creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Master's Thesis of Economics

On the role of the design phase in a linear regression

선형회귀분석에서의 설계 단계의 역할에 관하여

August 2023

Graduate School of Economics
Seoul National University
Economics Major

Junho Choi

On the role of the design phase in a linear regression

Academic advisor: Seojeong Lee

Submitting a master's thesis of Economics

July 2023

Graduate School of Economics
Seoul National University
Economics Major

Junho Choi

Confirming the master's thesis written by
Junho Choi
July 2023

Chair Myung Hwan Seo (Seal)
Vice Chair Seojeong Lee (Seal)
Examiner Jungmin Lee (Seal)

On the role of the design phase in a linear regression

Junho Choi[*]

Seoul National University

econjunho@gmail.com

August 2023

Abstract

The "design phase" refers to a stage in observational studies, during which a researcher constructs a subsample that achieves a better balance in covariate distributions between the treated and untreated, in order to conduct more robust and credible inference for the parameter of interest. This article studies the role of this preliminary phase in the context of linear regression, and provides justification for its utility. To this end, we first formalize the design phase as a process of selecting a subsample, where a researcher adjusts the estimand of her regression. Then, we justify covariate balance as a valid criterion for this selection process, in that it informs for a given subsample the maximum degree of misspecification that can be allowed for the regression model, when we aim to restrict the distance between our estimand and the parameter of interest within a target level of precision. Consequently, the pursuit of a subsample with improved covariate balance is interpreted as identifying an estimand that is less susceptible to bias in the face of possible misspecification of her regression model.

Keywords: covariate balance, conditional estimand, design phase, linear regression

[*] The author acknowledges the financial support from the College of Social Sciences of Seoul National University. Earlier versions of this article were presented at the Applied Micro Brown Bag sessions (2022 Fall, 2023 Spring) in the Department of Economics at Seoul National University. This particular version was presented at the 6th International Conference on Econometrics and Statistics. Special thanks are extended to Seojeong Lee for his continuous and invaluable guidance throughout this project, and also to Ryo Okui, Jungmin Lee, and Myung Hwan Seo, whose feedback contributed to the development of this article.

Contents

1 Introduction

In their influential book, Imbens and Rubin (2015, Chapter 15) define the term "design phase," which refers to a process of constructing a subsample on which "the treatment and control samples are more balanced (in their covariate distributions) than in the original full sample" so that "within this selected subsample, inferences are most robust and credible."

This article validates these authors' claim about the design phase, specifically in the context of linear regression with binary treatment. For that, the design phase is first conceptualized as a process of subsample selection, wherein a researcher fine-tunes her estimand. Covariate balance is then established as a legitimate criterion for this selection process, in that for a given subsample, it quantifies the maximum degree of misspecification that we may allow for our regression model, when pursuing a target closeness between our estimand and the parameter of interest. As such, constructing a subsample with improved balance in covariate distributions can be understood as finding an estimand that is less prone to bias, i.e., "credible," despite possible misspecification of our regression model, i.e., "robust."

To fix ideas, let $\mathcal{S}$ be a sample given to a researcher. Suppose that she uses least squares estimator $\hat{\beta}_{\mathcal{S}}$ derived from the sample $\mathcal{S}$ to estimate the parameter of interest, which we denote by $\tau_{\mathcal{S}}$. In general, because of misspecification of her regression model, the estimand $\beta_{\mathcal{S}}$ that $\hat{\beta}_{\mathcal{S}}$ identifies is different from $\tau_{\mathcal{S}}$. Thus, valid inference for $\tau_{\mathcal{S}}$ based on $\hat{\beta}_{\mathcal{S}}$ is not feasible. Nevertheless, there may exist a "subsample" of $\mathcal{S}$, which we denote by $\mathcal{S}^*$, on which $\beta_{\mathcal{S}^*}$ and $\tau_{\mathcal{S}^*}$ are close to each other. If so, we may perform valid inference for $\tau_{\mathcal{S}^*}$ based on the least squares estimator $\hat{\beta}_{\mathcal{S}^*}$ derived from $\mathcal{S}^*$.

This is our formalization of the design phase; it is a process of selecting the subsample $\mathcal{S}^*$ whose $\beta_{\mathcal{S}^*}$ and $\tau_{\mathcal{S}^*}$ are close enough that we can conduct inference for $\tau_{\mathcal{S}^*}$ using $\hat{\beta}_{\mathcal{S}^*}$. The problem is, how one knows beforehand whether the two, which depend on the population, are close enough based on a finite sample. In the design phase, a researcher explores subsamples, compares their covariate balance, and selects the one that exhibits better covariate balance. Given our formalization, this can be justified only if covariate balance of a subsample informs us of their closeness, i.e., $\beta_{\mathcal{S}^*} - \tau_{\mathcal{S}^*}$, which we refer to as the bias of a linear regression.

