Potential weights and implicit causal designs in linear regression

Jiafeng Chen
Stanford University
jiafeng@stanford.edu

Abstract. When do linear regressions estimate causal effects in quasi-experiments? This paper provides a generic diagnostic that assesses whether a given linear regression specification on a given dataset admits a design-based interpretation. To do so, we define a notion of *potential weights*, which encode counterfactual decisions a given regression makes to unobserved potential outcomes. If the specification does admit such an interpretation, this diagnostic can find a vector of unit-level treatment assignment probabilities—which we call *an implicit design*—under which the regression estimates a causal effect. This diagnostic also finds the implicit causal effect estimand. Knowing the implicit design and estimand adds transparency, leads to further sanity checks, and opens the door to design-based statistical inference. When applied to regression specifications studied in the causal inference literature, our framework recovers and extends existing theoretical results. When applied to widely-used specifications not covered by existing causal inference literature, our framework generates new theoretical insights.

Date: July 30, 2024. I thank Isaiah Andrews, Dmitry Arkhangelsky, Ambarish Chattopadhyay, Matthew Gentzkow, Guido Imbens, Lihua Lei, Ashesh Rambachan, David Ritzwoller, Brad Ross, Jonathan Roth, Jesse Shapiro, Jann Spiess, Davide Viviano, Jose Zubizarreta, and participants at the Harvard graduate student workshop on econometrics for helpful comments.

1. Introduction

Much of economic research, particularly in applied microeconomics, has put a focus on research design and causal identification.[1] The key distinguishing feature of this school of thought—which we will refer to as "design-based"—is an experimentalist focus on reasoning with plausible randomization of treatment, so as to mimic some randomized control trial.[2] In observational studies, researchers often argue that the treatment is plausibly as-if random, so that causal effects are identified as they are in randomized experiments. An advantage of this mode of identification is validity without requiring functional form or homogeneity assumptions on potential outcomes.

Despite the popularity of the design-based perspective in terms of identification, the same is not true for estimation. Economic research continues to overwhelmingly prefer linear regression methods to other methods that are arguably more "natively" design-based (Imbens, 2015). These other methods—including inverse propensity-score weighting, matching (Zubizarreta, Paredes and Rosenbaum, 2014; Abadie and Imbens, 2011), doubly robust (Bang and Robins, 2005), or balancing approaches (Ben-Michael, Feller, Hirshberg and Zubizarreta, 2021)—explicitly use the probability of treatment assignment to target some explicit treatment effect. In contrast, linear regressions are historically motivated by treating the linear model literally as a structural model of potential outcomes (e.g., linear simultaneous equation models). When read in this way, regression methods usually impose some form of treatment effect homogeneity—and are thus more natively model-based. Indeed, the influential article by Freedman (2008) criticizes regressions in experimental settings precisely because 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—placing restrictions on treatment assignment but few to no restrictions on potential outcomes. In this view, linear

[1]By May 2024, over 50% of papers in the NBER working paper series mentions experimental or quasi-experimental methods (Goldsmith-Pinkham, 2024; Currie, Kleven and Zwiers, 2020).

[2]We refer to "design-based" as modeling treatment assignment, in contrast to modeling potential outcomes (which is often called "model-based"). This is consistent with the terminology adopted in Borusyak, Hull and Jaravel (2024a), Borusyak and Hull (2024), and Angrist and Pischke (2010). In particular, Angrist and Pischke (2010) observe, "In applied micro fields such as development, education, environmental economics, health, labor, and public finance, 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."

regression is essentially a machine that compares different groups of units. Under the right treatment randomization assumption, these comparisons recover causal effects. Indeed, much has been written on the properties of various linear specifications under known restrictions on treatment assignment (among others, Imbens and Angrist, 1994; Angrist, 1998; Lin, 2013; Słoczyński, 2022; Blandhol, Bonney, Mogstad and Torgovitsky, 2022; Aronow and Samii, 2016; Goldsmith-Pinkham, Hull and Kolesár, 2022; Borusyak and Hull, 2024; Athey and Imbens, 2022). This line of research analyzes particular regression specifications under a known set of designs and examines whether the specification makes sensible comparisons under the design. A limitation is that this literature results in a hodge-podge of results: Certain specifications are design-based 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 whether, given a dataset and a regression specification, there exists a design—that is, a distribution over treatment assignment—that rationalizes the specification, without restrictive outcome models. Suppose we are presented with a regression for which a design-based interpretation is ascribed: An empirical researcher hopes that this regression estimates a causal effect under some *design*—some process of random treatment assignment—without restrictions on potential outcomes. This paper provides generic tools and diagnostics that answer the following three questions:

(Q1) First, are the hopes of the researcher obviously misplaced and can we reject design-based interpretation altogether?

(Q2) Second, if a design-based interpretation is not obviously rejected, suppose we are willing to presume a design-based interpretation. Can we estimate or infer what the design is when it is not explicitly specified?

(Q3) Third, given such a design, what is the estimand that the regression specification targets?

By a causal effect, we mean the following: An estimand is a *causal contrast* if it satisfies *level irrelevance* (Blandhol *et al.*, 2022) under the true treatment assignment process. That is, shifting every potential outcome of some unit by the same number should not alter the estimand. This is a minimal requirement for a causal comparison when we do not impose restrictions on potential outcomes. For instance, if the treatment is binary, then a quantity satisfies level irrelevance if and only if it is a function of individual treatment effects.

We answer the questions (Q1)–(Q3) by connecting the algebra of linear regression estimators to their behavior under random treatment assignment. Every linear regression estimator is a weighted average of *observed* outcomes, where weights depend on *observed* treatment status and other covariates. Analogous to potential outcomes, we define a notion of *potential weights* for the linear specification. The potential weight for unit i, evaluated at treatment $\mathbf{w}$, is the weight put on i's *potential* outcome by the regression under the counterfactual treatment $\mathbf{w}$. The observed weights can then be viewed as the potential weights evaluated at the realized treatment assignment.

In other words, the potential weights encode what the linear estimator would have done to unit i's outcomes, had unit i received different treatment value. Thus, potential weights encode the precise comparisons regression estimators make under treatment randomization. These potential weights are readily computable in practice, and estimators for them are consistent under no stronger assumptions than typically assumed for the consistency of the OLS estimator. We show that, given the potential weights, the answers to all three questions are immediate.

If a regression estimand satisfies level irrelevance, then the potential weights for every unit must be mean zero under the true design. Otherwise, shifting all potential outcomes by one unit would shift the estimand by the expected value of the potential weights. Fixing the potential weights, this then yields linear equations that the true design must satisfy, assuming that the regression estimand is a causal contrast. We call the solution to these equations *implicit designs*—vectors that are orthogonal to the potential weights. In many cases, the implicit designs, if they exist at all, are often unique.

A key observation is that a regression estimand is a causal contrast if and only if the true design is an implicit design; therefore, computing the set of implicit designs gives information on the true design, under a presumption of innocence for the regression specification. This diagnostic then helps answer the three questions (Q1)–(Q3).

For (Q1), we can immediately reject a causal interpretation altogether if no implicit designs exist, or if the only solutions to the level irrelevance equations are not valid probability measures. If that happens, then we know that the regression specification could not have been a causal contrast under the true design, since otherwise the true design would have solved the equations. For (Q2), when the implicit design is unique, we immediately have a single plausible candidate for the true design. We can then inspect whether such a design accords with the researcher's economic intuition, and whether it is well-calibrated for the observed treatment statuses.[3] Finally, for (Q3), given a design, we can compute the estimand under that design. The estimand is a weighted average of potential outcomes, where the weight put on unit i's potential outcome under treatment w is simply the product of the corresponding treatment assignment probability and the potential weight. Thus, for such an *implicit estimand*, we can inspect whether it is sensible—e.g., whether it is a positively weighted average of individual treatment effects, whether it is contaminated by potential outcomes that should be irrelevant (Goldsmith-Pinkham *et al.*, 2022), or whether it is of policy interest.

Our diagnostic provides a useful and universal sanity check for both empirical researchers and their critics. If a plausible implicit design is found and the corresponding implicit estimand is sensible, then the regression specification is assured to estimate a reasonable quantity under some reasonable model of treatment assignment—thereby dispelling concerns of "negative weights" altogether. When that happens, the regression can even be called doubly robust: It retains a causal interpretation either when the treatment is assigned under the implicit design or when the regression correctly specifies the conditional means of the potential outcomes.

If we fail to find a plausible implicit design, researchers should then reexamine the interpretation of the regression specification. If the researcher remains confident of a outcome-model-free, design-based identification strategy, then they should perhaps be more explicit about the design—including using methods that are more natively design-based. In that case, implicit designs may be used to guide the model of treatment assignment. On the other hand, if the researcher insists on the regression specification, then a purely design-based interpretation is insufficient and the researcher should justify the corresponding functional form assumptions.[4] Moreover, the implicit design also opens the door to design-based inference in observational settings where the true design is unknown (Rambachan and Roth, 2020); design-based inference is useful to quantify uncertainty especially when a sampling model is difficult to conceptualize or justify.

When applied to some common regression specifications, our framework uncovers new theoretical insights. In the literature, many specifications have been shown to

[3]A set of probability forecasts are said to be well-calibrated if, among the events predicted to occur with probability p, a proportion of approximately p of them actually occur.

[4]For instance, in a panel context, one might justify that treatment is not necessarily randomly assigned but mean potential outcomes satisfy parallel trends.

have design-based interpretation under certain restrictions on the treatment assignment process. For many of these specifications, our framework can recover these results and supply a converse of sorts: We show that *the only* designs under which these specifications estimate causal contrasts are exactly those analyzed in the literature. Concretely, we supply these converses for Angrist (1998), Goldsmith-Pinkham *et al.* (2022), and Athey and Imbens (2022). We emphasize that we obtain all of these results—themselves about rather distinct settings[5]—by simply mechanically computing the potential weights, implicit design, and implicit estimands, with little tailored analysis.

Our framework also uncovers novel results for specifications that are less well-understood. First, it is well-known that Lin (2013)'s specification—where the outcome is regressed on an interaction of a binary treatment and demeaned covariates—recovers the average treatment effect if the covariates are saturated. We recover a version of this result, but additionally show that the saturation is essentially *necessary* in observational settings. Without saturating, even when the propensity score is linear in the covariates, this regression generally fails to retain a design-based interpretation—in contrast to the specification without interaction (Angrist, 1998). Second, we show that, in a panel setting with staggered adoption, two-way fixed effect (TWFE) regressions with time-varying covariates also generally do not have design-based interpretations, in the sense that the set of implicit designs is empty. Lastly, with staggered adoption and an *unbalanced* panel, we show that TWFE regressions also fail to have design-based interpretation unless the distribution of treatment timing is invariant across units with different missingness patterns.

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 by treating these weights as the observed counterpart to some potential weights, much like observed outcomes are potential outcomes evaluated at the observed treatment.

This paper proceeds as follows. Section 2 introduces our framework in population, without estimation error. Section 3 discusses estimation of potential weights and implicit design and establishes their consistency. Section 4 then applies our framework

[5] Angrist (1998) studies regressions that adjust for covariates in a binary-treatment cross-sectional setting. Goldsmith-Pinkham *et al.* (2022) study similar specifications with multiple treatments. Athey and Imbens (2022) study two-way fixed effects specifications in panel settings under a design-based framework.

to a litany of regression specifications, recovering known results and providing new ones. Section 5 concludes with further discussion and practical recommendations.

2. Potential weights and implicit designs

Consider a finite population of units $i \in [n] \equiv \{1, \ldots, n\}$. Each individual receives one of $J+1$ treatments $\mathbf{w} \in \mathcal{W}$, where $|\mathcal{W}| = J+1 < \infty$. 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, $\{\mathbf{y}_i(\mathbf{w}) \in \mathbb{R}^T : \mathbf{w} \in \mathcal{W}\}$.[6] Upon treatment realization, we observe a realized outcome $\mathbf{Y}_i = \mathbf{y}_i(\mathbf{W}_i)$.

We consider a design-based framework in which the potential outcomes and covariates are treated as fixed,[7] and the only randomness comes from the random assignment of $\mathbf{W}_i$. To that end, let Π^* denote the joint distribution of $(\mathbf{W}_1, \ldots, \mathbf{W}_n)$ and let $\boldsymbol{\pi}^*$ denote the associated marginal treatment assignment probabilities:[8]

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

We refer to such a vector $\boldsymbol{\pi}$ of marginal treatment probabilities as a *design*,[9] In particular, to $\boldsymbol{\pi}^*$ as the *true design*, and to a joint distribution Π on $\mathbf{W}_1, \ldots, \mathbf{W}_n$ as a *joint design*.

This setup is sufficiently general to encompass both cross-sectional ($T = 1$) and panel settings ($T > 1$) as well as both binary treatment ($J + 1 = 2$) and multivalued treatment ($J + 1 > 2$) settings. When we are in a cross-sectional setting ($T = 1$), we shall identify $\mathcal{W} = \{0, \ldots, J\}$ and, to simplify notation, we shall not bold the symbols Y_i, y_i, w, W_i, x_i. If the treatment is binary, then we additionally use π_i to denote $\pi_i(1)$, where $\pi_i(0) = 1 - \pi_i$. On the other hand, in a panel setting, it is often natural to identify $\mathcal{W} \subset \{0, 1\}^T$ as the set of *treatment paths* (Arkhangelsky

[6]For expositional clarity, we assume that the dimension of the outcome vector is the same across individuals (i.e., balanced panels). Our results in Section 4.2 do discuss imbalanced panels.

[7]Here, "design-based" is used in contrast to "sampling-based." In the sampling-based framework, uncertainty is assumed to have come from sampling from an infinite superpopulation, rather than solely from the random assignment of the treatment to a finite population (Abadie, Athey, Imbens and Wooldridge, 2020; Li and Ding, 2017; Rambachan and Roth, 2020). This is a distinct meaning from "design-based" justifications for an estimand; there, "design-based" is in contrast to "model-based", referring instead to the fact that causal identification rests on assumptions about treatment assignment rather than assumptions about functional forms of potential outcomes.

[8]So far, we do not assume the analogue of unconfoundedness in this finite population context (i.e., units with the same covariates have the same π_i^*). The probabilities π_i^* simply define treatment assignment probabilities conditional on potential outcomes and covariates, and as a result can accommodate rich forms of selection. However, as we shall see, our results often imply that many regression specifications are only justified under designs that satisfy unconfoundedness.

[9]Following broadly adopted terminology, we also often refer to π_i^* as a propensity score.

and Imbens, 2023a). Throughout this section, we illustrate with the standard cross-sectional, binary treatment setting as a running example.

Example 2.1 (Cross-sectional setting, binary treatment)**.** Here, units are associated with potential outcomes $y_i(1)$ and $y_i(0)$. We observe treatment assignments W_i, covariates x_i, and observed outcomes $Y_i = y_i(W_i)$. The covariates and potential outcomes are assumed to be fixed. The design $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_n^*) \in [0,1]^n$ describes the probabilities that $W_i = 1$ for every unit $i \in [n]$, i.e., $\pi_i^* = \mathrm{P}(W_i = 1)$. ■

Suppose we are presented with a regression specification by a researcher; our colleague hopes that this specification estimates a causal effect under the true design $\boldsymbol{\pi}^*$. In observational settings, the true design $\boldsymbol{\pi}^*$ is typically unknown, but the researcher usually supplies some economic argument that their regression specification estimates a causal parameter under the unknown design. This paper shows that merely presuming the regression specification estimates a causal effect is often quite informative of $\boldsymbol{\pi}^*$ and of the estimand of the regression under $\boldsymbol{\pi}^*$. In particular, it is sufficient to answer questions (Q1)–(Q3) posed in the introduction.

