'\right)\theta_{1t}\} \\
&= -\frac{1}{n}\sum_{i=1}^{n}(1 - \pi_i^*)(x_i - \overline{x}_t)(\theta_{0t} + \theta_{1t}'(x_i - \overline{x}))
\end{aligned}$$

Thus, if $\pi_i^* = \pi_i$,

$$\theta_{0t} + \theta_{1t}'(x_i - \overline{x}) = (1 - \Gamma_{2t}'(x_i - \overline{x}_t))\pi_i^*.$$

We thus have that

$$0 = \frac{1}{n}\sum_{i=1}^{n}\pi_i^*(1 - \pi_i^*)(x_i - \overline{x}_t)(1 - \Gamma_{2t}'(x_i - \overline{x}_t))$$

$$= t\frac{1}{n}\sum_{i=1}^{n}\pi_i^*(x_i-\overline{x}_1)(1-\pi_i^*)(1-\Gamma_{2t}'(x_i-\overline{x}_t)) + (1-t)\frac{1}{n}\sum_{i=1}^{n}(1-\pi_i^*)(x_i-\overline{x}_1)\pi_i^*(1-\Gamma_{2t}'(x_i-\overline{x}_t))$$

where we may plug in

$$\pi_i^*(1-\Gamma_{2t}'(x_i-\overline{x}_t)) = \theta_{0t}+\theta_{1t}'(x-\overline{x}) \quad (1-\pi_i^*)(1-\Gamma_{2t}'(x_i-\overline{x}_t)) = 1-\theta_{0t}-\Gamma_{2t}'(x_i-\overline{x}_t)-\theta_{1t}'(x_i-\overline{x}).$$

Simplifying by noting that $\sum_{i=1}^n \pi_i^*(x_i-\overline{x}_1)=0=\sum_{i=1}^n(1-\pi_i^*)(x_i-\overline{x}_0)$ yields the condition

$$tV_1(\theta_{1t}+\Gamma_{2t}) = (1-t)U_0\theta_{1t}.$$

□

Proposition C.10. *In Proposition C.9, for $t=1$, the implicit design is equal to*

$$\pi_i = \frac{\alpha_0 + ((I-\Gamma_1')^{-1}\alpha_1)'(x-\overline{x})}{1+((I-\Gamma_1')^{-1}\alpha_1)'(x-\overline{x}_1)}.$$

The associated odds is

$$\frac{\pi_i}{1-\pi_i} = \frac{\alpha_0 + ((I-\Gamma_1')^{-1}\alpha_1)'(x-\overline{x})}{1-\alpha_0+((I-\Gamma_1')^{-1}\alpha_1)'(\overline{x}-\overline{x}_1)} \equiv \delta_0+\delta_1'(x_i-\overline{x}_0).$$

Here,

$$\delta_1 = \left(\frac{1}{n}\sum_{i=1}^{n}(1-\pi_i^*)(x_i-\overline{x}_0)(x_i-\overline{x}_0)\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}(1-\pi_i^*)(x_i-\overline{x}_0)\frac{\pi_i^*}{1-\pi_i^*}$$

$$\delta_0 = \frac{\alpha_0}{1-\alpha_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\pi_i^*)}{1-\alpha_0}\frac{\pi_i^*}{1-\pi_i^*}$$

Proof. Here, we omit the t subscript for objects defined in Proposition C.9. Observe that $\Gamma_0=0$. Hence

$$\Gamma_2 = -(I-\Gamma_1')^{-1}\alpha_1 \text{ and } \theta_0=\alpha_0$$

and thus

$$\theta_1 = \alpha_1-\Gamma_1'\Gamma_2 = (I+\Gamma_1'(I-\Gamma')^{-1})\alpha_1 = (I-\Gamma')^{-1}\alpha_1.$$

It remains to show that δ_0,δ_1 take the form stated. First, observe that

$$\theta_1 = (I-\Gamma_1')^{-1}\alpha_1 = (A-V_1)^{-1}\alpha_0(\overline{x}_1-\overline{x}) = (A-V_1)^{-1}\alpha_0(1-\alpha_0)(\overline{x}_1-\overline{x}_0).$$

by plugging in $\alpha_1 = A^{-1}(\alpha_0(\overline{x}_1-\overline{x}))$. Here,

$$\begin{aligned}
A &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x}_0)' \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_0)(x_i-\overline{x}_0)' - (\overline{x}-\overline{x}_0)(\overline{x}-\overline{x}_0)' \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_0)(x_i-\overline{x}_0)' - \alpha_0^2(\overline{x}_1-\overline{x}_0)(\overline{x}_1-\overline{x}_0)'
\end{aligned}$$