In this article, we formally show that this is indeed the case. That is, covariate balance of a subsample does contain information on the closeness of the two population quantities. For that, in Section 3, we first derive a general representation of their difference, i.e., $\beta - \tau$, for arbitrary population. Specifically, we show that it is an inner product of two functions, each of which results from model misspecification of a regression and covariate imbalance. Using this result, in Section 4, we bound the bias by the multiplication of two quantities, which we denote by m and c, each of which captures the overall degree of misspecification and covariate imbalance.

These results, which also apply to finite samples, are then employed in Section 5 when formalizing the aim of assessing covariate balance of subsamples during the design phase. Since the bias of a subsample $\mathcal{S}^*$, i.e., $\beta_{\mathcal{S}^*} - \tau_{\mathcal{S}^*}$, is bounded by $m_{\mathcal{S}^*}$ times $c_{\mathcal{S}^*}$, the balance in covariate distributions, $1/c_{\mathcal{S}^*}$, which can directly be calculated from $\mathcal{S}^*$, determines the order of misspecification $m_{\mathcal{S}^*}$ that a researcher may compromise, when she aims to restrict the bias within a desired range; for instance, below a pre-specified tolerance $\epsilon > 0$. That is, if she wants to be confident of $\tau_{\mathcal{S}^*}$ at the precision of ϵ, $m_{\mathcal{S}^*}$ can be allowed for up to $\epsilon/c_{\mathcal{S}^*}$. In this regard, constructing a subsample with improved covariate balance is equivalent to searching for an estimand, with which a researcher can be assured of less bias resulting from misspecification of her regression model.

1.1 Related literature

This article is related to the conditional estimand literature, given that $\beta_{\mathcal{S}^*}$ is the conditional linear projection of the outcome given the in-subsample empirical distribution of the treatment and covariates. We extend the full-sample results of Abadie et al. (2014) to arbitrary subsamples in Section 5, where we establish the asymptotic properties of $\hat{\beta}_{\mathcal{S}^*} - \beta_{\mathcal{S}^*}$. In fact, our formalization of the design phase can provide another motivation for conditional estimands. They allow a researcher to assess their population properties based on finite sample statistics.

The importance of covariate balance has been extensively emphasized in the literature. (Crump et al., 2009; Imbens, 2015; Imbens and Rubin, 2015; Abadie and Spiess, 2022, among others.) The literature has proposed various schemes for improving covariate balance, that is, systematic procedures for selecting $\mathcal{S}^*$; for example, matching or trimming. In this article, we do not restrict ourselves to a specific balancing method. Our results do not restrict the functional form of $\mathcal{S}^*$. The focus of this article is rather on formalizing and justifying the use of a general balancing process in the design phase, especially when we use linear regression.

2 Framework

We observe a random sample $\mathcal{S} \equiv \{1, \ldots, N\}$ consisting of N units. Each unit i is characterized by a scalar outcome $Y_i \in \mathbb{R}$, a binary treatment $D_i \in \{0, 1\}$, and p-dimensional control variables $X_i \in \mathbb{R}^p$. The parameter of our interest is

$$\tau \equiv \mathbf{E}[Y|D = 1] - \mathbf{E}[\mathbf{E}[Y|X, D = 0]|D = 1], \tag{1}$$

where we omit the subscript i in the variables for the sake of conciseness. τ simplifies into the average treatment effect on the treated, once we assume the conditional independence between $Y(0)$ and D given X, where $Y(0)$ denotes the potential outcome when untreated.

A popular estimation method for τ is matching (Abadie and Imbens, 2012), wherein the outcome values of treated units are compared to the averaged ones of the untreated units with similar values of control variables. In this article, however, our discussion is restricted to linear regression. Our focus is on understanding the mechanism whereby the design phase improves the ability of a linear regression to estimate τ.

Let $\mathcal{X}^d \subseteq \mathbb{R}^p$ denote the support of the conditional distribution $\mathcal{L}_{X|D=d}$ of X given $D = d$, which we denote by G^d. Assume that each G^d is dominated by a common measure μ, which can be either a counting measure or a Lebesgue measure. Then, we can rewrite τ as

$$\int \left(\mathbf{E}[Y|X=x,D=1]-\mathbf{E}[Y|X=x,D=0]\right)\mathbf{Pr}[X=x|D=1]\mu(dx), \tag{2}$$

where $\mathbf{Pr}[X=x|D=d] \equiv (dG^d/d\mu)(\mathrm{x})$ denotes the Radon-Nikodym derivative of G^d with respect to μ. This representation is immediate but yields a pivotal observation. Note that, in the current setting, (i) the joint distribution between X and D, which we denote by G, and (ii) the conditional distribution $\mathcal{L}_{Y|X,D}$ of Y given (X,D), which we denote by F, provide a complete description of the population. Thus, equation (2) reveals how each component of the population (G,F) interacts with τ. In this regard, henceforth, whenever there is a need to explicitly indicate the dependence on G or F, we incorporate a relevant subscript to signify that specific relationship; for example, $\tau_{G,F}$, $\mathbf{Pr}_G$, or $\mathbf{E}_F$.