As an overview, our approach proceeds in two steps. First, we characterize the counterfactual behavior of the regression specification under the design $\boldsymbol{\pi}^*$. That is, we examine how the regression makes decisions with respect to each unit's *potential outcomes*. We define a notion of *potential weights* that encode these decisions, where the potential weight for unit i and treatment $\mathbf{w}$ is the weight that the regression *would assign* to the potential outcome $y_i(\mathbf{w})$, if the realized treatment $\mathbf{W}_i$ were equal to $\mathbf{w}$.

Second, we consider restrictions on the potential weights that make the regression estimand admit a causal interpretation. We observe that when we require the regression estimand to be insensitive to equal level shifts in each unit's potential outcomes, the expected potential weight under $\boldsymbol{\pi}^*$ must be zero for every unit. This leads to linear restrictions on $\boldsymbol{\pi}^*$ that often allow us to exactly pinpoint $\boldsymbol{\pi}^*$.

2.1. Population regression specification and potential weights.

To proceed, we first describe what we mean by a regression specification from a population perspective. Consider a regression of $\mathbf{Y}_{it}$ on a covariate transform $z_t(\mathbf{x}_i, \mathbf{W}_i) \in \mathbb{R}^K$.[10]

[10] In a panel setting, this setup accommodates unit-level fixed effects by defining $z_t(\mathbf{x}_i, \mathbf{W}_i)$ such that $\sum_{t=1}^T z_t(\mathbf{x}_i, \mathbf{W}_i) = 0$. In particular, for any given covariate transform $\tilde{z}_t$ that excludes a unit fixed effect, we can incorporate the unit fixed effect by defining $z_t = \tilde{z}_t - \frac{1}{T}\sum_{t=1}^T \tilde{z}_t$ as the within-transformed version of $\tilde{z}_t$. Importantly, our subsequent arguments do not accommodate unit fixed effects parametrized as a unit-level dummy variable in $\mathbf{x}_i$. See Remark 2.3 for an explanation.

Let β denote the K-dimensional coefficient vector associated with the covariate vector z_t—whose precise population definition is specified next. The empirical researcher hopes that a subvector of dimension k, $\boldsymbol{\tau} = \Lambda\beta$, has a causal interpretation, for a known matrix $\Lambda \in \mathbb{R}^{k\times K}$.

Example 2.2 (Example 2.1 continued)**.** As an example, consider the regression $Y_i = \alpha + \tau W_i + \epsilon$, where the coefficient of interest is τ. Then the covariate transform is $z(x_i, w) = [1, w]'$ and the coefficient contrast corresponds to $\Lambda = [0, 1]$. For a more general regression specification with covariate transform $z(x_i, w)$ with corresponding coefficients β, if we are interested in a scalar contrast $\tau = \lambda'\beta$, then $\Lambda = \lambda'$. ■

Such a regression has the following population (i.e., expectation over Π^*) first-order condition

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

where we define $G_n(\boldsymbol{\pi}^*)$ as the *population Gram matrix* of the regression, and $H_n(\boldsymbol{\pi}^*)$ as the population covariance between the outcome and the covariate transform.[11] The regression estimand β is then defined to be $\beta \equiv G_n(\boldsymbol{\pi}^*)^{-1}H_n(\boldsymbol{\pi}^*)$, and the coefficients of interest are in turn defined to be $\boldsymbol{\tau} \equiv \Lambda G_n(\boldsymbol{\pi}^*)^{-1}H_n(\boldsymbol{\pi}^*)$.

Formally, we make the following definition of a population regression specification. We take the perspective that when we are presented with a regression specification, we are:

(a) shown how to transform the covariates and treatment values to put on the right-hand side of the regression (i.e., $z_t(\mathbf{x}_i, \cdot)$),[12]
(b) told which coefficient contrasts to interpret causally (i.e., Λ), and
(c) shown the population Gram matrix $G_n = G_n(\boldsymbol{\pi}^*)$ of the regression.

The following definition formalizes this, which we illustrate with the running example.

Definition 2.1 (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

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

[11]Throughout, we assume that the population Gram matrix $G_n(\boldsymbol{\pi}^*)$ is invertible.

[12]It is important that counterfactual values of $z_t(\mathbf{x}_i, \mathbf{w})$ are known, which rules out specifications that include, say, lagged outcomes on the right-hand side. See Remark 2.4 for a discussion.

Example 2.3 (Example 2.1 continued)**.** Consider regressing Y_i on $z(x_i, W_i)$, where β is the coefficient vector corresponding to $z(\cdot)$. We are interested in the target parameter $\tau = \lambda'\beta$. This describes the following population regression specification:

$$\begin{aligned}
\Lambda &= \lambda' \\
G_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[z(x_i, W_i) z(x_i, W_i)'] \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi_i^* z(x_i, 1) z(x_i, 1)' + (1 - \pi_i^*) z(x_i, 0) z(x_i, 0)' \\
\mathbf{z}(x_i, w) &= z(x_i, w)', w \in \{0, 1\}\,.
\end{aligned}$$

■

It is worth remarking here that our perspective effectively treats the population Gram matrix G_n as known. This is a modeling device that resembles identification analysis (Manski, 2022), in which we often assume the population distribution of the observed data is known and abstract away from estimation. Despite G_n being a function of $\boldsymbol{\pi}^*$, this perspective is reasonable, as G_n, along with $H_n \equiv H_n(\boldsymbol{\pi}^*)$, admits a natural estimator

$$\hat{G}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} z_t(\mathbf{x}_i, \mathbf{W}_i) z_t(\mathbf{x}_i, \mathbf{W}_i)' \qquad \hat{H}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} z_t(\mathbf{x}_i, \mathbf{W}_i)\mathbf{Y}_i,$$

which we use to replace the unknown G_n in finite samples. It is quite reasonable to hope that $\hat{G}_n \approx G_n$. In fact, we ought to believe $\hat{G}_n, \hat{H}_n$ are close to G_n, H_n if we think the ordinary least-squares (OLS) estimator is a good estimator. Indeed, $\hat{G}_n, \hat{H}_n$ are directly used to substitute the unknown G_n and H_n when we estimate the regerssion via OLS. An empirical researcher who estimates the regression in question via OLS exactly justifies the estimator by hoping that $\hat{G}_n \approx G_n$ and $\hat{H}_n \approx H_n$. More formally, these hopes are usually validated, as under mild conditions, we can usually embed the finite population in a sequence of populations with growing sizes n, such that $\hat{G}_n - G_n \xrightarrow{p} 0$ and $\hat{H}_n - H_n \xrightarrow{p} 0$. Some examples of these conditions are presented in Section 3 and Appendix B.2.

There is also sometimes reason to think that G_n is directly observable and equal to $\hat{G}_n$; we discuss this in the following remark.

Remark 2.1 (Fixed Gram designs)**.** Note that for some regression specifications, there may be joint designs Π^* with the property that $\hat{G}_n = G_n$ almost surely under Π^*. We call such joint designs fixed Gram designs. If the treatments are assigned under a fixed Gram design, then the observed sample Gram matrix is equal to the population Gram matrix, making the latter known in a more literal sense. OLS estimators in design-based causal inference are frequently analyzed under fixed Gram designs, as they are unbiased for the estimand when the Gram matrix is fixed (see, e.g., Rambachan and Roth, 2020; Athey and Imbens, 2022; Zhao and Ding, 2022; Neyman, 1923/1990).

Fixed Gram designs can arise naturally. For instance, suppose we regress Y_i on a constant and the binary treatment indicator $W_i \in \{0,1\}$ as in Example 2.2. Then completely randomized experiments are fixed Gram designs.[13] In general, such designs often correspond to randomly choosing certain permutations of the treatment assignments, and are thus naturally related to randomized experiments and permutation inference.[14] ■

Given a population regression specification as defined by Definition 2.1, we can now unpack exactly how $\boldsymbol{\tau}$ compares different potential outcomes:

$$
\begin{aligned}
\boldsymbol{\tau} \equiv \Lambda G_n^{-1} H_n &= \frac{1}{n}\sum_{i=1}^n \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) \left(\sum_{t=1}^T \Lambda G_n^{-1} z_t(\mathbf{x}_i,\mathbf{w}) \mathbf{y}_{it}(\mathbf{w})\right) \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) \cdot \underbrace{\left[\Lambda G_n(\boldsymbol{\pi}^*)^{-1} \mathbf{z}(\mathbf{x}_i,\mathbf{w})'\right]}_{\boldsymbol{\rho}_i(\mathbf{w})\in\mathbb{R}^{k\times T}} \mathbf{y}_i(\mathbf{w}) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{W}\sim\pi_i^*}[\boldsymbol{\rho}_i(\mathbf{W})\mathbf{y}_i(\mathbf{W})]. \qquad (2.1)
\end{aligned}
$$

Equation (2.1) reveals that the vector of coefficients $\boldsymbol{\tau}$ is a comparison of potential outcomes $\mathbf{y}_i(\mathbf{w})$ under $\boldsymbol{\pi}^*$ in the sense that each potential outcome vector is weighted by the matrix $\boldsymbol{\rho}_i(\mathbf{w})$. These weighted potential outcomes are then aggregated under

[13] Under a completely randomized experiment with N_1 treated units all $\binom{n}{N_1}$ subsets of size N_1 are equally likely to be the treated units. When this is the case, the Gram matrix for this regression is equal to

$$\hat{G}_n = \begin{bmatrix} 1 & N_1/n \\ N_1/n & N_1/n \end{bmatrix} = G_n.$$

[14] More precisely speaking, given a realized vector of treatments $(\mathbf{W}_1,\ldots,\mathbf{W}_n)$, a fixed Gram design exists if there are a set of permutations $\sigma \in \mathcal{S}_n$ such that $(\mathbf{W}_1,\ldots,\mathbf{W}_n) \neq (\mathbf{W}_{\sigma(1)},\ldots,\mathbf{W}_{\sigma(n)})$ and that

$$\hat{G}_n = \hat{G}_n(\sigma) \equiv \frac{1}{n}\sum_{i=1}^n \sum_{t=1}^T z_t(\mathbf{x}_i,\mathbf{W}_{\sigma(i)}) z_t(\mathbf{x}_i,\mathbf{W}_{\sigma(i)})'.$$

for any $(\mathbf{W}_1,\ldots,\mathbf{W}_n)$. Any distribution over elements of $\mathcal{S}_n$ then induces a fixed Gram design by sampling the corresponding permutation of the treatment assignments.

the design $\boldsymbol{\pi}^*$ to produce $\boldsymbol{\tau}$. Motivated by this observation, we refer to $\boldsymbol{\rho}_i(\mathbf{w})$ as the *potential weights* associated with the regression specification.

Definition 2.2 (Potential weights)**.** Given a population regression specification, the *potential weights* for individual i are matrices defined by $\boldsymbol{\rho}_i(\mathbf{w}) \equiv \Lambda G_n^{-1}\mathbf{z}(\mathbf{x}_i, \mathbf{w})'$.

Since the potential weights are solely functions of the population regression specification, they are known under our perspective.

Before moving on to define causal restrictions on the potential weights, we summarize and illustrate our description of the regression specification and the potential weights in the running example Example 2.1.

Example 2.4 (Example 2.3 continued)**.** Each unit has two scalar potential weights $\rho_i(1)$ and $\rho_i(0)$, where

$$\rho_i(w) = \boldsymbol{\rho}_i(w) = \lambda' G_n^{-1} z(x_i, w) \in \mathbb{R}, \quad w \in \{0, 1\}\,.$$

The population regression estimand is decomposed according to (2.1) as

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

■

2.2. **Causal restrictions, implicit designs, and implicit estimands.** Not all regression estimands admit a causal interpretation. In other words, imposing a causal interpretation requires that the potential weights satisfy certain properties. We limit ourselves to considering causal *contrasts*, where treatment arms are compared against each other. A minimal requirement for a causal contrast is *level irrelevance*.[15] That is, if we added c_{it} to $\mathbf{y}_{it}(\mathbf{w})$ for all $\mathbf{w}$, then we should expect the causal contrast to be unchanged.

Definition 2.3 (Causal contrast)**.** We say that $\boldsymbol{\tau}$ is a vector of causal contrasts 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 $c_{it} \in \mathbb{R}$.

On the one hand, Definition 2.3 is a weak requirement for $\boldsymbol{\tau}$ in the sense that it does not rule out certain problematic behaviors. For instance, in a cross-sectional, binary-treatment setting, a weighted average treatment effect with negative weights[16] is a causal contrast by Definition 2.3. Whether stronger properties hold can be

[15] We follow the terminology in Blandhol *et al.* (2022).

[16] That is, $\tau = \frac{1}{n}\sum_{i=1}^{n} \omega_i(y_i(1) - y_i(0))$, where there exists (i, j) such that $\omega_i < 0 < \omega_j$

assessed under our framework as well, once we use level irrelevance to pinpoint $\boldsymbol{\pi}^*$. On the other hand, Definition 2.3 is a strong requirement because it allows any perturbation c_{it} to the potential outcomes. Thus, estimands whose validity hinges on a model for the potential outcomes usually fail to satisfy Definition 2.3. In this sense, Definition 2.3 encodes a requirement that causal identification purely comes from design-based assumptions.

This paper finds that assuming $\boldsymbol{\tau}$ are causal contrasts under $\boldsymbol{\pi}$ in the sense of Definition 2.3 is often sufficient to characterize $\boldsymbol{\pi}^*$. The reason is that $\boldsymbol{\rho}_i(\mathbf{w})$ specifies exactly how the regression estimands weight different potential outcomes $\mathbf{y}_i(\mathbf{w})$ under randomized treatment assignment. For $\boldsymbol{\tau}$ to satisfy Definition 2.3, these weights should be on average zero for every unit—otherwise, shifting $\mathbf{y}_{it}(\mathbf{w})$ by 1 would increase $\boldsymbol{\tau}$ by $\mathbb{E}[\boldsymbol{\rho}_{it}(\mathbf{W}_i)]$. Indeed, if $\boldsymbol{\tau}$ is vector of causal contrasts in the sense of Definition 2.3, then by examining (2.1), we immediately have that for all i, the average potential weight under $\boldsymbol{\pi}^*$ is zero:

$$\sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w})\boldsymbol{\rho}_i(\mathbf{w}) = 0, \quad \sum_{\mathbf{w}\in\mathcal{W}} \pi_i^*(\mathbf{w}) = 1. \tag{2.2}$$

Viewed as a set of restrictions on $\boldsymbol{\pi}^*$ with known potential weights, (2.2) defines a set of $kT+1$ linear equations in $J+1$ variables for π_i^*. We refer to the solutions of these restrictions as *implicit designs*.