$$
\begin{aligned}
V_1 &= \frac{1}{n}\sum_{i=1}^{n} \pi_i^*(x_i - \overline{x}_1)(x_i - \overline{x}_0) \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi_i^*(x_i - \overline{x}_0)(x_i - \overline{x}_0)' + (\overline{x}_0 - \overline{x}_1)\cdot\frac{1}{n}\sum_{i=1}^{n}\pi_i^*(x_i - \overline{x}_0)' \\
&= V_0 + \alpha_0(\overline{x}_0 - \overline{x}_1)(\overline{x}_1 - \overline{x}_0)'
\end{aligned}
$$

Therefore,

$$
A - V_1 = U_0 + \alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)(\overline{x}_1 - \overline{x}_0)'
$$

By Sherman–Morrison,

$$
(A - V_1)^{-1} = U_0^{-1} - \alpha_0(1-\alpha_0)\frac{U_0^{-1}(\overline{x}_1 - \overline{x}_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}}{1 + \alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)}
$$

Then,

$$
\begin{aligned}
\frac{\theta_1}{\alpha_0(1-\alpha_0)} &= (A - V_1)^{-1}(\overline{x}_1 - \overline{x}_0) \\
&= \left(1 - \frac{\alpha_0(1-\alpha_0)\cdot(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)}{1 + \alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)}\right) U_0^{-1}(\overline{x}_1 - \overline{x}_0) \\
&= \frac{1}{1 + \alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)} U_0^{-1}(\overline{x}_1 - \overline{x}_0) \\
\implies \theta_1'(\overline{x}_1 - \overline{x}_0) &= \frac{\alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)}{1 + \alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)}
\end{aligned}
$$

and

$$
\begin{aligned}
\theta_1'(\overline{x} - \overline{x}_1) &= -(1-\alpha_0)\theta_1'(\overline{x}_1 - \overline{x}_0) \\
&= -(1-\alpha_0)\frac{\alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)}{1 + \alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)}
\end{aligned}
$$

Hence,

$$
1 - \alpha_0 + ((I - \Gamma_1')^{-1}\alpha_1)'(\overline{x} - \overline{x}_1) = (1-\alpha_0)\frac{1}{1 + \alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)'U_0^{-1}(\overline{x}_1 - \overline{x}_0)},
$$

and therefore

$$
\delta_1 = \frac{\theta_1}{1 - \alpha_0 + ((I - \Gamma_1')^{-1}\alpha_1)'(\overline{x} - \overline{x}_1)} = U_0^{-1}\alpha_0\,(\overline{x}_1 - \overline{x}_0) = U_0^{-1}\frac{1}{n}\sum_{i=1}^{n}(1-\pi_i)\frac{\pi_i}{1-\pi_i}(x_i - \overline{x}_0),
$$

as claimed.

Now,

$$
\delta_0 = \frac{\alpha_0 + \theta_1'(\overline{x}_0 - \overline{x})}{1 - \alpha_0 + ((I - \Gamma_1')^{-1}\alpha_1)'(\overline{x} - \overline{x}_1)}
$$

$$= \frac{\alpha_0 - \theta_1' \alpha_0 (\overline{x}_1 - \overline{x}_0)}{(1-\alpha_0) \frac{1}{1+\alpha_0(1-\alpha_0)(\overline{x}_1 - \overline{x}_0)' U_0^{-1} (\overline{x}_1 - \overline{x}_0)}}$$
$$= \frac{\alpha_0}{1-\alpha_0},$$

as claimed. □

Remark C.1. Here we consider the ramifications of (C.3) for the model-based ATE (τ_{α_0}). For some true design $\boldsymbol{\pi}^*$ to justify causal interpretation, we must have that this design is fractional linear:

$$\pi_i^* = \frac{\theta_0 + \theta_1'(x_i - \overline{x})}{1 - \Gamma_2'(x_i - \overline{x})}$$

for some $\theta_0, \theta_1, \Gamma_2$. Moreover, the choices $\theta_0, \theta_1, \Gamma_2$ needs to satisfy (C.3), which depends on the covariates $x_1, \ldots, x_n$. We may thus consider the parameter space as

$$\Theta(t; x_{1:n}) = \{(\theta_0, \theta_1, \Gamma_2) : \ \tau_t \text{ is a causal contrast under the corresponding } \boldsymbol{\pi}^* \text{ under } (\theta_0, \theta_1, \Gamma_2) \}.$$