Now, consider a linear regression model:

$$Y_i = \alpha + \beta D_i + s(X_i)'\gamma + E_i, \tag{3}$$

where E_i denotes a regression error and $s(\cdot)$ is a function that maps X_i to a vector, resulting in a vector $s(X_i)$ of covariates.

The starting point of this article is to establish a general relationship between β and τ. Specifically, we derive a useful representation of their difference, $\beta-\tau$, whose absolute value is called the "bias" of a linear regression. In the appendix, we explore an extended regression model that incorporates interaction terms between D_i and $s(X_i)$.

Example 1 To provide a concrete illustration of our discussion, we consider a toy example in which Y_i is generated as

$$Y_i = D_i + D_iX_i + X_i + U_i, \tag{4}$$

where D_i and X_i are binary variables such that $\mathbf{Pr}[D_i = d, X_i = x] = (1/2 - p)\mathbf{1}\{d = x\} + p\mathbf{1}\{d \neq x\}$ for some $p \in (0, 1/2)$ and U_i is any random variable that has a finite first moment and is independent from (D_i, X_i). Since X_i is binary conditional on D_i, $\mu = \delta_0 + \delta_1$ is a counting measure, where δ_x denotes a Dirac measure defined by $\delta_x(B) \equiv \mathbf{1}\{x \in B\}$ for any borel set B, and G^1 and G^0 are both Bernoulli distributions, where $\mathbf{Pr}[X = 1|D = 1] = (1/2 - p)/(1/2)$ and $\mathbf{Pr}[X = 1|D = 0] = p/(1/2)$. p is a parameter that represents the dependence between D_i and X_i; the more it deviates from $1/4$, the more dependent the two variables become. Note also that, in this setting, $\tau = \mathbf{E}[Y|D = 1] - \mathbf{E}[\mathbf{E}[Y|X, D = 0]|D = 1] = (1 + 2(1 - 2p)) - (1 - 2p) = 2 - 2p$.

We assume that the two following specifications are considered for the regression model:

Specification A: $Y_i = \alpha_A + \beta_A D_i + E_i$ and

Specification B: $Y_i = \alpha_B + \beta_B D_i + \gamma_B X_i + E_i$.

Both of the models are "misspecified" in the sense that their functional forms are different from that of the conditional expectation, which additionally includes the interaction term between D_i and X_i. We will utilize this setting to illustrate the implication of our results.

3 Representation of the bias of a linear regression

In this section, we provide a novel representation of the bias of a linear regression, which will be used in formalizing the role of covariate balance in Section 4. Our result shows that the bias of a linear regression is an inner product of two functions, each of which results from the deviation from the ideal scenarios where the regression model is correctly specified or the treatment is randomly assigned.

To begin, we anchor the interpretation of β by imposing a set of regularity conditions.

Assumption 1. *Let $Z \equiv (1, D, s(X)')'$ denote a vector of regressors in model* (3). *Then, $\mathbf{E}[\|Z\|^2]$ and $\mathbf{E}[\|ZE\|]$ are finite, $\mathbf{E}[ZZ']$ is positive definite, and* $\mathbf{E}[ZE] = 0$.

Let $\theta \equiv (\alpha, \beta, \gamma)$ be a vector of the coefficients of regression model (3). Under Assumption

1, θ is identified by $(\mathbf{E}[ZZ'])^{-1}\,\mathbf{E}[ZY]$, which is a vector of linear projection coefficients in cases where the second moment of Y exists. In particular,

$$\beta = \frac{\mathbf{E}[\tilde{D}Y]}{\mathbf{E}[\tilde{D}^2]}, \tag{5}$$

where $\tilde{D}$ denotes the population residual from a linear projection of D on $(1, s(X)')'$.

We tighten our previous notations. Let $Y = f(X, D) + U$, where $\mathbf{E}[U|X, D] = 0$. That is, $f(x, d)$ denotes the conditional expectation of Y given $(X, D) = (x, d)$. Denote by $g^d(x) \equiv \mathbf{Pr}[X = x|D = d]$ the conditional density of X given $D = d$. Let $l(x, d) \equiv \alpha + \beta d + s(x)'\gamma$ be the population regression function.