Definition 2.4 (Implicit design)**.** Given the potential weights, we call a vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$ that satisfies (2.2) 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 $\boldsymbol{\pi}$ is proper and it generates the same Gram matrix G_n,

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

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

Implicit designs are solutions to (2.2), viewed as linear equations in $\boldsymbol{\pi}^*$. They admit an interpretation as a plausible marginal distribution of treatment if they are proper and Gram-consistent—meaning that they are probability measures on $\mathcal{W}$ and are consistent with the known population Gram matrix G_n.

We can often expect that implicit designs are unique, if they exist—at least modulo minor exceptions. This is because the number of equations is at least the number of unknowns ($kT \geq J$) in many common situations. For instance, in a binary treatment, cross-sectional setting, we have that $k = T = J = 1$. In a multivalued-treatment setting, we often have $k = J$ contrasts. In a panel setting with staggered adoption, we usually have fewer unique treatment times ($J+1$) than time horizon T. Lemma B.6 in the appendix gives conditions for the uniqueness of implicit design in cross-sectional settings ($T = 1$).

To be concrete, we explicitly compute the restriction in the binary treatment, cross-sectional setting.

Example 2.5 (Example 2.1, continued)**.** Specialized to the setting of Example 2.1, the restrictions on the potential weights—which we compute in Example 2.4—reduce to the single equation

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

This equation admits the sole solution $\pi_i = \frac{-\rho_i(0)}{\rho_i(1)-\rho_i(0)}$ if $\rho_i(1) \neq \rho_i(0)$. Thus, the set of improper implicit designs is empty if any unit has $\rho_i(1) = \rho_i(0) \neq 0$; otherwise, the set of improper implicit designs are real vectors $\boldsymbol{\pi}$ such that $\pi_i = \frac{-\rho_i(0)}{\rho_i(1)-\rho_i(0)}$ for all units with nonzero potential weights. Among these, a proper implicit design exists if $\rho_i(1)\rho_i(0) \leq 0$ for all i. ■

Relative to a proper and Gram-consistent implicit design $\boldsymbol{\pi}$, the regression specification estimates the $\boldsymbol{\pi}$-weighted average of the potential outcomes, which we term the corresponding *implicit estimand*, indexed by potential outcome weights $\boldsymbol{\omega}_i(\boldsymbol{\pi}, \mathbf{w}) = \pi_i \rho_i(\mathbf{w})$.

Definition 2.5 (Implicit estimand)**.** Let $\boldsymbol{\pi}$ denote a proper and Gram-consistent implicit design of a given regression specification. The corresponding *implicit estimand* is the $\boldsymbol{\pi}$-weighted average of the potential weights

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

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

2.3. **Implicit designs as diagnostics.** The following theorem summarizes our discussion of causal interpretations of the population regression specification.

Theorem 2.1. *$\boldsymbol{\tau}$ is a vector of causal contrasts in the sense of Definition 2.3 if and only if it is equal to some implicit design. When this happens, if, furthermore, there is a unique implicit design $\boldsymbol{\pi}$, then $\boldsymbol{\pi}$ is proper, Gram-consistent, and equal to $\boldsymbol{\pi}^*$.*

Having computed the potential weights, implicit designs, and implicit estimands, Theorem 2.1 allows us to answer the three questions (Q1)–(Q3) for the regression our colleague presents us with. If $\boldsymbol{\tau}$ is a vector of causal contrasts under $\boldsymbol{\pi}^*$, then the true design $\boldsymbol{\pi}^*$ satisfies (2.2). Thus, inspecting the solutions of (2.2)—or lack thereof—gives us information about $\boldsymbol{\pi}^*$.

To answer (Q1), we can reject any design-based interpretation of the regression if no proper and Gram-consistent implicit weights exist. When that happens, there is no treatment assignment process—consistent with the population Gram matrix—that rationalizes the regression estimand as a causal contrast. As a result, the regression estimand cannot be interpreted causally, at least when we are unwilling to restrict the behavior of potential outcomes. Researchers committing to such a regression specification thus must entertain and justify models of potential outcomes.

On the other hand, if a proper and Gram-consistent implicit design does exist and is unique, then we know that $\boldsymbol{\pi}^*$ must be equal to it—if we presume innocence for the regression specification. This answers (Q2). A candidate implicit design also leads to further diagnostics and insights. We can inspect, for instance, whether the implicit design conforms to our substantive knowledge of the treatment assignment process. Moreover, when the implicit design is unique, units with same covariates $\mathbf{x}_i$ must necessarily have the same implicit design π_i. This means that the regression specification must be justified by a finite population analogue of selection on observables.[17]

Lastly, we can compute the implicit estimand under $\boldsymbol{\pi}^*$ by computing the weights $\boldsymbol{\omega}(\boldsymbol{\pi}^*, \mathbf{w})$, thereby answering (Q3). Having computed the estimand, we can then inspect whether it is a reasonable quantity—we thus far only know that it satisfies level irrelevance. For instance, we might inspect whether it is a positively weighted average of individual treatment contrasts and whether it aligns with our intuition of the regression estimand. Section 4.1 contains examples where either requirement may fail even when an implicit design exists.

If these checks all pass—we find a proper and Gram-consistent implicit design under which the regression estimates a reasonable causal contrast—then we can rest assured that, under this treatment assignment process, our regression aggregates a sensible set of comparisons. This immediately dispels worries about "negative weights" when we are willing to defend the implicit design as reflecting reality. This also automatically imbues the regression specification with a double robustness property: The regression specification estimates causal quantities either if the treatment is assigned according

[17] See Remark 2.5 for a detailed discussion, and see Proposition 3.5 and Lemma B.6 for some primitive conditions under which the implicit design is unique.

to the implicit design, or if the regression correctly specifies the conditional means of the potential outcomes given treatment and covariates.[18]

In sample, however, these takeaways are complicated by the need to estimate G_n with $\hat{G}_n$, which we consider in detail in Section 3. That said, if the design Π^* is a fixed Gram design (Remark 2.1), then the potential weights are directly observable, which yields direct answers to (Q1)–(Q3) when combined with Theorem 2.1 without needing to discuss estimation. In particular, for (Q1), if we fail to find a proper and Gram-consistent implicit design, then the regression specification cannot be interpreted causally under any fixed Gram design.[19] For (Q2) and (Q3), if we do find such designs $\boldsymbol{\pi}$—and if we further find a joint design Π whose marginals are $\boldsymbol{\pi}$—then the regression has a design-based interpretation under Π with implicit estimands given by Definition 2.5.

When applied to the leading setting of binary treatments in a cross section (Example 2.1), Theorem 2.1 results in a sharp characterization of the design-based interpretations of regression estimands, as the implicit design is unique up to units with zero potential weights.

Corollary 2.2 (Binary treatment, cross-sectional setting). *In the cross-sectional, binary treatment setting of Example 2.1, τ is a causal contrast under the true design $\boldsymbol{\pi}^*$ if and only if $\rho_i(1)\rho_i(0) \le 0$ for all i and*

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

for all i with one of $\rho_i(1)$ and $\rho_i(0)$ nonzero.

When this happens, the implicit estimand is defined by $\omega_i(\boldsymbol{\pi}^*, 1) = \pi_i^* \rho_i(1)$ and $\omega_i(\boldsymbol{\pi}^*, 0) = -\omega_i(\boldsymbol{\pi}_i^*, 1)$, 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 $\rho_i(1) < 0 < \rho_i(0)$.

Namely, the regression estimand τ is a causal contrast if and only if the true design is equal to the implicit design, which is uniquely defined up to units with equal zero potential weights. Moreover, the implicit estimand is a weighted average treatment

[18]The meaning of these causal quantities may be different under design-based and model-based justifications, however.

[19]Even if we find such implicit designs $\boldsymbol{\pi}$, but if we cannot find a fixed Gram Π corresponding to $\boldsymbol{\pi}$, then we similarly cannot interpret the regression causally under any fixed Gram design.

effect. All unit-level weights ω_i in the implicit estimand are nonnegative if no unit has the "wrongly ordered" potential weights $\rho_i(1) < 0 < \rho_i(0)$.[20]

Before discussing estimation of G_n, potential weights, and implicit designs in Section 3, we conclude this section with a few remarks on invariances of potential weights to equivalent specifications, on restrictions placed on the covariate transform, and on interpretation of $\boldsymbol{\pi}$.

Remark 2.2 (Invariances of potential weights)**.** There are two ways in which the potential weights are invariant. First, the potential weights are invariant under reparametrizations of the covariate transform. That is, if we replace $z_t(\mathbf{x}_i, \mathbf{W}_i)$ with $\tilde{z}_t(\mathbf{x}_i, \mathbf{W}_i) \equiv M z_t(\mathbf{x}_i, \mathbf{W}_i)$ for an invertible $M \in \mathbb{R}^{K \times K}$ and correspondingly replace Λ with $\tilde{\Lambda} = \Lambda M'$—so that $\Lambda\beta$ and $\tilde{\Lambda}\tilde{\beta}$ are the same contrast—then the corresponding potential weights remain unchanged.

Second, the potential weights satisfy a Frisch–Waugh–Lovell property. That is, suppose we partition $z_t(\mathbf{x}, \mathbf{w})$ into $z_{t1}(\mathbf{x}, \mathbf{w})$ and $z_{t2}(\mathbf{x}, \mathbf{w})$ with dimensions K_1, K_2 respectively. Assume that Λ loads solely on coefficients of $z_{t1}(\mathbf{x}, \mathbf{w})$. Then the potential weights $\boldsymbol{\rho}_i(\mathbf{w})$ are the same as the potential weights from the regression of $\mathbf{Y}_{it}$ on $\tilde{z}_t(\mathbf{x}_i, \mathbf{W}_i) \equiv z_{t1}(\mathbf{x}_i, \mathbf{W}_i) - B_{1\to 2} z_{t2}(\mathbf{x}_i, \mathbf{W}_i)$. Here, $B_{1\to 2} \in \mathbb{R}^{K_1 \times K_2}$ is the matrix of population regression coefficients of z_{t1} on z_{t2}:

$$B_{1\to 2} = \left(\frac{1}{n} \sum_{i,t} \mathbb{E}[z_{t1}(\mathbf{x}_i, \mathbf{W}_i) z_{t2}(\mathbf{x}_i, \mathbf{W}_i)'] \right) \left(\frac{1}{n} \sum_{i,t} \mathbb{E}[z_{t2}(\mathbf{x}_i, \mathbf{W}_i) z_{t2}(\mathbf{x}_i, \mathbf{W}_i)'] \right)^{-1},$$

where both expectations are taken under $\mathbf{W}_i \sim \pi_i^*$. Both claims are stated and proven in Appendix B.1. ■

Remark 2.3 (Unit fixed effects)**.** Including unit fixed effects in a regression specification can be parametrized by ensuring that the covariate transform $z_t(\mathbf{x}_i, \mathbf{w})$ sum to zero: $\sum_{t=1}^{T} z_t(\mathbf{x}_i, \mathbf{w}) = 0$ for all i and $\mathbf{w} \in \mathcal{W}$. This holds by construction if we preprocess the covariates via the within transformation.

It may be tempting to parametrize unit fixed effects as unit-level dummy variable in the covariate transform, so that the dimension of z_t grows with the size of the population. We can define potential weights relative to this parametrization, but these potential weights are going to be different in general to the potential weights defined under a within-transformation. Moreover, the Gram matrix G_n associated

[20]If unit i has wrongly ordered potential weights, then the regression compares unit i's untreated potential outcomes against its treated potential outcome, rather than the other way around.

with this parametrization is in general no longer consistently estimable under reasonable conditions.[21] This is a manifestation of the incidental parameter problem. As a result, we maintain that unit-level fixed effects are accommodated by applying the within-transformation to the covariates z_t. ■

Remark 2.4 (Completion of $\mathbf{z}(\mathbf{x},\mathbf{w})$)**.** To compute the potential weights, it is crucial that $\mathbf{z}(\mathbf{x}_i, \mathbf{w})$ is known for all counterfactual values of the treatment assignment. This rules out specifications in which post-treatment variables are included as covariates (e.g., lagged outcomes in panel settings). Similarly, this rules out specifications that include endogenous treatment variables on the right-hand side in an instrumental variables (IV) or mediation setting.[22] In such cases, the counterfactual weight put on counterfactual outcomes depend on other counterfactual outcomes, and thus we cannot "complete" the regression representation in (2.1). ■

Remark 2.5 (Unconfoundedness)**.** When the implicit design is unique and $\boldsymbol{\tau}$ is a causal contrast under $\boldsymbol{\pi}^*$, Theorem 2.1 shows that the true design $\boldsymbol{\pi}^*$ is equal to the implicit design. The implicit design, however, is a function solely of the potential weights $\rho_i(\mathbf{w})$, which only depends on the covariates $\mathbf{x}_i$ in the sense that two individuals with the same covariate values $\mathbf{x}_i = \mathbf{x}_j$ must also then have the same potential weights $\rho_i(\mathbf{w}) = \rho_j(\mathbf{w})$ for all $\mathbf{w}$. If the implicit design is unique, then these two individuals must also have $\pi_i = \pi_i^* = \pi_j^* = \pi_j$. This property of $\boldsymbol{\pi}^*$ is a finite-population version of the unconfoundedness assumption, where the treatment is randomly assigned "conditional on $\mathbf{x}_i$."

Therefore, if a design-based interpretation is possible for parameters $\boldsymbol{\tau}$ under a $\mathbf{Y}_i$-on-$\mathbf{z}(\mathbf{x}_i, \mathbf{W}_i)$ regression, then this interpretation must be under a design that satisfies this unconfoundedness property—where selection propensities into treatment only depends on the covariates. ■

[21] For unit i, let $\tilde{z}_t(\mathbf{x}_i, \mathbf{W}_i)$ denote the covariate transforms that exclude the 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.

[22] For IV estimators that are ratios of two linear treatment effect estimators, we can analyze the numerator (i.e., reduced form, or the treatment effect of the instrument on the outcome) and the denominator (i.e., first stage, or the treatment effect of the instrument on the endogenous treatment) separately under our framework. This accommodates, for instance, two-stage least squares.

3. **Estimating potential weights**

This section studies the properties of the estimated potential weights

$$\hat{\boldsymbol{\rho}}_i(\mathbf{w}) = \Lambda \hat{G}_n^{-1}\mathbf{z}(\mathbf{x}_i, \mathbf{w})' \text{ where } \hat{G}_n = \frac{1}{n}\sum_{i=1}^n \sum_{t=1}^T z_t(\mathbf{x}_i, \mathbf{W}_i) z_t(\mathbf{x}_i, \mathbf{W}_i)'. \tag{3.1}$$

By considering finite populations of increasing size, it is often possible to show that $\hat{G}_n - G_n \xrightarrow{p} 0$—which is often used to show consistency for the OLS estimator. This, coupled with an assumption that G_n is suitably invertible, allows us to conclude that the estimated potential weights are consistent for their population counterparts. Consistency of the inverse Gram matrix also implies that the OLS estimator is close to an oracle Horvitz–Thompson estimator, at an order of $O(\|\hat{G}_n - G_n\|)$. 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. Quantitative bounds on the deviation $\|\hat{G}_n - G_n\|$ also allows us to derive tests for the existence of proper implicit designs. Practically, these results assure us that the estimated potential weights and implicit designs—by simply plugging in the sample Gram matrix—are close to their population counterparts in large samples, and further provides formal inference procedures.