Note that for $t = 0$ and $t = 1$, this parameter space does not depend on $x_{1:n}$ aside from the obvious support restriction. For instance, if $x_{1:n}$ and $\tilde{x}_{1:n}$ are two sets of covariate values, and

$$\pi_i^*/(1-\pi_i^*) = \delta_0 + \delta_1'(x_i - \overline{x}_1) \geq 0 \quad \tilde{\pi}_i^*/(1-\tilde{\pi}_i^*) = \delta_0 + \delta_1'(\tilde{x}_i - \overline{\tilde{x}}_1) \geq 0,$$

then the corresponding parameters are in $\Theta(1; x_{1:n}) \cap \Theta(1; \tilde{x}_{1:n})$—similarly for $t = 0$. This is true by Proposition C.10.

However, for other t, and in particular for $t = \alpha_0$, it is possible that some $(\theta_0, \theta_1, \Gamma_2) \in \Theta(\alpha_0; x_{1:n})$, still does not belong to $\Theta(\alpha_0; \tilde{x}_{1:n})$, even though no support restriction is violated

$$\tilde{\pi}_i^* \equiv \frac{\theta_0 + \theta_1'(\tilde{x}_i - \overline{\tilde{x}})}{1 - \Gamma_2'(\tilde{x}_i - \overline{\tilde{x}})} \in [0,1] \text{ for all } i.$$

It is not very difficult to come up with a numerical example. See Listing 1. ■

LISTING 1. Example for Remark C.1

```
import numpy as np

n = 100
x = (np.arange(n) / n) ** 3

lin_form = 1 + (x - 1 / 2) * 0.25

# Search for a valid pi* from this value
pi_init = lin_form / (1 + 2 * lin_form)

# Regression is y ~ 1 + w + x + wx
z0 = np.c_[np.ones(n), np.zeros(n), x, np.zeros(n)]
z1 = np.c_[np.ones(n), np.ones(n), x, x]

gram = ((1 - pi_init[:, None, None]) * z0[:, :, None] * z0[:, None, :]).mean(0) + (
    (pi_init[:, None, None]) * z1[:, :, None] * z1[:, None, :]
).mean(0)
lmbda = np.array([0, 1, 0, x.mean()]) # for the model-based ATE
rho0 = lmbda @ np.linalg.inv(gram) @ z0.T
```

```
rho1 = lmbda @ np.linalg.inv(gram) @ z1.T
pi = -rho0 / (rho1 - rho0)

# Iterate to find a fixed point
for _ in range(1000):
    new_gram = ((1 - pi[:, None, None]) * z0[:, :, None] * z0[:, None, :]).mean(0) + (
        (pi[:, None, None]) * z1[:, :, None] * z1[:, None, :]
    ).mean(0)
    new_rho0 = lmbda @ np.linalg.inv(new_gram) @ z0.T
    new_rho1 = lmbda @ np.linalg.inv(new_gram) @ z1.T
    new_pi = -new_rho0 / (new_rho1 - new_rho0)
    pi = new_pi

# Decompose to find the fractional linear parameters
A = ((x - x.mean()) ** 2).mean()
v = (pi * (x - x.mean()) ** 2).mean()
gamma1 = v / A
gamma0 = (pi * (x - x.mean())).mean()
a0 = pi.mean()
a1 = gamma0 / A
gamma2 = 1 / (v - v * gamma1 - gamma0 * gamma0) * (1 - a0 - v * a1 / gamma0) * gamma0
theta0 = a0 - gamma2 * gamma0
theta1 = a1 - gamma1 * gamma2

print(theta0, theta1, gamma2) # 0.3254446337090774 0.014146176384094272
    0.044263163071495276

# pi \in [0,1]
assert (pi > 0).all() and (pi < 1).all()

# the fractional linear representation is correct
assert (
    np.abs(pi - (theta0 + theta1 * (x - x.mean())) / (1 - gamma2 * (x - x.mean()))) < 1e-11
).all()

# Now, consider a new set of x's, will show that the parameters theta0, theta1, gamma2 no
# longer form a design justifying level irrelevance for the corresponding estimand
new_x = np.arange(n) / n
new_z0 = np.c_[np.ones(n), np.zeros(n), new_x, np.zeros(n)]
new_z1 = np.c_[np.ones(n), np.ones(n), new_x, new_x]
new_design = (theta0 + theta1 * (new_x - new_x.mean())) / (1 - gamma2 * (new_x -
    new_x.mean()))

# \tilde \pi \in [0,1]
assert (new_design > 0).all() and (new_design < 1).all()

gram_new_design = ((1 - new_design[:, None, None]) * z0[:, :, None] * z0[:, None,
    :]).mean(0) + (
    new_design[:, None, None] * z1[:, :, None] * z1[:, None, :]
).mean(0)
new_lmbda = np.array([0, 1, 0, new_x.mean()])
new_rho0 = new_lmbda @ np.linalg.inv(gram_new_design) @ new_z0.T
new_rho1 = new_lmbda @ np.linalg.inv(gram_new_design) @ new_z1.T

# Failure of level irrelevance
print(
    np.abs(new_design * new_rho1 + (1 - new_design) * new_rho0).max()
) # Equal to 0.08293240507391064, not zero
```