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 $\pi^*_{1,n}, \ldots, \pi^*_{n,n}$ 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 $\pi^*_{1,n}, \ldots, \pi^*_{n,n}$.

3.1. **Consistency of potential weights.**

We begin by establishing the consistency of $\hat{\boldsymbol{\rho}}_i$ for $\boldsymbol{\rho}_i$. The results in this subsection are based on the following assumption—that the sample Gram matrix is close to the population Gram matrix—regardless of whether the population regression specification has a causal interpretation.

Assumption 3.1. $\hat{G}_n - G_n \xrightarrow{p} 0$.

Assumption 3.1 holds whenever the underlying populations are such that a law of large numbers hold. If the treatments are independently assigned, for instance, then Assumption 3.1 is true under standard laws of large number (e.g., Theorem 2.2.6 in Durrett (2019)), which we state in the following lemma.

Lemma 3.1. *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 $n \to \infty$, $\frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\pi_{i,n}^*}[G_{i,j\ell}^2]\right) \to 0$. Then Assumption 3.1 holds.*

$\hat{G}_n$ is consistent for G_n under more complex designs as well. Appendix B.2 verifies Assumption 3.1 for designs that involve sampling without replacement (Hájek, 1964; Rambachan and Roth, 2020).

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.[23]

Assumption 3.2. *The sequence of population Gram matrices G_n is such that their minimum eigenvalues are bounded below: For some $\epsilon > 0$, for all n, $\lambda_{\min}(G_n) \ge \epsilon > 0$.*

Assumption 3.2 makes sure that the population regression specification is strongly identified, in the sense that the Gram matrix is bounded away from singularity. Under Assumption 3.1 and Assumption 3.2, we can show that the estimated potential weights (3.1) are consistent for their population counterparts.

Proposition 3.2. *Under Assumptions 3.1 and 3.2, the estimated potential weights are consistent: $\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w}) \xrightarrow{p} 0$. 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:*

$$\max_{i \in [n], \mathbf{w} \in \mathcal{W}} |\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty \xrightarrow{p} 0. \tag{3.2}$$

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

3.2. **OLS as approximately Horvitz–Thompson.** Under Assumptions 3.1 and 3.2, we can also make a connection between the OLS estimator and an oracle Horvitz–Thompson estimator. Note that the OLS estimator for $\boldsymbol{\tau}$ is an average of the observed outcome $\mathbf{y}_i$ weighted by the observed estimated potential weights $\hat{\boldsymbol{\rho}}(\mathbf{W}_i)$:

$$\hat{\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\rho}}_i(\mathbf{W}_i)\mathbf{y}_i = \frac{1}{n}\sum_{i=1}^n \sum_{\mathbf{w} \in \mathcal{W}} \mathbb{1}(\mathbf{W}_i = \mathbf{w})\hat{\boldsymbol{\rho}}_i(\mathbf{w})\mathbf{y}_i(\mathbf{w}). \tag{3.3}$$

The corresponding population estimand is $\boldsymbol{\tau} = \frac{1}{n}\sum_{i=1}^n \sum_{\mathbf{w} \in \mathcal{W}} \pi_i^*(\mathbf{w})\boldsymbol{\rho}_i(\mathbf{W}_i)\mathbf{y}_i$ irrespective of whether it is a causal contrast. With knowledge of $\boldsymbol{\rho}_i(\mathbf{W}_i)$ and $\pi_i^*(\mathbf{w})$, a

[23]Under Assumption 3.1 and Assumption 3.2, 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.

natural Horvitz–Thompson estimator targeting $\boldsymbol{\tau}$ is

$$\hat{\boldsymbol{\tau}}_{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}(\mathbf{W}_i = \mathbf{w})}{\pi_i^*(\mathbf{w}_i)} \pi_i^*(\mathbf{w}) \boldsymbol{\rho}_i(\mathbf{W}_i)\mathbf{y}_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}\in\mathcal{W}} \mathbb{1}(\mathbf{W}_i = \mathbf{w})\boldsymbol{\rho}_i(\mathbf{w})\mathbf{y}_i(\mathbf{w}). \tag{3.4}$$

Thus, the OLS estimator (3.3) can be viewed as the analogue of the Horvitz–Thompson estimator (3.4) with the estimated potential weights replacing their population counterparts. The following proposition makes that precise under mild boundedness conditions on the sequence of populations.

Proposition 3.3. *Assume that the sequence of populations is such that the average covariance between the covariates and potential outcomes is bounded*

$$\frac{1}{n}\sum_{i=1}^{n} \max_{\mathbf{w}\in\mathcal{W}} \|\mathbf{z}(\mathbf{x}_i, \mathbf{w})'\mathbf{y}_i(\mathbf{w})\|_\infty = O(1).$$

Under Assumptions 3.1 and 3.2, let constants $b_n > 0$ be such that $|\hat{G}_n - G_n|_\infty = O_P(b_n)$. Then $\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}_{\mathrm{HT}}\|_\infty = O_P(b_n)$.

Thus, in this sense, the OLS estimator is close to a natural design-based estimator for the estimand $\boldsymbol{\tau}$.[24] When $\boldsymbol{\tau}$ has a causal interpretation in the sense of Definition 2.3, the OLS estimator can be interpreted as an estimator of a causal quantity, asymptotically justified under design-based assumptions.

With a proper estimated implicit design $\hat{\boldsymbol{\pi}}$, we can also write the OLS estimator in a form suggestive of inverse propensity weighting:

$$\hat{\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{w}\in\mathcal{W}} \frac{\mathbb{1}(\mathbf{W}_i = \mathbf{w})}{\hat{\pi}_i(\mathbf{w})} \hat{\pi}_i(\mathbf{w})\hat{\boldsymbol{\rho}}_i(\mathbf{w})\mathbf{y}_i(\mathbf{w}). \tag{3.5}$$

When $\hat{\boldsymbol{\pi}}$ is consistent for $\boldsymbol{\pi}^*$, this representation is directly informative of the design-based interpretation of $\hat{\boldsymbol{\tau}}$. In that case, the OLS estimator can be interpreted as a Horvitz–Thompson estimator for the estimand $\frac{1}{n}\sum_i\sum_{\mathbf{w}} \hat{\pi}_i(\mathbf{w})\hat{\boldsymbol{\rho}}_i(\mathbf{w})y_i(\mathbf{w})$, with propensity score $\hat{\pi}_i(\cdot)$. Crucially, however, we emphasize that the OLS estimator is close to the Horvitz–Thompson estimator regardless of whether $\hat{\boldsymbol{\pi}}$ is close to $\boldsymbol{\pi}^*$, as long as the estimated potential weights are consistent for their population counterparts.

3.3. **Consistency of implicit designs.** Naturally, we define the set of estimated (improper) implicit designs as the set of solutions to the sample counterpart of (2.2):

[24] While $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\tau}}_{\mathrm{HT}}$ are close, since $|\hat{G}_n - G_n|_\infty$ is usually on the order of $1/\sqrt{n}$, the two estimators are not generally first-order asymptotically equivalent.

$\hat{\boldsymbol{\pi}}$ is an estimated implicit design if for every i,

$$\sum_{\mathbf{w}\in\mathcal{W}} \hat{\pi}_i(\mathbf{w})\hat{\boldsymbol{\rho}}_i(\mathbf{w}) = 0, \quad \sum_{\mathbf{w}\in\mathcal{W}} \hat{\pi}_i(\mathbf{w}) = 1. \tag{3.6}$$

The next proposition, Proposition 3.4, verifies that the estimated implicit designs are consistent in the sense that the population implied weights have zero expectation in the limit under members of $\hat{\boldsymbol{\pi}}$. In that sense, the estimated implicit designs are limiting solutions to the population level irrelevance restriction (2.2).

Proposition 3.4. *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 probability vector 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 (3.2) holds, then the convergence is uniform:

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

Coupled with an assumption about the uniqueness of population implicit design (i.e., uniqueness of solutions to (2.2)), the estimated implicit designs are then consistent for $\boldsymbol{\pi}^*$ in the usual sense. We illustrate with the following result in the cross-sectional case ($T = 1$) where the system of equations (2.2) 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 3.5. *Assume that $k = J$. Suppose that:*

(i) The estimand $\boldsymbol{\tau}$ is a vector of causal contrasts in the sense of Definition 2.3 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 (3.2).

Then, for some $C > 0$, a unique estimated implicit design $\hat{\boldsymbol{\pi}}$ that is bounded by 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* $\boldsymbol{\pi}^*$ *in the sense that*

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

The key condition in Proposition 3.5 is (ii), which ensures that the solution to the population level irrelevance restriction (2.2) 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 B.6 further provides sufficient conditions for (ii) to hold in the cross-sectional case.

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. For instance, it is possible for noise in the coefficients of a underdetermined system to turn the system into an overdetermined one. Thus, it is possible that a population implicit design exists—meaning that (2.2) is under or exactly determined—but no estimated implicit designs do. For instance, the theoretical applications in Section 4 (Propositions 4.4 and 4.5) may exhibit this phenomenon. Nevertheless, the estimated implicit designs—or their existence thereof—are directly informative about any interpretations under fixed-Gram designs.

Finally, it is in principle possible to perform statistical inference for implicit designs. For a confidence set of implicit designs, one simple method is to project a confidence set for G_n, perhaps itself derived from large-deviation inequalities for the sample Gram matrix. For a given $(1-\alpha)$-confidence set $\hat{\mathcal{G}}_n$ for G_n,[25] we can project it to form a confidence set of proper and Gram-consistent implicit designs:

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

$$\text{for all } i,\ \pi_i(\cdot) \geq 0,\ \sum_{\mathbf{w}\in\mathcal{W}} \pi_i(\mathbf{w}) = 1, \qquad \text{(Proper)}$$

[25]For instance, such a confidence set can be obtained from the simultaneous confidence set for all entries of G_n—which are themselves population means of sample averages and obey central limit theorems under typically mild conditions.

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

Unfortunately, computing this set is a nonconvex, nonlinear problem,[26] which we conjecture to be computationally challenging. Nonetheless, checking whether a candidate $\boldsymbol{\pi}$ belongs to this confidence set 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.[27]

4. **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. To emphasize, for both purposes, our results essentially reduce the problem to *computing* the (estimated) potential weights and the (estimated) set of implicit designs.

This section computes the potential weights and implicit designs for a number of population regression specifications. 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.*, 2022; Athey and Imbens, 2022). We show that the corresponding implicit designs essentially rediscover these properties and—moreover—supply a converse of sorts. Specifically, the implicit design is often unique given the population regression specification, meaning that there is a unique design that is (a) consistent with a given Gram matrix G_n and (b) grants the regression a causal interpretation in the sense of Definition 2.3. In this sense, then, the design assumptions justifying these specifications in the literature are the *only* such assumptions.

When applied to a number of examples that are less known, our results uncovers some new insights. We show that the following regression specifications generally do not have causal interpretations, in the sense that their sets of implicit designs are either empty or do not include $\boldsymbol{\pi}^*$, except under certain knife-edge cases:

[26] To be precise, checking whether this set is nonempty is a nonconvex and nonlinear feasibility problem.

[27] In particular, given a proper implicit design, we can compute the corresponding Gram matrix under that design and check whether it belongs to $\hat{\mathcal{G}}_n$.

- In a cross-section, binary-treatment setting, consider τ in the specification $Y_i = \alpha + \tau W_i + \gamma'(x_i - \overline{x}) + \delta' W_i (x_i - \overline{x}) + \epsilon_i$ (Lin, 2013) where $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. The covariates x_i do not represent saturated categorical covariates.
- In a balanced panel setting with staggered adoption, consider τ in the TWFE specification $Y_{it} = \alpha_i + \beta_t + \tau W_{it} + \gamma' x_{it} + \epsilon_{it}$, where inclusion of the covariates changes the coefficient on W_{it}.
- In an unbalanced panel setting with staggered adoption, consider τ in the TWFE specification $Y_{it} = \alpha_i + \beta_t + \tau W_{it} + \epsilon_{it}$, where the distributions of treatment times are different across units with different observable time periods.

While we present a litany of results here in starkly distinct settings, we emphasize that in each example we uncover these results by mechanically computing the potential weights, implicit designs, and estimands, with little tailored analysis. Thus our results provide a unified understanding of design-based interpretations of regression specifications. For ease of exposition, we divide these results into cross-sectional settings and panel settings. Throughout this section, we assume that the population Gram matrix is invertible, and our results about estimated potential weights and estimated implicit designs are limited to the event that the sample Gram matrix is invertible.

4.1. Cross-sectional applications.

4.1.1. *Binary treatments, Angrist (1998).*

We begin with the leading case of cross-sections with binary treatments. We first study the regression specification in Angrist (1998), where one "controls for" the covariates x_i in a outcome-on-treatment regression:

$$Y_i = \tau W_i + \gamma' x_i + \epsilon_i. \tag{4.1}$$

Loosely speaking, Angrist (1998) finds that (4.1) estimates the $\pi_i^*(1-\pi_i^*)$-weighted average treatment effect when the propensity score π_i^* is linear in x_i. The following proposition computes the set of implicit weights and shows a converse: Given any true design $\boldsymbol{\pi}^*$, the set of implicit designs contains only $\boldsymbol{\pi}^*$'s linear projection onto the covariates. Thus, for $\boldsymbol{\pi}^*$ to justify a causal interpretation of τ, it must then be linear in the covariates to begin with. Thus, a researcher who runs the regression (4.1) hoping for a design-based, outcome-model-free interpretation of τ must then assume that propensity scores are linear in x_i.

Proposition 4.1. *Let the population regression specification be described by (4.1) under the design $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_n^*)$, where the coefficient of interest is τ. Let $\pi_i =$*

$x_i'\beta_{w\to x}$ be the linear projection of $\boldsymbol{\pi}^*$ onto x_i:

$$\beta_{w\to x} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n x_i \pi_i^*.$$

Then the improper implicit design is unique and equal to $\boldsymbol{\pi}$. If τ is a causal contrast under $\boldsymbol{\pi}^*$, then (i) $\boldsymbol{\pi}^* = \boldsymbol{\pi}$, (ii) $\boldsymbol{\pi}^*$ is linear in x_i, and (iii) $\boldsymbol{\pi}$ is proper. When that happens, the implicit estimand is the $\pi_i(1-\pi_i)$-weighted average treatment effect

$$\tau = \frac{1}{n}\sum_{i=1}^n \underbrace{\frac{\pi_i(1-\pi_i)}{\frac{1}{n}\sum_{j=1}^n \pi_j(1-\pi_j)}}_{\omega_i}(y_i(1) - y_i(0)),$$

and ω_i is nonnegative if and only if $\pi_i \in [0,1]$.

Proposition B.7 presents the sample analogue of Proposition 4.1. We show that the estimated implicit design is uniquely $\hat{\pi}_i = x_i'\hat{\beta}_{w\to x}$, where $\hat{\beta}_{w\to x}$ is the regression coefficient of the observed treatment indicators $W_1, \ldots, W_n$ on the covariates. The estimated estimand is then the $\hat{\pi}_i(1-\hat{\pi}_i)$-weighted average treatment effect.

Proposition 4.1 also immediately implies a few corollaries. First, if the covariates only include a constant—in which case τ is a simple mean comparison between treated and non-treated units—then the implicit design is necessarily constant ($\pi_i = \pi$). We thus know that for τ to have a design-based causal interpretation, the true design must be a randomized experiment. Second, if the covariates represent saturated versions of an underlying categorical variable c_i,[28] then the implicit design is simply a stratified randomized experiment on levels of c_i, which is then required if τ were to be causal under $\boldsymbol{\pi}^*$. Moreover, when the design is a stratified randomized experiment, a fixed Gram design exists by running a completely randomized experiment within each stratum—when that is true, the estimated implicit design (described by the within-stratum realized proportions of treatment) is exactly the true design. Finally, if the linear probability model is obviously misspecified in the sense that $\pi_i \notin [0,1]$ for some i, then we must reject the causal interpretation of τ under $\boldsymbol{\pi}^*$.

4.1.2. *Binary treatment, interacted specification.* For a researcher interested in the average treatment effect, the regression specification (4.1) can be unsatisfying, as

[28] That is, $x_i = e_m$ if $c_i = m$ for c taking values in $1, \ldots, M$ and e_m the m^{th} standard basis vector.

it upweights units with propensity scores closer to 1/2.[29] It is well-known that when the covariates are saturated versions of some categorical variable, the following regression—which interacts the treatment indicator with the covariates—recovers the average treatment effect with τ under a stratified randomized experiment (Lin, 2013; Imbens and Wooldridge, 2009):

$$Y_i = \alpha + \tau W_i + \gamma'(x_i - \overline{x}) + \delta' W_i (x_i - \overline{x}) + \epsilon_i \quad \overline{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i \tag{4.2}$$

Analogous to Proposition 4.1, Proposition B.8 likewise shows that, when x_i represents saturated categorical covariates, stratified randomized experiments are the only designs justifying a causal interpretation of τ; when this happens, the corresponding implicit estimand is the average treatment effect.

Given Proposition 4.1, one might conjecture that the same is true when x_i is not saturated—for instance when x_i is continuous—as long as the true propensity scores π_i^* are linear in the covariates. Surprisingly, we show that (4.2) does not in general have a causal interpretation justified by design, even when π_i^* is linear in the covariates.

To describe this result, we define a few population projection coefficients. Let the population regression specification be described by (4.2) under π_i^*. Let $\tilde{\pi}_i$ be the linear projection of π_i^* onto a constant and x_i. Moreover, let Γ_0, Γ_1 be the population projection coefficient of $\pi_i^*(x_i - \overline{x})$ on a constant and $x_i - \overline{x}$, respectively. Let Γ_2 be the population projection coefficient of $W_i - \tilde{\pi}_i$ on the residual $W_i(x_i - \overline{x}) - \Gamma_0 - \Gamma_1(x_i - \overline{x})$.[30] In other words, the population linear projection of W_i on $[1, x_i - \overline{x}, W_i(x_i - \overline{x})]$ can be written as

$$\begin{aligned} W_i &= \tilde{\pi}_i + \Gamma_2' \left(W_i(x_i - \overline{x}) - \Gamma_0 - \Gamma_1(x - \overline{x}) \right) + u_i \\ &= -\Gamma_2'\Gamma_0 + \Gamma_2' W_i(x_i - \overline{x}) + (\tilde{\pi}_i - \Gamma_2'\Gamma_1(x - \overline{x})) + u_i, \end{aligned}$$

[29] The interpretation of the estimand thus depends on the experimental design, which may be unrelated to policy objectives. On the other hand, Goldsmith-Pinkham *et al.* (2022) do give this estimand an interpretation as the easiest-to-estimate weighted average treatment effect—the treatment effect with the most favorable semiparametric efficiency bound.

[30] That is,

$$\Gamma_2 = \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W_i}\left[W_i(x_i - \overline{x}) \left(W_i(x_i - \overline{x}) - \Gamma_0 - \Gamma_1(x_i - \overline{x}) \right)' \right] \right)^{-1} \cdot \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W_i}\left[(W_i - \tilde{\pi}_i) \cdot \left(W_i(x_i - \overline{x}) - \Gamma_0 - \Gamma_1(x_i - \overline{x}) \right) \right] \right).$$

$$\mathbb{E}[u_i\,[1,(x_i-\overline{x})',W_i(x_i-\overline{x})']'] = 0.$$

Finally, we say that vectors $v_1,\ldots,v_n$ are *generically non-affine* in $x_1,\ldots,x_n$ if the affine projection $v_i = A_0 + A_1 x_i + r_i$ has residuals that are full rank: $\frac{1}{n}\sum_{i=1}^n r_i r_i' \succ 0$.

Proposition 4.2. *Under the setting described above for the specification (4.2), the implicit design is equal to*

$$\pi_i = \frac{\tilde{\pi}_i - \Gamma_2'(\Gamma_0 + \Gamma_1(x_i-\overline{x}))}{1-\Gamma_2'(x_i-\overline{x})}. \tag{4.3}$$

for units with $\Gamma_2'(x_i-\overline{x}) \neq 1$*. Moreover,*

(i) Suppose $\pi_i^*(x_i-\overline{x})$ *is affine in* x_i*. Then* τ *has a causal interpretation under* π_i^* *if and only if* π_i^* *is affine in* x_i*. When this happens,* $\pi_i^* = \tilde{\pi}_i$ *for all* i*,* $\pi_i = \tilde{\pi}_i = \pi_i^*$ *for all* i *with* $\Gamma_2'(x_i-\overline{x}) \neq 1$*, and the estimand is a weighted average treatment effect with weights proportional to* $\omega_i \propto \pi_i(1-\pi_i)(1-\Gamma_2'(x_i-\overline{x}))$*.*

(ii) Suppose π_i^* *is affine in* x_i *but* $\pi_i^*(x_i-\overline{x})$ *is generically non-affine in* x_i*. Then* τ *has a causal interpretation under* π_i^* *if and only if* $\frac{1}{n}\sum_{i=1}^n \pi_i^*(1-\pi_i^*)(x_i-\overline{x}) = 0$*. When this happens, the estimand is a weighted average treatment effect with weights proportional to* $\pi_i^*(1-\pi_i^*)$*.*

Proposition 4.2 gives two scenarios where the regression specification (4.2) has causal interpretation under affine propensity scores. The upshot is that neither scenario is plausible, unless either π_i^* is constant or x_i represents saturated discrete covariates. This thus suggests that a design-based interpretation for τ in (4.2) requires a randomized experiment or a saturated discrete covariate. In general, therefore, the "saturate" in interact-and-saturate can not be omitted; the landscape of causal interpretation in binary-treatment regressions is summarized in Table 1.

	Saturated x_i	Randomized Exp.	Non-saturated x_i but linear π_i^*
No interaction	$\pi_i^*(1-\pi_i^*)$-ATE (Angrist, 1998)	ATE	$\pi_i^*(1-\pi_i^*)$-ATE (Angrist, 1998)
Interacted spec.	ATE	ATE (Lin, 2013)	No interpretation (outside of knife-edge)

TABLE 1. Design-based interpretation of controlling for covariates in binary-treatment regressions

The first scenario (i) requires not only that the propensity score is affine in x_i, but also that $\pi_i^*(x_i-\overline{x})$ is affine in x_i. This scenario is unlikely since $\pi_i^*(x_i-\overline{x})$ is a quadratic expression in x_i if π_i^* is affine in x_i. Nevertheless, $\pi_i^*(x_i-\overline{x})$ can be affine when x_i takes at most $\dim(x_i)+1$ distinct values, which effectively means that

x_i represent saturated discrete covariates. It also holds when π_i^* is constant in x_i—i.e., the true design is a randomized experiment (Lin, 2013). On the other hand, if $\pi_i^*(x_i - \overline{x})$ is not affine in x_i, then (ii) states that, under linear propensity scores, τ has a causal interpretation only when the mean covariate value $\overline{x}$ also happens to be the weighted mean under $\pi_i^*(1-\pi_i^*)$. This is a knife-edge scenario[31] and is generically violated.

The reason for this difficulty is that the implicit design (4.3) is defined by a strange expression. This expression is a rearrangement of the condition that the potential weights have mean zero under the true design:

$$\underbrace{\pi_i^* - \tilde{\pi}_i}_{\text{Linear approximation error of } \pi_i^*} = \Gamma_2' \underbrace{\left(\pi_i^*(x_i - \overline{x}) - \Gamma_0 - \Gamma_1(x_i - \overline{x})\right)}_{\text{Linear approximation error of } \pi_i^*(x_i - \overline{x})}. \tag{4.4}$$

Equation (4.4) states that the linear approximation error in the true propensity score just so happens to be linearly related to the linear approximation error in $\pi_i^*(x_i - \overline{x})$, where the coefficient in that relationship is exactly Γ_2. If the true design is linear, then $\pi_i^* - \tilde{\pi}_i = 0$, and the two scenarios (i) and (ii) come from either $\pi_i^*(x_i - \overline{x}) = \Gamma_0 + \Gamma_1(x_i - \overline{x})$ or $\Gamma_2 = 0$. Equation (4.4) also therefore suggests that causal interpretation under nonlinear propensity scores is difficult, since it relies on a coincidental relationship between two approximation errors.

4.1.3. *Multiple treatments, Goldsmith-Pinkham et al. (2022).* Our results also recover insights from the recent work by Goldsmith-Pinkham *et al.* (2022). They consider the following regression specification for a multivalued treatment $W_i \in \{0, \ldots, J\}$, analogous to (4.1):

$$Y_i = \alpha_0 + \sum_{j=1}^{J} \tau_j \mathbb{1}(W_i = j) + \gamma' x_i + \epsilon_i. \tag{4.5}$$

In this specification, we might be tempted to interpret τ_j as a treatment effect between the j^{th} treatment value and the baseline treatment value $W_i = 0$. Goldsmith-Pinkham *et al.* (2022) show that, under propensity scores linear in x_i, τ_j does not solely measure the contrast between treatment j and control ($W = 0$). In contrast, τ_j is contaminated by treatment effects between other treatments ($\ell \notin \{0, j\}$) and the control condition.

It is simple to analyze this regression specification 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.* (2022). Consider the specification (4.5) with

[31]This scenario is not impossible: It holds, say, when x_i has support on $[0, 1]$ that is symmetric about $1/2$, and $\pi_i^*(x) = x$, for instance.

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 population regression specification is described by (4.5) under these assignment probabilities, where 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 (2.2), and thus $\boldsymbol{\pi}^*$ is the only implicit design for this regression specification. The corresponding implicit estimand for τ_j is then

$$\tau_j = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{J} \underbrace{\pi_i(j)\rho_i(j)}_{\omega_i(\boldsymbol{\pi}, j)} y_i(j),$$

which is the sum of potential outcomes weighted by $\omega_i(\boldsymbol{\pi}, j)$.

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.* (2022). For each x_i value and each estimand, we tabulate $\omega_i(\cdot)$ for individuals with those x_i:

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

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.* (2022).[32] Our framework automatically finds qualitatively and quantitatively the same contamination issue—we show that, for this population regression specification, if both τ_1 and τ_2 satisfy level irrelevance, then the design must be the stratified randomized experiment under $\boldsymbol{\pi}^*$, under which τ_1 and τ_2 are contaminated estimands even from a design-based perspective.

Our framework also allows researchers to verify that the interacted version of (4.5) retains a design-based interpretation *in this case with binary* x_i. If we compute the

[32] 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.* (2022).

same potential weights for $\boldsymbol{\tau} = [\tau_1, \tau_2]'$ in the regression with interaction

$$Y_i = \alpha_0 + \sum_{j=1}^{J} \tau_j \mathbb{1}(W_i = j) + \gamma x_i + \sum_{j=1}^{J} \delta_j W_j (x_i - \overline{x}) + \epsilon_i,$$

we find that (a) the only implicit design is $\boldsymbol{\pi}^*$ and (b) under this design, the implicit estimand is the ATE.[33] However, consistent with a natural extension of Proposition 4.2, the same is not true if the covariates are not saturated—we generally find a unique implicit design that does not equal $\boldsymbol{\pi}^*$.

4.2. **Panel settings.** We now turn to panel settings ($T > 1$). We identify the set of treatments as treatment paths $\mathcal{W} \subset \{0,1\}^T$ for a binary treatment. Since we write potential outcomes as $\mathbf{y}_i(\mathbf{w})$, defined with respect to the treatment path, we do not rule out dynamic or anticipated treatment effects. Let us first consider a balanced panel setting for notational clarity; the last result in this subsection concerns unbalanced panels. We define a leading treatment assignment pattern in panel settings—staggered adoption.

Definition 4.1. We say $\mathcal{W} \subset \{0,1\}^T$ satisfies *staggered adoption* if $\mathbf{w}_t \le \mathbf{w}_s$ for all $t \le s$ and for all $\mathbf{w} \in \mathcal{W}$.

4.2.1. *Two-way fixed effects.* A popular specification is the two-way fixed effect (TWFE) regression,[34] where one regresses the outcome on some transformations of the treatment path $f_t(\mathbf{w})$ as well as individual and time treatment effects. This encompasses both the simple regression where $f_t(\mathbf{w}) = \mathbf{w}_t$ is the contemporaneous treatment status and more complex event-study specifications, where $f_t(\mathbf{w})$ is the vector of lags relative to treatment time. We find that the implicit designs of such TWFE specifications contain—sometimes uniquely—the design where units are randomly assigned treatment paths.

Proposition 4.3. *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:*

[33] That is, $\boldsymbol{\omega}_i(\boldsymbol{\pi}^*, 1) = 1, \boldsymbol{\omega}_i(\boldsymbol{\pi}^*, 0) = -1, \boldsymbol{\omega}_i(\boldsymbol{\pi}^*, 2) = 0$ for τ_1 and $\boldsymbol{\omega}_i(\boldsymbol{\pi}^*, 1) = 0, \boldsymbol{\omega}_i(\boldsymbol{\pi}^*, 0) = -1, \boldsymbol{\omega}_i(\boldsymbol{\pi}^*, 2) = 1$ for τ_2.

[34] Proposition B.9 states that one-way fixed effect, in contrast, usually does not have causal interpretation under staggered adoption and compares with recent work by Arkhangelsky and Imbens (2023b).

(i) Whether or not $\boldsymbol{\tau}$ satisfies level irrelevance under the true design $\boldsymbol{\pi}^*$, one proper and Gram-consistent implicit design is

$$\pi_i(\mathbf{w}) \equiv \pi(\mathbf{w}) \equiv \frac{1}{n}\sum_{j=1}^{n} \pi_j^*(\mathbf{w}). \tag{4.6}$$

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) When $f_t(\mathbf{W}_i) = W_{it}$, so that the regression is $Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \epsilon_{it}$, if $\mathcal{W}$ satisfies staggered adoption and $\mathcal{W}$ excludes an always-treated unit, then the set of implicit designs is a singleton equal to (4.6). Correspondingly, the only solution to (3.6) sets $\hat{\pi}_i(\mathbf{w})$ as the empirical frequency of treatment path $\mathbf{w}$ for all i.