Proposition C.11. *Let $n \geq 3$. Consider binary W_i, covariates $x_i \in \mathbb{R}^d$, and scalar outcomes. Suppose the true design is linear $\pi_i^* = \delta_0 + \delta_1' x_i$ for some $\delta_0 \in \mathbb{R}, \delta_1 \in \mathbb{R}^d$. There is no regression $(\Lambda, z(x,w))$—where Λ may[45] depend on $x_{1:n}, \pi^*_{1:n}$—such that:*

(1) (Regression is linear in covariates) For all m and all w, the m^{th} entry of $z(x_i, w)$ is of the form $a_m(w) + b_m(w)' x_i$ for some fixed conformable $a_m(\cdot), b_m(\cdot)$.

(2) ($\Lambda\beta$ is the ATE) The corresponding estimand $\Lambda\beta \in \mathbb{R}$ is equal to the ATE, regardless of the configuration of $d, x_{1:n}, \delta_0, \delta_1$ (such that $\pi_i^* \in [0,1]$ for all i)

Proof. It suffices to show that for certain given choices of $x_{1:n}, \delta_0, \delta_1$, there is no regression that satisfies (1) and estimates the ATE.

Suppose $x_i \in [-1, 1]$ is scalar ($d = 1$) and suppose $\{x_1, \ldots, x_n\}$ includes $\{-1, 0, 1\}$. Set $\delta_0 = 1/2, \delta_1 = 1/4$ so that $\pi_i^* \in (0,1)$ for all i.

Fix some covariate transform $z(x_i, w)$. By assumption, its m^{th} entry is of the form $a_{mw} + b_{mw} x_i$. Suppose this regression estimates the ATE. In particular, it estimates a causal contrast. Thus

$$\pi_i^* \rho_i(1) + (1 - \pi_i^*)\rho_i(0) = 0. \tag{C.5}$$

Because the estimand is the ATE, we also have that $\omega_i = \rho_i(1)\pi_i^* = 1$. Hence, by (C.5), we have that for all i

$$1 + \left(1 - \frac{1}{\rho_i(1)}\right)\rho_i(0) = 0 \iff \rho_i(1)\rho_i(0) = \rho_i(1) + \rho_i(0)$$

Now, the potential weights are of the form

$$\rho_i(w) = \Lambda G_n^{-1} z(x_i, w) = c_{0w} + c_{1w} x_i$$

where the coefficients c_{0w}, c_{1w} may depend on $x_{1:n}, \pi^*_{1:n}$, but are the same for all i. As a result, for the regression to estimate the ATE, we must have

$$(c_{00} + c_{10}x)(c_{01} + c_{11}x) = (c_{00} + c_{10}x) + (c_{01} + c_{11}x)$$

for $x = -1, 0, 1$: That is,

$$c_{00}c_{01} = c_{00} + c_{01} \qquad (x = 0)$$

$$c_{00}c_{01} + c_{10}c_{11} + c_{10}c_{01} + c_{00}c_{11} = c_{00} + c_{01} + c_{10} + c_{11} \qquad (x = 1)$$

$$c_{00}c_{01} + c_{10}c_{11} - c_{10}c_{01} - c_{00}c_{11} = c_{00} + c_{01} - c_{10} - c_{11} \qquad (x = -1)$$

Now, adding the last two equations yield

$$c_{00}c_{01} + c_{10}c_{11} = c_{00} + c_{01} = c_{00}c_{01} \implies c_{10}c_{11} = 0 \implies \text{one of } c_{10}, c_{11} \text{ is zero}$$

Suppose $c_{10} = 0$. Then the bottom two equations simplify to

$$c_{00}c_{11} = c_{11}.$$

[45] This is to accommodate for estimands like the model-based ATT, where we may consider contrasts that depend on $\overline{x}_1 = \sum_i \pi_i^* x_i / \sum_i \pi_i^*$

This implies that $c_{00} = 1$ or $c_{11} = 0$. $c_{00} = 1$ contradicts the $x = 0$ equation. Thus, we must have $c_{11} = 0$. Suppose instead we assumed $c_{11} = 0$, by a symmetric argument we would derive $c_{10} = 0$. Thus, $c_{10} = c_{11} = 0$.