The first claim, Proposition 4.3(i), simply states that, algebraically, the design where $\mathbf{W}_i$ is randomly assigned according to their marginal probability solves the level irrelevance condition (2.2) and is Gram-consistent. Thus, if $\boldsymbol{\pi}^*$ does correspond to random assignment, then $\boldsymbol{\tau}$ is a vector of causal contrasts: That is, coefficients in *any* two-way fixed effect specification correspond to causal contrast under random assignment of treatment paths. An empirical consequence is that using (3.6) in practice to find an implicit design will always at least return the design where units are randomly assigned $\mathbf{w}$ according to their empirical frequency. The second claim, Proposition 4.3(ii), shows that, at least in the staggered adoption case, random assignment of treatment time is the *only* design that generates parameters satisfying level irrelevance, given a population regression specification.[35]

Proposition 4.3 echoes and complements the analysis of Athey and Imbens (2022), who analyze the specification $Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \epsilon_{it}$ under staggered adoption and *complete random assignment* of treatment timing. This design is a fixed Gram design, and thus the potential weights are directly observable without estimation error. In particular, following (3.5), the TWFE estimator can be written as

$$\hat{\tau} = \sum_{t=1}^{T}\sum_{\mathbf{w}\in\mathcal{W}} \boldsymbol{\omega}_t(\boldsymbol{\pi}^*, \mathbf{w})\overline{y}_{t,\mathbf{w}} \quad \boldsymbol{\omega}_t(\boldsymbol{\pi}^*, \mathbf{w}) \equiv \pi^*(\mathbf{w})\boldsymbol{\rho}_t(\mathbf{w})$$

where $\overline{y}_{t,\mathbf{w}}$ is the sample mean of time-t outcomes of individuals receiving treatment path $\mathbf{w}$.[36] This decomposition is identical to Lemma 5 in Athey and Imbens (2022). The expectation of $\hat{\tau}$, under completely random treatment timing, is the following

[35] There is nothing particularly special about staggered adoption—the same argument for uniqueness extends to any case where the nonzero paths in $\mathcal{W}$ are linearly independent and do not span 1_T.

[36] We drop the i subscript since all π_i and $\boldsymbol{\omega}_i$ are the same.

weighted average of mean potential outcomes

$$\sum_{t=1}^{T} \sum_{\mathbf{w} \in \mathcal{W}} \boldsymbol{\omega}_t(\boldsymbol{\pi}^*, \mathbf{w}) \left(\frac{1}{n} \sum_{i=1}^{n} y_{it}(\mathbf{w}) \right),$$

which echoes Theorem 1 in Athey and Imbens (2022). While Athey and Imbens (2022) show that the two-way fixed effect estimand has this interpretation under random treatment timing, Proposition 4.3(3) shows a converse: τ fails to satisfy level irrelevance under any other design with the same population Gram matrix.

Furthermore, we can often inspect the weights $\boldsymbol{\omega}_t(\boldsymbol{\pi}^*, \mathbf{w})$ and verify if the weights satisfy some additional requirements. For instance, 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, Sant'Anna, Bilinski and Poe, 2023; Borusyak, Jaravel and Spiess, 2024b; De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021). Proposition B.10 shows that when $\mathcal{W}$ has only two elements and includes a never treated unit (i.e., classical difference-in-difference), all weights post treatment are non-negative, but such forbidden comparisons are possible in other cases.

4.2.2. *TWFE with time varying covariates.* It is common empirical practice to include time-varying covariates in a two-way fixed effect regression. Surprisingly, by our criteria, such a specification usually does not have a design-based interpretation—even when we allow for negative weights on treated potential outcomes—unless the time-varying covariates do not covary with treatment paths. This is in the sense that the set of implicit designs can be and often is empty. In practice, this also means that the equation (3.6) would fail to admit any (proper and Gram-consistent) solutions $\hat{\boldsymbol{\pi}}$.

Proposition 4.4. *Let the population regression specification be described by the regression $Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \delta' x_{it}$ under $\boldsymbol{\pi}^*$ where τ is the coefficient of interest. Let $\beta_{w \to x}$ be the population projection coefficient of W_{it} on x_{it}, including individual and time fixed effects. For the set of implicit designs to be non-empty, a necessary condition is*

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

Proposition 4.4 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. Under staggered adoption, when there are relatively few adoption dates, $\text{span}(\mathcal{W} \cup \{1_T\})$ is a small linear subspace of $\mathbb{R}^T$. If $\beta_{w\to x}$ is nonzero under $\boldsymbol{\pi}$, it is thus highly 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, barring special structure on the covariates $\mathbf{x}_i$. On the other hand, if $\beta_{w\to x}$ is zero under $\boldsymbol{\pi}^*$, then by the Frisch–Waugh–Lovell theorem, including the covariates makes no difference to the coefficient on W_{it} in population. In any case, it is highly unlikely that the sample analogue of $\beta_{w\to x}$ is zero, and so we would usually fail to find any solution to (3.6).

These caveats aside, we interpret Proposition 4.4 to mean that two-way fixed effects regression with time-varying covariates does not in general have a design-based interpretation, in the sense that the set of implicit designs is empty. 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.

4.2.3. *TWFE with unbalanced panels.* Lastly, we consider unbalanced panels. Suppose units are observed in some subperiod $\mathcal{T}_i \subset \{1, \ldots, T\}$. We consider the TWFE specification

$$Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \epsilon_{it} \tag{4.7}$$

over $i = 1, \ldots, n$ and $t \in \mathcal{T}_i$. The objects $\mathcal{T}_1, \ldots, \mathcal{T}_n$ are considered fixed and observed. Much like the case with time-varying covariates, we find that TWFE with unbalanced panels also does not have a causal interpretation under staggered adoption—unless treatment time is uniformly randomly assigned and uncorrelated with $\mathcal{T}_i$.

To introduce the result, 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. 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, $\mathcal{T} = \{t_{\min} - 1, \ldots, t_{\max}\}$—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.

Proposition 4.5. *Let the population regression specification be characterized by* (4.7) *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* $t \in [T]$*:*

$$\sum_{\mathbf{w}\in\mathcal{W}} \overline{\pi}(\mathbf{w}) \sum_{i:t\in\mathcal{T}_i} (\mathbf{w}_t - Q_i(\mathbf{w})) = \sum_{\mathbf{w}\in\mathcal{W}} \sum_{i:t\in\mathcal{T}_i} \pi_i^*(\mathbf{w})(\mathbf{w}_t - Q_i(\mathbf{w})). \tag{4.8}$$

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

Proposition 4.5 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. When it exists, the implicit design is equal to a design where the treatment timing is randomly assigned. Therefore, if $\boldsymbol{\pi}^*$ justifies τ as a design-based estimand in the sense of Definition 2.3, then $\boldsymbol{\pi}^*$ must equal to the implicit design and thus must describe random assignment.

Proposition 4.5 generalizes Proposition 4.3. When applied to a balanced panel, the condition (4.8) is always satisfied, and we recover the characterization of the implicit design in Proposition 4.3. With unbalanced panels, the condition (4.8) is in general difficult to satisfy unless $\boldsymbol{\pi}^*$ describes random assignment, and when it fails the set of implicit designs is empty. In this sense, (4.7) usually does not have a causal interpretation unless the treatment timing is uniformly randomly assigned.

Proposition 4.5 assumes that there is a commonly observed period where the treatment paths have rich variation. This is not a strong condition. If every unit's observed period includes the period spanning all adoption dates, then this condition is automatically satisfied. We also emphasize that this condition is only a sufficient condition for a clean theoretical characterization, and we expect the set of implicit designs to be generally empty when $\boldsymbol{\pi}^*$ is not a random assignment of treatment timing, outside of certain knife-edge configurations.

The sample analogue of (4.8) is that for all $t \in [T]$,

$$\underbrace{\sum_{\mathbf{w}\in\mathcal{W}} \hat{\pi}(\mathbf{w}) \sum_{i:t\in\mathcal{T}_i} (\mathbf{w}_t - Q_i(\mathbf{w}))}_{\substack{\text{Expected sum (among units observing } t\text{) of demeaned} \\ \text{treatment, if treatment is randomly assigned according to } \hat{\pi}(\cdot)}} = \underbrace{\sum_{i:t\in\mathcal{T}_i} (W_{it} - Q_i(\mathbf{W}_i))}_{\substack{\text{Sum of realized demeaned} \\ \text{treatment among units observing } t}},$$

where $\hat{\pi}(\mathbf{w})$ is the empirical frequency of $\mathbf{W}_i = \mathbf{w}$. This is a difficult condition to satisfy unless the empirical frequency of each path is identical across units with different observation patterns $\mathcal{T}_i$. Thus, in practice, the set of estimated implicit designs is usually empty, unless the treatment timing is completely randomized within each observation pattern. Thus, even if the true design $\boldsymbol{\pi}^*$ is random assignment, the estimated implicit designs usually do not exist, illustrating the difficulty of estimating implicit designs that we allude to in Section 3. Nevertheless, when the set of estimated implicit designs is empty, we can still reject the existence of a fixed-Gram design that rationalizes τ as a causal contrast—thus ruling out a result like Athey and Imbens (2022) for the unbalanced setting.

5. Discussion and conclusion

This paper examines the design-based interpretation of linear regression estimators. Rather than asking whether a specification estimates a causal parameter under a given design, we fix the linear specification and ask what design exists to rationalize the linear regression as a causal estimator. We define potential weights and find that they encode the causal comparisons the linear estimator takes, and seek a design that rationalizes these comparisons. We call these designs implicit designs. The potential weights themselves are also readily estimable in practice. We show that these potential weights provide a simple diagnostic for whether a linear estimator has a design-based interpretation.

When applied to a number of specifications encompassing binary treatment, multivalued treatment, and panel settings, the implicit designs and estimands conform either to intuition or to existing theoretical results. For instance, the implicit design of a outcome-on-treatment-and-covariates regression is a stratified randomized experiment, whose treatment assignment probability is linear in the covariates (Angrist, 1998). The implicit estimand of this regression, with multivalued treatments, turns out to suffer the same contamination bias as shown in (Goldsmith-Pinkham *et al.*, 2022). In panel settings with staggered adoption, the two-way fixed effect regression has implicit design equal to randomization of treatment timing (Athey and Imbens, 2022).

These results supply converses to these existing intuition and results. Namely, not only do these regressions have design-based interpretations under these treatment assignment processes, but these processes are the only processes consistent with a causal interpretation and the population Gram matrix of the regression estimator. Our framework uncovers new results as well. We find that a number of common regression specifications do not have design-based interpretations, unless certain knife-edge conditions hold.

An advantage of our approach is that it is extremely widely applicable. The potential weights of any regression can be estimated by simple formulae—and some potential weights are already calculated when one calculates the OLS estimator. We hope these potential weights and the corresponding implicit design and implicit estimand are useful to practitioners, and we conclude the paper by discussing a few ways these objects can be used.

A diagnostic for whether a design-based interpretation is possible. First, whether a plausible implicit design exists in the first place provides a binary diagnostic for the design-based interpretation of a regression specification, especially for those not covered by existing theoretical results. Finding a plausible implicit design that generates a sensible implicit estimand shows that there is *some* treatment assignment mechanism under which the regression specification is sensible, regardless of potential outcomes.[37] If the researcher believes that the implicit design accurately describes treatment assignment probabilities, then they can rest assured that the regression estimator would not suffer a "negative weights" problem.

On the other hand, if one cannot find an implicit design that generates a sensible implicit estimand, then this calls for either defending outcome-modeling assumptions or for more explicitly modeling the treatment assignment process instead. Regression specifications are naturally interpreted as outcome models (e.g., in a cross-sectional context, we may model $\mathbb{E}[Y(w) \mid X = x] = z(x, w)$). Though these models often build in implicit assumptions about treatment effect homogeneity and other functional form assumptions, they nevertheless can be useful and plausible. One way forward is thus explicitly stating the regression specification as structural assumptions on the joint distribution of *potential outcomes* (consistent with recommendations in Rubin, 2019).

If the researcher is nevertheless confident in a design-based interpretation free of outcome modeling, then one way forward is to abandon the regression altogether and use methods that put design first. These methods include weighting, matching, balancing, and doubly robust approaches that have been recently influential in the causal inference literature. Of course, these methods in turn require an explicit model of the design and often require explicit estimation of the relevant propensity scores.

[37]Even if a proper implicit design exists, it might not be plausible. Perhaps it is poorly calibrated—meaning that those with treatment assignment probability approximately p do not appear to receive treatment with proportion p. Or perhaps the implicit design relates to the covariates x_i in a way that is not economically plausible.

Here, the implicit design might guide and anchor our modeling choices. Perhaps the researcher, by choosing the regression specification considered, is encoding some substantive knowledge of the design. In that case, we may have reason to suspect that the implicit design of the regression is "close" to the true design, despite the implicit design being implausible (e.g. containing probabilities that lie outside of $[0, 1]$). In that case, the implicit design might serve as a reasonable starting point for modeling the treatment assignment process. Perhaps a simple way forward is to simply re-calibrate the implicit design. Practically, for a binary treatment, we can discretize the values of $\hat{\pi}_i$ into bins, and for an observation i in a given bin, recompute $\tilde{\pi}_i \in [0, 1]$ as the observed frequency of treatment within the bin.[38] One can then estimate causal effects by, say, inverse propensity weighting with $\tilde{\pi}_i$. Such a procedure would admit a design-based interpretation, and it is in a sense close to the original regression specification.

Assessing whether the implicit design is reasonable. Suppose an implicit design—as well as the associated estimand—does exist and the regression passes the basic diagnostic. Having the implicit design means researchers can inspect it and check whether they conform to economic intuition. One way is to check whether the relationship between $\hat{\pi}_i$ and covariates x_i—preferably those excluded from the regression specification—conform to economic intuition. Another is to check whether the $\hat{\pi}_i$ are well calibrated, by comparing observed values W_i against $\hat{\pi}_i$, expecting that $\mathbb{E}[W_i \mid \hat{\pi}_i] \approx \hat{\pi}_i$. If $\hat{\pi}_i$ are poorly calibrated, then the regression specification is correspondingly suspect.

Design-based inference. Finally, if we are willing to take the regression specification—and therefore the implicit design—seriously, then we might consider conducting design-based inference, particularly in settings (e.g., settings where units are geographies) where the superpopulation sampling thought experiment is difficult to rationalize (Rambachan and Roth, 2020; Borusyak and Hull, 2023). 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

[38] $\tilde{\pi}_i$ can be viewed as calibrated versions of $\hat{\pi}_i$ via histogram binning. We can likewise consider other methods from the machine learning literature on calibrating binary classifiers (Naeini, Cooper and Hauskrecht, 2014). This procedure can also be viewed as blocking on an estimated propensity score (Chapter 17, Imbens and Rubin, 2015) where the estimated propensity score is the implicit design of a regression specification.

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 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, if one uses the regression specification and is internally consistent, then one is 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., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, **88** (1), 265–296. 7

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

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, 6, 24, 25, 28, 36

— and Pischke, J.-S. (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. (2023a). *Causal models for longitudinal and panel data: A survey*. Tech. rep., National Bureau of Economic Research. 7

— and Imbens, G. W. (2023b). Fixed effects and the generalized mundlak estimator. *Review of Economic Studies*, p. rdad089. 31, 63

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, 11, 24, 32, 33, 36

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

Blandhol, C., Bonney, J., Mogstad, M. and Torgovitsky, A. (2022). *When is TSLS actually late?* Tech. rep., National Bureau of Economic Research. 3, 12

Borusyak, K. and Hull, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica*, **91** (6), 2155–2185. 38

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

—, — 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. 33

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

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. 33

Durrett, R. (2019). *Probability: theory and examples*, vol. 49. Cambridge university press. 19, 44

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

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

—, Hull, P. and Kolesár, M. (2022). *Contamination bias in linear regressions*. Tech. rep., National Bureau of Economic Research. 3, 5, 6, 24, 27, 29, 30, 36

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of econometrics*, **225** (2), 254–277. 33

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35** (4), 1491–1523. 20, 58

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

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

— and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press. 38

— and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, **47** (1), 5–86. 27

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, **112** (520), 1759–1769. 7

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. 3, 6, 25, 27, 28, 29

Manski, C. F. (2022). Identification and statistical decision theory. *Econometric Theory*, pp. 1–17. 10

Naeini, M. P., Cooper, G. F. and Hauskrecht, M. (2014). Binary classifier calibration: Non-parametric approach. *arXiv preprint arXiv:1401.3390*. 38

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472. 11

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*. 5, 7, 11, 20, 38, 58

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. 33

Rubin, D. B. (2019). Essential concepts of causal inference: a remarkable history and an intriguing future. *Biostatistics & Epidemiology*, **3** (1), 140–155. 37

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

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. 11

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. Proofs

Theorem 2.1. *$\boldsymbol{\tau}$ is a vector of causal contrasts in the sense of Definition 2.3 if and only if it is equal to some implicit design. When this happens, if, furthermore, there is a unique implicit design $\boldsymbol{\pi}$, then $\boldsymbol{\pi}$ is proper, Gram-consistent, and equal to $\boldsymbol{\pi}^*$.*

Proof. For the first statement and the $\Longrightarrow$ direction, if $\boldsymbol{\tau}$ is a vector of causal contrasts for the regression specification, then $\pi_i^*(\cdot)$ solves (2.2), and thus $\boldsymbol{\pi}^*$ is an implicit design. For the $\Longleftarrow$ direction, if $\boldsymbol{\pi}^*$ is an implicit design, then $\boldsymbol{\pi}^*$ solves (2.2), 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. Moreover, $\boldsymbol{\pi}^*$ is proper and Gram-consistent by definition. □

Corollary 2.2 (Binary treatment, cross-sectional setting)**.** *In the cross-sectional, binary treatment setting of Example 2.1, τ is a causal contrast under the true design $\boldsymbol{\pi}^*$ if and only if $\rho_i(1)\rho_i(0) \le 0$ for all i and*

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

for all i with one of $\rho_i(1)$ and $\rho_i(0)$ nonzero.

When this happens, the implicit estimand is defined by $\omega_i(\boldsymbol{\pi}^*, 1) = \pi_i^* \rho_i(1)$ and $\omega_i(\boldsymbol{\pi}^*, 0) = -\omega_i(\boldsymbol{\pi}_i^*, 1)$, 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 $\Longrightarrow$ 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 $\Longleftarrow$ direction, suppose $\rho_i(1)\rho_i(0) \le 0$ 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 other units, the condition holds trivially since both potential weights are zero. Thus τ satisfies level irrelevance under π^*.

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. 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 $\rho_i(1)\rho_i(0) \leq 0$. □

Lemma 3.1. *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 $1 \leq j \leq K, 1 \leq \ell \leq K$, the average second moment of $G_{i,j\ell}$ grows slower than n: as $n \to \infty$, $\frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\pi_{i,n}^*}[G_{i,j\ell}^2]\right) \to 0$. Then Assumption 3.1 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

$$\sigma_n^2 \leq \sum_{i=1}^n \mathbb{E}[G_{i,jk}^2].$$

Thus the condition that $\frac{1}{n^2}\sum_i \mathbb{E}[G_{i,jk}^2] \to 0$ 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 3.2. *Under Assumptions 3.1 and 3.2, the estimated potential weights are consistent: $\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w}) \xrightarrow{p} 0$. 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:*

$$\max_{i \in [n], \mathbf{w} \in \mathcal{W}} |\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty \xrightarrow{p} 0. \tag{3.2}$$

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

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

The first term converges to zero by Assumptions 3.1 and 3.2 and the Hoffman–Wielandt inequality. Note that, by the submultiplicativity of the Frobenius norm, when $\lambda_{\min}(\hat{G}_n) > \epsilon/2$,

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

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

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

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

For the uniformity in i, note that, if $\lambda_{\min}(\hat{G}_n) > \epsilon/2$, up to constants,

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

Therefore,

$$\mathrm{P}\left(\max_i |\hat{\boldsymbol{\rho}}_i(\mathbf{w}) - \boldsymbol{\rho}_i(\mathbf{w})|_\infty > \eta\right) \le \mathrm{P}[\lambda_{\min}(\hat{G}_n) \le \epsilon/2] + \mathrm{P}\left[\|\hat{G}_n^{-1} - G_n^{-1}\|_F \gtrsim \eta\right] \to 0.$$

□

Proposition 3.3. *Assume that the sequence of populations is such that the average covariance between the covariates and potential outcomes is bounded*

$$\frac{1}{n}\sum_{i=1}^{n} \max_{\mathbf{w}\in\mathcal{W}} \|\mathbf{z}(\mathbf{x}_i, \mathbf{w})'\mathbf{y}_i(\mathbf{w})\|_\infty = O(1).$$

Under Assumptions 3.1 and 3.2, let constants $b_n > 0$ be such that $|\hat{G}_n - G_n|_\infty = O_P(b_n)$. Then $\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}_{\mathrm{HT}}\|_\infty = O_P(b_n)$.

Proof. Fix $M > 0$ and sequence $a_n > 0$. We compute

$$\mathrm{P}\left[a_n\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}_{\mathrm{HT}}\|_\infty > M\right] \le \mathrm{P}[\lambda_{\min}(\hat{G}_n) \le \epsilon/2] + \mathrm{P}\left[a_n\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}_{\mathrm{HT}}\|_\infty > M, \lambda_{\min}(\hat{G}_n) > \epsilon/2.\right]$$

Note that if $\lambda_{\min}(\hat{G}_n) > \epsilon/2$, then, since the Frobenius norm is submultiplicative,

$$\begin{aligned}\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}_{\mathrm{HT}}\|_2 &\le \|\Lambda\|_F \|\hat{G}_n^{-1} - G_n^{-1}\|_F \cdot \underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}(\mathbf{x}_i, \mathbf{W}_i)'\mathbf{y}_i\right\|_F}_{O(1) \text{ by assumption}} \\ &\le \frac{2}{\epsilon^2} O(\|\hat{G}_n - G_n\|_F) \\ &= O\left(|\hat{G}_n - G_n|_\infty\right).\end{aligned}$$

Thus, for some C,

$$\mathrm{P}\left[a_n\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}_{\mathrm{HT}}\|_\infty > M\right] \le \mathrm{P}[\lambda_{\min}(\hat{G}_n) \le \epsilon/2] + \mathrm{P}\left[Ca_n|\hat{G}_n - G_n|_\infty > M\right].$$

Take $a_n = b_n^{-1}$, we have that for every $\eta > 0$, there exists M such that the right-hand side is bounded by η for all sufficiently large n. This is equivalent to

$$\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}_{\mathrm{HT}}\|_\infty = O_P(b_n).$$

□

Proposition 3.4. *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 probability vector 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 (3.2) *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

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

Thus,

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

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

Proposition 3.5. *Assume that $k = J$. Suppose that:*

(i) The estimand $\boldsymbol{\tau}$ is a vector of causal contrasts in the sense of Definition 2.3 *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 (3.2).

Then, for some $C > 0$, a unique estimated implicit design $\hat{\boldsymbol{\pi}}$ that is bounded by 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 $\boldsymbol{\pi}^*$ in the sense that

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

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 $R_i\pi_i^* = 0$. 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. 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

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

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

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

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

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

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

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

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 so that $1/C < 0.207$, then

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

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

Now,

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

Therefore,

$$\|\hat{\pi}_i - \pi_i\|_2 \leq \underbrace{\left(\frac{1}{0.207}(\sqrt{J+1} + 1)\frac{2(1+\sqrt{2})}{\epsilon}\right)}_{M/\epsilon}\|R_i - \hat{R}_i\|_F.$$

Hence, for $\eta > 0$,

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

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. □

Proposition 4.1. *Let the population regression specification be described by* (4.1) *under the design* $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_n^*)$*, where the coefficient of interest is* τ*. Let* $\pi_i = x_i'\beta_{w\to x}$ *be the* *linear projection of $\boldsymbol{\pi}^*$ onto x_i:*

$$\beta_{w\to x} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n x_i \pi_i^*.$$

Then the improper implicit design is unique and equal to $\boldsymbol{\pi}$. If τ is a causal contrast under $\boldsymbol{\pi}^*$, then (i) $\boldsymbol{\pi}^* = \boldsymbol{\pi}$, (ii) $\boldsymbol{\pi}^*$ is linear in x_i, and (iii) $\boldsymbol{\pi}$ is proper. When that happens, the implicit estimand is the $\pi_i(1-\pi_i)$-weighted average treatment effect

$$\tau = \frac{1}{n}\sum_{i=1}^n \underbrace{\frac{\pi_i(1-\pi_i)}{\frac{1}{n}\sum_{j=1}^n \pi_j(1-\pi_j)}}_{\omega_i}(y_i(1)-y_i(0)),$$

and ω_i is nonnegative if and only if $\pi_i \in [0,1]$.

Proof. By Theorem B.1, the potential weights for this regression specification are the same as the regression specification with the covariate $W_i - x_i'\beta_{w\to x}$, where $\beta_{w\to x}$ is the population projection of W_i (and hence π_i^*) on x_i. Thus the potential weights are

$$\rho_i(w) = V^{-1}\left[w - x_i'\beta_{w\to x}\right].$$

where $V > 0$ by Theorem B.1. Note that since $V > 0$, $\rho_i(1) \neq \rho_i(0)$ for all i, and thus the implicit design is unique and equal to $\pi_i = x_i'\beta_{w\to x}$. The corresponding implicit weight ω_i is then $V^{-1}\pi_i(1-\pi_i)$.

Lastly, we verify that $V = \frac{1}{n}\sum_{i=1}^n \pi_i(1-\pi_i)$:

$$0 < V = \frac{1}{n}\sum_i \pi_i(1-\pi_i)^2 + (1-\pi_i)\pi_i^2 = \frac{1}{n}\sum_i \pi_i(1-\pi_i).$$

Note that $\omega_i = \pi_i(1-\pi_i)/V$ is negative if and only if $\pi_i \notin [0,1]$. □

Proposition 4.2. *Under the setting described above for the specification (4.2), the implicit design is equal to*

$$\pi_i = \frac{\tilde{\pi}_i - \Gamma_2'(\Gamma_0 + \Gamma_1(x_i - \overline{x}))}{1 - \Gamma_2'(x_i - \overline{x})}. \tag{4.3}$$

for units with $\Gamma_2'(x_i - \overline{x}) \neq 1$. Moreover,

(i) Suppose $\pi_i^*(x_i - \overline{x})$ is affine in x_i. Then τ has a causal interpretation under π_i^* if and only if π_i^* is affine in x_i. When this happens, $\pi_i^* = \tilde{\pi}_i$ for all i, $\pi_i = \tilde{\pi}_i = \pi_i^*$ for all i with $\Gamma_2'(x_i - \overline{x}) \neq 1$, and the estimand is a weighted average treatment effect with weights proportional to $\omega_i \propto \pi_i(1-\pi_i)(1 - \Gamma_2'(x_i - \overline{x}))$.

(ii) Suppose π_i^* is affine in x_i but $\pi_i^*(x_i - \overline{x})$ is generically non-affine in x_i. Then τ has a causal interpretation under π_i^* if and only if $\frac{1}{n}\sum_{i=1}^n \pi_i^*(1-\pi_i^*)(x_i - \overline{x}) = 0$. When this happens, the estimand is a weighted average treatment effect with weights proportional to $\pi_i^*(1-\pi_i^*)$.

Proof. By Theorem B.1, the potential weights for this regression specification are equivalent to the potential weights in the specification

$$Y_i = \tau(W_i - \tilde{\pi}_i - \Gamma_2'(W_i(x_i - \overline{x}) - \Gamma_0 - \Gamma_1(x_i - \overline{x}))) + \epsilon_i.$$

We can then readily compute that

$$\rho_i(0) = V^{-1}(-\tilde{\pi} - \Gamma_2'(-\Gamma_0 - \Gamma_1(x_i - \overline{x}))) \quad \rho_i(1) = V^{-1}(1 - \tilde{\pi} - \Gamma_2'(x_i - \overline{x} - \Gamma_0 - \Gamma_1(x_i - \overline{x}))).$$

where $V > 0$ by Theorem B.1.

Thus, all implicit designs must satisfy

$$\pi_i = \frac{-\rho_i(0)}{\rho_i(1) - \rho_i(0)} = \frac{\tilde{\pi}_i - \Gamma_2'(\Gamma_0 + \Gamma_1(x_i - \overline{x}))}{1 - \Gamma_2'(x_i - \overline{x})}$$

for those with $1 - \Gamma_2'(x_i - \overline{x}) \neq 0$.

For (i), τ is a causal contrast under $\boldsymbol{\pi}^*$ if and only if $\boldsymbol{\pi}^*$ belongs to the set of implicit designs. This is further equivalent to (4.4). When $\pi_i^*(x_i - \overline{x})$ is affine, then the right-hand side of (4.4) is zero, and thus (4.4) is equivalent to $\boldsymbol{\pi}^* = \tilde{\boldsymbol{\pi}}$, meaning that $\boldsymbol{\pi}^*$ is affine in x_i. When this happens,

$$\pi_i = \frac{\pi_i^* - \pi_i^* \Gamma_2'(x_i - \overline{x})}{1 - \Gamma_2'(x_i - \overline{x})} = \pi_i^*$$

when $1 - \Gamma_2'(x_i - \overline{x}) \neq 0$. Computing the implicit weight in this case yields

$$\begin{aligned}
\omega_i &= \pi_i^* \rho_i(1) \\
&= V^{-1}\left[\pi_i^*(1 - \pi_i^*) - \Gamma_2'(\pi_i^*(x_i - \overline{x}) - \pi_i^* \Gamma_0 - \pi_i^* \Gamma_1(x_i - \overline{x}))\right] \\
&= V^{-1}\left[\pi_i^*(1 - \pi_i^*) - \Gamma_2'(1 - \pi_i^*)(\Gamma_0 + \Gamma_1(x_i - \overline{x}))\right] && (\pi_i^*(x_i - \overline{x}) = \Gamma_0 + \Gamma_1(x_i - \overline{x})) \\
&= V^{-1}\pi_i^*(1 - \pi_i^*)(1 - \Gamma_2'(x_i - \overline{x})). && (\pi_i^*(x_i - \overline{x}) = \Gamma_0 + \Gamma_1(x_i - \overline{x}))
\end{aligned}$$

Hence, ω_i is proportional to $\pi_i^*(1 - \pi_i^*)(1 - \Gamma_2'(x_i - \overline{x}))$.