However, in that case, $\rho_i(w) = c_{0w}$ does not depend on x. For (C.5) to hold (for $x = -1, 0, 1$) then, we must have $c_{0w} = 0$, since π_i^* does depend on x. This means $\rho_1(w) = 0$ regardless of x, but this contradicts the fact that we estimate the ATE ($\rho_1(1)\pi_i^* = 1$). Thus, this regression does not estimate the ATE. □

C.5. **Additional results for panel specifications.**

Proposition C.12 (One-way fixed effects)**.** *Consider the regression $Y_{it} = \alpha_i + \tau W_{it} + \epsilon_{it}$. If the span of $\mathcal{W} \subset \{0,1\}^T$ excludes 1_T, then the set of implicit designs is empty.*

Proof. Let $\dot{\mathbf{w}} = \mathbf{w} - 1_T \frac{1_T' \mathbf{w}}{T}$. Then, by Theorem C.1, the potential weights are equal to

$$\rho_i(w) = \frac{\dot{\mathbf{w}}}{\frac{1}{n}\sum_{i=1}^n \sum_{\mathbf{w} \in \mathcal{W}} \pi_i^*(\mathbf{w}) \dot{\mathbf{w}}' \dot{\mathbf{w}}}.$$

Thus, if π_i corresponds to some implicit design $\boldsymbol{\pi}$,

$$\sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) \left(\mathbf{w} - 1_T \frac{1_T' \mathbf{w}}{T} \right) = 0.$$

In particular, this implies that 1_T is in the linear span of $\mathcal{W}$. By assumption, this is not the case. Therefore the set of implicit designs must be empty. □

Intuitively, this is because one-way fixed effects fail to account for the case where the treatment path correlates with underlying potential outcomes. If there are no treatment effects whatsoever, but the baseline potential outcome correlates with the treatment path (e.g., later potential outcomes tend to be larger in staggered adoption), then one-way fixed effects would in general estimate a nonzero coefficient, violating level irrelevance.

These results contrast with the results in Arkhangelsky and Imbens (2023), where the one-way fixed effect regression admits a design-based interpretation under within-unit random assignment of W_{it}. To reconcile, Arkhangelsky and Imbens (2023) consider a sampling-based setup and the unconfoundedness restriction

$$W_{it} \perp\!\!\!\perp (Y_{it}(0), Y_{it}(1)) \mid \frac{1}{T}\sum_{s=1}^T W_{is},$$

which is their (2.2) in our notation. Under staggered adoption, for instance, the associated propensity score $\mathrm{P}\left(W_{it} = 1 \mid \frac{1}{T}\sum_{s=1}^T W_{is}\right)$ is always degenerate, since $\frac{1}{T}\sum_{s=1}^T W_{is}$ perfectly distinguishes which treatment path unit i is assigned. Conversely, if W_{it} is randomly assigned within a unit and that permuting the time index results in valid counterfactual assignments, then $\mathcal{W}$ is large enough to contain 1_T.

Proposition C.13 (Event-study and more complex TWFE)**.** *Consider the regression* $Y_{it} = \alpha_i + \gamma_t + \beta' f_t(\mathbf{W}_i)$ *with target estimand* $\boldsymbol{\tau} = \Lambda\beta$ *under some design* $\boldsymbol{\pi}^*$*. We have:*

(i) Whether or not $\boldsymbol{\tau}$ *satisfies level irrelevance under the true design* $\boldsymbol{\pi}^*$*, one proper and Gram-consistent implicit design is*

$$\pi_i(\mathbf{w}) \equiv \pi(\mathbf{w}) \equiv \frac{1}{n}\sum_{j=1}^{n} \pi_j^*(\mathbf{w}). \tag{C.6}$$

Correspondingly, one estimated implicit design sets $\hat{\pi}_i(\mathbf{w})$ *to be the empirical frequency of treatment path* $\mathbf{w}$ *for all* i*.*

(ii) Suppose $\Lambda = I$*. If any column* k *of* $f_t(\mathbf{W}_i)$ *is such that the multiset*

$$\left(\begin{bmatrix} f_{1k}(\mathbf{w}) \\ \vdots \\ f_{Tk}(\mathbf{w}) \end{bmatrix} : \mathbf{w} \in \mathcal{W} \right)$$

(a) contains the zero vector at most once, (b) has nonzero vectors that are linearly independent, and (c) does not span the subspace $\{c1_T : c \in \mathbb{R}\}$*, then* $\overline{\pi}$ *is also the unique implicit design.*

Proof. (i) By Theorem C.1, it suffices to compute potential weights for the two-way residualized specification. That is, let

$$\mathbf{z}_i(\mathbf{w}) = \begin{bmatrix} f_1(\mathbf{w})' \\ \vdots \\ f_T(\mathbf{w})' \end{bmatrix} \equiv \mathbf{z}(\mathbf{w})$$

be the covariate transform. Note that the population residual of projecting $\mathbf{z}_i(\mathbf{W}_i)$ on unit and time fixed effect is

$$\begin{aligned}
\ddot{\mathbf{z}}(\mathbf{w}) &= \mathbf{z}(\mathbf{w}) - 1_T\frac{1_T'}{T}\mathbf{z}(\mathbf{w}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}'\in\mathcal{W}} \pi_i^*(\mathbf{w}')\mathbf{z}(\mathbf{w}') + 1_T\frac{1_T'}{T}\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}'\in\mathcal{W}} \pi_i^*(\mathbf{w}')\mathbf{z}(\mathbf{w}') \\
&= \mathbf{z}(\mathbf{w}) - 1_T\frac{1_T'}{T}\mathbf{z}(\mathbf{w}) - \sum_{\mathbf{w}'\in\mathcal{W}} \pi(\mathbf{w}')\mathbf{z}(\mathbf{w}') + 1_T\frac{1_T'}{T}\sum_{\mathbf{w}'\in\mathcal{W}} \pi(\mathbf{w}')\mathbf{z}(\mathbf{w}')
\end{aligned}$$

Thus the potential weights are

$$\boldsymbol{\rho}_i(\mathbf{w}) = \Lambda\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ddot{\mathbf{z}}(\mathbf{w})'\ddot{\mathbf{z}}(\mathbf{w})]\right)^{-1}\ddot{\mathbf{z}}(\mathbf{w})'$$

Note that

$$\sum_{\mathbf{w}\in\mathcal{W}}\pi(\mathbf{w})\boldsymbol{\rho}_i(\mathbf{w}) = \Lambda\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ddot{\mathbf{z}}(\mathbf{w})'\ddot{\mathbf{z}}(\mathbf{w})]\right)^{-1}\underbrace{\left(\sum_{\mathbf{w}\in\mathcal{W}}\pi(\mathbf{w})\ddot{\mathbf{z}}(\mathbf{w})'\right)}_{=0} = 0.$$

Therefore $\boldsymbol{\pi}$ defined by $\pi_i(\mathbf{w}) = \pi(\mathbf{w})$ is a valid implicit design. It is proper by definition. Now, let $\dot{\mathbf{z}}(\mathbf{w})$ be the within-transformed covariate transform for this regression (which includes the time

fixed effects), which does not depend on i. Note that the Gram matrix is

$$G_n(\boldsymbol{\pi}^*) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\dot{\mathbf{z}}(\mathbf{w})'\dot{\mathbf{z}}(\mathbf{w})] = \sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w})\dot{\mathbf{z}}(\mathbf{w})'\dot{\mathbf{z}}(\mathbf{w}) = G_n(\boldsymbol{\pi}).$$

Thus $\boldsymbol{\pi}$ is Gram-consistent.

The argument for the estimated implicit design is analogous. The sample residuals of $\mathbf{z}(\mathbf{W}_i)$ regressing on unit and time fixed effects is

$$\mathbf{z}(\mathbf{W}_i) - 1_T\frac{1_T'}{T}\mathbf{z}(\mathbf{W}_i) - \sum_{\mathbf{w}'\in\mathcal{W}} \hat{\pi}(\mathbf{w}')\mathbf{z}(\mathbf{w}') + \sum_{\mathbf{w}'\in\mathcal{W}} \hat{\pi}(\mathbf{w}')1_T\frac{1_T'}{T}\mathbf{z}(\mathbf{w}')$$

where $\hat{\pi}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\mathbf{W}_i = \mathbf{w})$. The estimated potential weights are proportional to these residuals evaluated at $\mathbf{W}_i = \mathbf{w}$. Thus $\hat{\boldsymbol{\pi}}$ defined by $\hat{\pi}_i(\cdot) = \hat{\pi}(\cdot)$ solves (A.2).