For (ii), τ is a causal contrast if and only if (4.4) holds. Since π_i^* is assumed to be affine, the left-hand side of (4.4) is zero. Since $\pi_i^*(x_i - \overline{x})$ is generically non-affine, the right-hand side is zero for every i if and only if $\Gamma_2 = 0$. By the definition of Γ_2 as a projection coefficient, it is zero if and only if

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W_i}\left[(W_i - \tilde{\pi}_i) \cdot (W_i(x_i - \overline{x}) - \Gamma_0 - \Gamma_1(x_i - \overline{x}))\right] = 0.$$

This condition is further equivalent to

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

When this happens, since $\Gamma_2 = 0$, the implicit weight is proportional to $\pi_i^*(1 - \pi_i^*)$. □

Proposition 4.3. *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*

$$\pi_i(\mathbf{w}) \equiv \pi(\mathbf{w}) \equiv \frac{1}{n}\sum_{j=1}^n \pi_j^*(\mathbf{w}). \tag{4.6}$$

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) When $f_t(\mathbf{W}_i) = W_{it}$*, so that the regression is* $Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \epsilon_{it}$*, if* $\mathcal{W}$ *satisfies staggered adoption and* $\mathcal{W}$ *excludes an always-treated unit, then the set of implicit designs is a singleton equal to (4.6). Correspondingly, the only solution to (3.6) sets* $\hat{\pi}_i(\mathbf{w})$ *as the empirical frequency of treatment path* $\mathbf{w}$ *for all* i*.*

Proof. (i) By Theorem B.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 (3.6).

(ii) In this case, the population residuals are

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

Any implicit design $\tilde{\pi}_i$ satisfies

$$\sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}_i(\mathbf{w})\mathbf{w} - 1_T\sum_{\mathbf{w}\in\mathcal{W}} \tilde{\pi}_i(\mathbf{w})\frac{1_T'\mathbf{w}}{T} - \sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w})\mathbf{w} + 1_T\sum_{\mathbf{w}\in\mathcal{W}} \pi(\mathbf{w})\frac{1_T'}{T}\mathbf{w} = 0$$

This implies that the vector

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

However, the only vector in $\operatorname{span}(1_T)$ and $\operatorname{span}(\mathcal{W})$ 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, $\pi(\mathbf{w})$ in (4.6) is the unique implicit design. The proof for the uniqueness of the estimated implicit design is analogous. □

Proposition 4.4. *Let the population regression specification be described by the regression* $Y_{it} = \alpha_i + \gamma_t + \tau W_{it} + \delta' x_{it}$ *under* $\boldsymbol{\pi}^*$ *where* τ *is the coefficient of interest. Let* $\beta_{w\to x}$ *be the population projection coefficient of* W_{it} *on* x_{it}*, including individual and time fixed effects. For the set of implicit designs to be non-empty, a necessary condition is*

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

Proof. By Theorem B.1, we can focus on the population regression of $\mathbf{y}_i$ on $\ddot{\mathbf{W}}_i - \ddot{\mathbf{x}}_i \beta_{w \to x}$. 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 B.1. Here, the notation $\ddot{\mathbf{z}}_i$ for $T \times m$ matrices $\mathbf{z}_1(\mathbf{w}), \ldots, \mathbf{z}_n(\mathbf{w})$ is defined as the population residual against unit and time fixed effects:

$$\ddot{\mathbf{z}}_i = \mathbf{z}_i - 1_T \frac{1_T' \mathbf{z}_i}{T} - \frac{1}{n} \sum_{j=1}^n \sum_{\mathbf{w}} \mathbf{z}_j(\mathbf{w}) \pi_j^*(\mathbf{w}) + 1_T \frac{1_T'}{T} \frac{1}{n} \sum_{j=1}^n \sum_{\mathbf{w}} \mathbf{z}_j(\mathbf{w}) \pi_j^*(\mathbf{w}).$$

The condition (2.2) is then

$$\begin{aligned} 0 &= \sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) \rho_i(\mathbf{w}) \\ \implies \ddot{\mathbf{x}}_i \beta_{w \to x} &= \sum_{\mathbf{w} \in \mathcal{W}} \pi_i(\mathbf{w}) \ddot{\mathbf{w}} \end{aligned}$$

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

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

Proposition 4.5. *Let the population regression specification be characterized by (4.7) 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 $t \in [T]$:

$$\sum_{\mathbf{w} \in \mathcal{W}} \overline{\pi}(\mathbf{w}) \sum_{i : t \in \mathcal{T}_i} (\mathbf{w}_t - Q_i(\mathbf{w})) = \sum_{\mathbf{w} \in \mathcal{W}} \sum_{i : t \in \mathcal{T}_i} \pi_i^*(\mathbf{w}) (\mathbf{w}_t - Q_i(\mathbf{w})). \tag{4.8}$$

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 $R_i(\mathbf{w}) = \sum_{s \in \mathcal{T}_i} \mathbf{w}_s$. 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 B.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 B.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, $t \in \mathcal{T}_i$,

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

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 (A.1), 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 $s\in\{1,\ldots,T\}$,

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\ddot{\mathbf{w}}_{it}\dot{\mathbb{1}}_{its} \\
&= \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\sum_{t\in\mathcal{T}_i}\left(\mathbf{w}_t - R_i(\mathbf{w})/L_i - \sum_{\ell\in\mathcal{T}_i}\delta_\ell\dot{\mathbb{1}}_{it\ell}\right)\dot{\mathbb{1}}_{its} \\
&= \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\sum_{t\in\mathcal{T}_i}\left(\mathbf{w}_t - R_i(\mathbf{w})/L_i\right)\dot{\mathbb{1}}_{its} - \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{t\in\mathcal{T}_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell\dot{\mathbb{1}}_{it\ell}\dot{\mathbb{1}}_{its} \\
&= \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\sum_{t\in\mathcal{T}_i}\left(\mathbf{w}_t - R_i(\mathbf{w})/L_i\right)\mathbb{1}(t=s) - \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{t\in\mathcal{T}_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell\mathbb{1}(t=\ell)\dot{\mathbb{1}}_{its} \\
&\hspace{20em}(\dot{\mathbf{w}} \text{ and } \dot{\mathbb{1}}_i \text{ sum to zero over } t\in\mathcal{T}_i) \\
&= \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\left(\mathbf{w}_s - R_i(\mathbf{w})/L_i\right) - \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell\dot{\mathbb{1}}_{i\ell s} \\
&= \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\left(\mathbf{w}_s - R_i(\mathbf{w})/L_i\right) - \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell(\mathbb{1}(\ell=s)-1/L_i)
\end{aligned}$$

Rearranging, we have that the orthogonality condition is equivalent to

$$\frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\sum_{\mathbf{w}\in\mathcal{W}}\pi_i^*(\mathbf{w})\left(\mathbf{w}_s - Q_i(\mathbf{w})\right) = \frac{1}{n}\sum_{i:s\in\mathcal{T}_i}\left(\delta_s - \frac{1}{L_i}\sum_{\ell\in\mathcal{T}_i}\delta_\ell\right) \tag{A.2}$$

for all $s\in[T]$.

Pick $t\in\mathcal{T}$ and consider the corresponding (A.2). Note that $t\in\mathcal{T}_i$ for all i, and hence

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

Therefore, again, given (A.1) and $\pi_i(\mathbf{w}) = \pi(\mathbf{w})$,

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

for some $M'(i)$ that does not depend $t \in \mathcal{T}$. By the same argument as above, 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 (A.1), and then plug the expression into (A.2). 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 (4.8). 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 (4.8) holds. For a given t, set

$$\delta_t = \sum_{\mathbf{w}\in\mathcal{W}} \overline{\pi}(\mathbf{w})\mathbf{w}_t, \tag{A.3}$$

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 (A.1). Thus $\overline{\pi}(\cdot)$ is an implicit design if (A.3) 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 (4.8). This verifies the orthogonality condition (A.2) for t. Therefore, our choice of (A.3) does indeed equal the projection coefficients. This concludes the proof. □

Appendix B. Additional results

B.1. Invariances.

Theorem B.1 (Frisch–Waugh–Lovell, in population)**.** *Consider a population regression specification in the sense of Definition 2.1. 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}: $\Lambda z_t = \Lambda_1 z_{t1}$. 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}. Observe that $\Gamma = G_{12} G_{22}^{-1}$ and that $\tilde{G}_n = G_{11} - \Gamma G_{21}$. This proves (i).

Since G_n is symmetric and positive definite by assumption, G_{11} and G_{22} are invertible. 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 B.2 (Frisch–Waugh–Lovell, in sample)**.** *Consider a population regression specification in the sense of Definition 2.1. 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 B.1 by setting $\pi_i^*(\mathbf{w}) = \mathbb{1}(\mathbf{w} = \mathbf{W}_i)$. □

Theorem B.3 (Invariance under reparameterization)**.** *Consider a population regression specification in the sense of Definition 2.1. 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 $\tilde{G}_n = M G_n M'$ (and $\widehat{\tilde{G}_n} = M \hat{G}_n M'$). Thus the potential weights for the second specification are

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

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

B.2. **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 $W_i \sim \text{Bern}(p_i)$ independently unconditionally. Correspondingly, let $\pi_{i,n}$ be the probability that $W_i = 1$ under Π_n^*.

Lemma B.4. *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* $n \to \infty$

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

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 B.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 B.5. *Under rejective sampling, suppose Assumption B.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 B.4 and Assumption B.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 $\text{Var}(\overline{z}_n(w)) \to 0$ 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. □

B.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 3.2.

Partition z into z_1 and z_2 where z_1 is J-dimensional. By Theorem B.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 B.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 greater than some $\eta > 0$, uniformly for all i.

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

Lemma B.6. *Under Assumptions 3.2, B.2, and B.3, condition (ii) in Proposition 3.5 is satisfied for some lower bound 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 3.2 is sufficient to show 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. Since $G_{n,1|2}^{-1}$ 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. 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 3.2 and $\lambda_{\max}(G_{n,1|2}) < M$ by Assumption B.3.

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

$$Z_i = U \Sigma V'$$

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

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

for an orthogonal matrix Q and diagonal D. Then

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

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,[39] 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. □

B.4. **Additional results for cross-sectional specifications.**

Proposition B.7. *Consider the cross-sectional specification $Y_i = \tau W_i + \gamma' x_i + \epsilon_i$ and a sample $(Y_i, W_i, x_i)_{i=1}^n$. Suppose x_i includes a constant. Suppose $\hat{G}_n$ is invertible on this realization. Let $\hat{\gamma}_{w \to x}$ be the OLS coefficient of W_i on x_i, and let $\hat{\pi}_i$ be the OLS fitted value $\hat{\pi}_i = \hat{\gamma}_{w \to x}' x_i$. Then:*

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

(i) *The potential weights are*

$$\hat{\rho}_i(w) = \frac{W_i - \hat{\pi}_i}{\frac{1}{n}\sum_{i=1}^n \hat{\pi}_i(1-\hat{\pi}_i)}$$

and $\frac{1}{n}\sum_{i=1}^n \hat{\pi}_i(1-\hat{\pi}_i) > 0$.

(ii) *The estimated implicit design exists and is uniquely equal to* $\hat{\pi}_i$.

Proof. (i) follows from Theorem B.2 immediately, upon verifying that

$$0 < \frac{1}{n}\sum_{i=1}^n (W_i - \hat{\pi}_i)^2 = \frac{1}{n}\sum_{i=1}^n \hat{\pi}_i(1-\hat{\pi}_i).$$

The equality follows because

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (W_i - \hat{\pi}_i)^2 &= \frac{1}{n}\sum_{i=1}^n W_i - 2W_i\hat{\pi}_i + \hat{\pi}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^n \hat{\pi}_i - 2(W_i - \hat{\pi}_i)\hat{\pi}_i - 2\hat{\pi}_i^2 + \hat{\pi}_i^2 \\
&\qquad\qquad (\tfrac{1}{n}\textstyle\sum_i \hat{\pi}_i = \tfrac{1}{n}\sum_i W_i \text{ since } x_i \text{ includes a constant}) \\
&= \frac{1}{n}\sum_{i=1}^n \hat{\pi}_i - \hat{\pi}_i^2 \quad (\tfrac{1}{n}\textstyle\sum_{i=1}^n (W_i - \hat{\pi}_i)\hat{\pi}_i = 0 \text{ by orthogonality of residuals}) \\
&= \frac{1}{n}\sum_{i=1}^n \hat{\pi}_i(1-\hat{\pi}_i).
\end{aligned}$$

For (ii), note that the implicit design is equal to

$$-\rho_i(0)/(\rho_i(1) - \rho_i(0)) = \hat{\pi}_i$$

uniquely by Corollary 2.2. □

Proposition B.8. *Consider the specification* (4.2) *where* x_i *is a saturated vector of underlying categorical covariates. That is, suppose* $c_i \in \{0, \ldots, L\}$ *and* x_i *is a vector such that* $x_{i\ell} = \mathbb{1}(c_i = \ell)$ *for* $\ell = 1, \ldots, L$. *Then the only implicit design sets* π_i, *for* $c_i = \ell$, *to be the mean* π_j^* *among those* j *with* $c_j = \ell$. *The corresponding implicit estimand is the average treatment effect.*

Proof. The specification (4.2) is equivalent to the following specification:

$$Y_i = \sum_{\ell=0}^L \alpha_\ell + \tau_\ell W_i$$

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, let us order the covariates to be

$$[\mathbb{1}(c_i = 0), \mathbb{1}(c_i = 0)W_i, \mathbb{1}(c_i = 1), \mathbb{1}(c_i = 1)W_i, \ldots, \mathbb{1}(c_i = L), \mathbb{1}(c_i = L)W_i].$$

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 $c_i = \ell$, 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

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

By Corollary 2.2, the only implicit design sets $\pi_i = \pi(\ell)$ for $c_i = \ell$. Note that the implicit design is described by $\omega_i = \pi(\ell)\rho_i(1) = 1$. Therefore the implicit estimand is the average treatment effect. □

B.5. **Additional results for panel specifications.**

Proposition B.9 (One-way fixed effects)**.** *Consider the regression $Y_{it} = \alpha_i + \tau W_{it} + \epsilon_{it}$. If $\mathcal{W}$ satisfies staggered adoption without an always treated unit, then the set of implicit designs is empty.*

Proof. Let $\dot{\mathbf{w}} = \mathbf{w} - 1_T \frac{1_T' \mathbf{w}}{T}$. Then 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. $\square$

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 (2023b), 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 (2023b) 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. In staggered adoption, however, 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 B.10. *Consider the TWFE specification $Y_{it} = \alpha_i + \gamma_t + \tau W_{it}$ under staggered adoption. 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 $\boldsymbol{\rho}_{it}(\mathbf{w}) < 0$ 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 $\mathbf{w}$. □