(ii) In this case, any implicit design $\tilde{\pi}_i$ needs to satisfy

$$0 = \sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}(\mathbf{w})\boldsymbol{\rho}_i(\mathbf{w}) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ddot{\mathbf{z}}(\mathbf{w})'\ddot{\mathbf{z}}(\mathbf{w})]\right)^{-1}\left(\sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}_i(\mathbf{w})\ddot{\mathbf{z}}(\mathbf{w})'\right)$$

Since $\Lambda = I$ and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ddot{\mathbf{z}}(\mathbf{w})'\ddot{\mathbf{z}}(\mathbf{w})]$ is positive definite by Theorem C.1, this implies that

$$\sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}_i(\mathbf{w})\ddot{\mathbf{z}}(\mathbf{w})' = 0$$

Inspecting this expression, we have

$$\sum_{\mathbf{w}\in\mathcal{W}} (\tilde{\pi}_i(\mathbf{w}) - \pi(\mathbf{w}))\mathbf{z}(\mathbf{w}) = 1_T \sum_{\mathbf{w}\in\mathcal{W}} (\tilde{\pi}_i(\mathbf{w}) - \pi(\mathbf{w}))\frac{1_T'\mathbf{z}(\mathbf{w})}{T}$$

Column k of this expression is in turn

$$\sum_{\mathbf{w}\in\mathcal{W}} (\tilde{\pi}_i(\mathbf{w}) - \pi(\mathbf{w}))\begin{bmatrix} f_{1k}(\mathbf{w}) \\ \vdots \\ f_{Tk}(\mathbf{w}) \end{bmatrix} = 1_T C_k$$

for some scalar C_k. The left hand side is in the span of $\begin{bmatrix} f_{1k}(\mathbf{w}) \\ \vdots \\ f_{Tk}(\mathbf{w}) \end{bmatrix}$, which excludes 1_T by assumption. Thus,

$$\sum_{\mathbf{w}\in\mathcal{W}} (\tilde{\pi}_i(\mathbf{w}) - \pi(\mathbf{w}))\begin{bmatrix} f_{1k}(\mathbf{w}) \\ \vdots \\ f_{Tk}(\mathbf{w}) \end{bmatrix} = 0.$$

Since at most only one $\mathbf{w}$ leads to $\begin{bmatrix} f_{1k}(\mathbf{w}) \\ \vdots \\ f_{Tk}(\mathbf{w}) \end{bmatrix} = 0$, and nonzero vectors $\begin{bmatrix} f_{1k}(\mathbf{w}) \\ \vdots \\ f_{Tk}(\mathbf{w}) \end{bmatrix}$ are linearly independent, we have that

$$\tilde{\pi}_i(\mathbf{w}) = \pi(\mathbf{w})$$

for all but one $\mathbf{w}$ by linear independence. Since both designs sum to 1, they also equal for remaining $\mathbf{w}$.

□

Proposition C.14 (Forbidden comparisons)**.** *Consider the TWFE specification $Y_{it} = \alpha_i + \gamma_t + \tau W_{it}$ under staggered adoption ($\mathbf{w}_t \leq \mathbf{w}_{t+1}$ for all $\mathbf{w} \in \mathcal{W}$ and all $t \in [T]$). Suppose the treatment time is randomly assigned so that $\pi_i^*(\mathbf{w}) = \pi^*(\mathbf{w})$. If there are two treatment paths with one being never-treated, $\mathbf{W} = \{0, \mathbf{w}\}$, then $\boldsymbol{\rho}_{it}(\mathbf{w}) \geq 0$ for all post treatment periods $\mathbf{w}_t = 1$ and $i \in [n]$. Otherwise, in all other configurations where $|\mathcal{W}| \geq 2$, there exists a choice of $\pi^*(\mathbf{w})$ such that the corresponding $\boldsymbol{\rho}_{it}(\mathbf{w}) < 0$ for some treatment path $\mathbf{w}$ with positive assignment probability ($\pi^*(\mathbf{w}) > 0$) and some post-treatment period t ($\mathbf{w}_t = 1$).*

Proof. The potential weight can be computed in closed form. For some $V > 0$,

$$\boldsymbol{\rho}_{it}(\mathbf{w}) = V^{-1} \left[\mathbf{w}_t - \frac{1'}{T}\mathbf{w} - \sum_{\tilde{\mathbf{w}} \in \mathcal{W}} \pi^*(\tilde{\mathbf{w}}) \tilde{\mathbf{w}}_t + \sum_{\tilde{\mathbf{w}} \in \mathcal{W}} \pi^*(\tilde{\mathbf{w}}) \frac{1'}{T} \tilde{\mathbf{w}} \right].$$

When $\mathcal{W} = \{0, \mathbf{w}\}$, then for a post-treatment t,

$$\boldsymbol{\rho}_{it}(\mathbf{w}) = V^{-1}(1 - \pi^*(\mathbf{w}))(1 - 1'\mathbf{w}/T) \geq 0.$$

Otherwise, let $\mathbf{w}$ be the treatment path with the earliest adoption date, and consider $t = T$. By assumption, $\mathcal{W}$ contains a path that adopts later than $\mathbf{w}$. Then

$$\boldsymbol{\rho}_{it}(\mathbf{w}) = V^{-1} \left((1 - \pi^*(\mathbf{w}))(1 - 1'\mathbf{w}/T) - \sum_{\tilde{\mathbf{w}} \neq \mathbf{w}} \pi^*(\tilde{\mathbf{w}}) \left(1 - 1'\tilde{\mathbf{w}}/T\right) \right)$$

Pick $\pi^*(\mathbf{w}) = 1/2 = \pi^*(\tilde{\mathbf{w}})$ where $\tilde{\mathbf{w}}$ adopts later than $\mathbf{w}$. Then

$$\boldsymbol{\rho}_{it}(\mathbf{w}) = V^{-1} \left(1/2(1 - 1'\mathbf{w}/T) - 1/2 \left(1 - 1'\tilde{\mathbf{w}}/T\right) \right) = \frac{1}{2} V^{-1} 1'(\tilde{\mathbf{w}} - \mathbf{w})/T < 0$$

since $\tilde{\mathbf{w}}$ adopts later than $\mathbf{w}$. □

Proposition C.15 (Time-varying covariates by interacting unit-specific covariates with time fixed effect)**.** *Suppose $\mathbf{W}$ satisfies staggered adoption and excludes an always-treated unit. Consider the TWFE specification*

$$\mathbf{Y}_{it} = \alpha_i + \gamma_t + \tau \mathbf{W}_{it} + x_i' \eta_t.$$

τ is a causal contrast only if there is some $\delta_0(\cdot), \delta_1(\cdot)$ such that

$$\pi_i^*(\mathbf{w}) = \delta_0(\mathbf{w}) + \delta_1(\mathbf{w})' x_i$$

for all $i \in [n], \mathbf{w} \in \mathcal{W}$.

Proof. Let $\mathbf{w}(t)$ be the path in $\mathcal{W}$ that is first treated in time t. The never-treated unit is denoted by $\mathbf{w}(\infty)$. By excluding the always-treated path, $\mathbf{w}(t) \in \mathcal{W}$ implies that $t \geq 2$.

By the same argument leading to (B.5), we have that if τ is a causal contrast, then for some $\rho_i, \lambda_t, \kappa_t$, for all $t \in [T]$ and $i \in [n]$,

$$\sum_{\mathbf{w} \in \mathcal{W}} \pi_i^*(\mathbf{w}) \mathbf{w}_t = \rho_i + \lambda_t + \kappa_t' x_i$$

For $s \geq 2$ a treatment time, consider taking the difference of the above equation between s and $s-1$. Because

$$\mathbf{w}_s - \mathbf{w}_{s-1} = \begin{cases} 1, & \text{if } \mathbf{w} = \mathbf{w}(s) \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\pi_i^*(\mathbf{w}(s)) = \lambda_s - \lambda_{s-1} + (\kappa_s - \kappa_{s-1})' \mathbf{x}_i \equiv \delta_0(\mathbf{w}(s)) + \delta_1(\mathbf{w}(s))' x_i.$$

This shows the desired representation for all $i \in [n]$ and all $0 \neq \mathbf{w} \in \mathcal{W}$. Lastly,

$$\pi_i^*(0) = 1 - \sum_{\mathbf{w} \in \mathcal{W} \setminus \{0\}} \pi_i(\mathbf{w})$$

can also be written as $\delta_0(0) + \delta_1(0)' x_i$. □