We now state our result:

Proposition 1. *Suppose that Assumption 1 holds. Then,*

$$\beta - \tau = \int \underbrace{(f(x,0) - l(x,0))}_{\textit{model misspecification}} \overbrace{(g^1(x) - g^0(x))}^{\textit{covariate imbalance}} \mu(dx). \tag{6}$$

Proposition 1 shows that the bias of a linear regression is an inner product of two basic functions. The first one is $f(\cdot, 0) - l(\cdot, 0)$, which results from model (3) being misspecified. If it were correctly specified, zero becomes a version of $f(X, D) - l(X, D)$, and thus (6) implies $\beta = \tau$. The second one is $g^1(\cdot) - g^0(\cdot)$ right next to it. Its value depends on the degree of overlap in covariate distributions between the treated and untreated, i.e., G^1 and G^0. When using observational data, this term will generally be non-zero. If D and X were independent, $g^1 - g^0 = 0$ μ-almost surely, and thus $\beta = \tau$.[1]

Example 1 (cont') We illustrate how equation (6) operates in our toy example. Let l_A and l_B denote the population regression functions for each specification. Then, it can be shown that $l_A(x, d) = 2p + (3 - 6p)d$ and $l_B(x, d) = -p + (3/2)x + (3/2)d$. Hence, for

[1] This particular result has been shown by Imbens and Rubin (2015, Chapter 7).

Specification A, equation (6) holds in the form of

$$
\begin{aligned}
1-4p=\beta-\tau=&\sum_{x\in\{0,1\}}(f(x,0)-l_A(x,0))(g^1(x)-g^0(x))\\
&=(1-2p)((1-2p)-2p)+(0-2p)(2p-(1-2p))=1-4p,
\end{aligned}
$$

and for Specification B, it takes the form of

$$
\begin{aligned}
-1/2+2p=\beta-\tau=&\sum_{x\in\{0,1\}}(f(x,0)-l_A(x,0))(g^1(x)-g^0(x))\\
&=(1-(-p+3/2))((1-2p)-2p)+(0-(-p))(2p-(1-2p))=-1/2+2p.
\end{aligned}
$$

Let F^d denote the conditional distribution $\mathcal{L}_{Y|X,D=d}$ of Y given X and $D=d$. We refer to the "support" of a conditional distribution as the support of the conditioning variable. For instance, the support of F^0 is $\mathcal{X}^0$.

Assumption 2. *$\mathcal{X}^1$ is contained in the support of F^0.*

Assumption 2 is not strictly required for our discussion. However, it facilitates a natural interpretation of τ.

4 Role of covariate balance in a linear regression

In this section, we formalize the role of covariate balance in regression, which serves as a basis for the formalization of that of the design phase in Section 5. Before starting, we note that the term "covariate balance" here denotes the population dependence between X and D, which is not a common usage in the literature. Nevertheless, given that our results for the population balance will be employed when justifying the use of the finite-sample balance during the design phase, we abuse the language for a moment.

In the first subsection, we justify the conventional notion that better covariate balance makes regression more robust to model misspecification. For that, we show that the bias of a regression is bounded by the product of model misspecification and covariate imbalance. Suppose that a researcher aims to know τ by β within a desired tolerance. Then, this result implies that the extent of misspecification to which she may compromise is proportional to the inverse of imbalance, that is, the balance in covariate distributions, which accords with our previous notion. In the second subsection, we use this bound analysis in demonstrating how inference is robustified against misspecification with better covariate balance.

4.1 Bounds for the bias of a linear regression

Proposition 1 yields useful bounds for the bias of a linear regression, all of which possess the structure of the multiplication between two quantities, m and c, which respectively capture the degree of misspecification of the regression model and covariate imbalance, that is,

$$|\beta - \tau| \leq mc. \tag{7}$$

This common structure of the bounds implies that misspecification of a regression model can be allowed for on the order of $1/c$; if a researcher wants to be confident of her β for τ at the precision of ϵ, she may not worry much about misspecification of her regression model up to ϵ/c. In the extreme case where $c = 0$, the specification of a regression model becomes irrelevant to the identification of τ.[2]

Another crucial observation from equation (7) is that c depends solely on the joint distribution of X and D, i.e., G, which plays an important role in formalizing the role of the design phase in Section 5.

In this subsection, we provide three forms of equation (7) that differ in how we define misspecification and covariate imbalance, i.e., m and c. They provide different descriptions

[2] We note that even when $c = 0$, the specification of a regression model can be important in terms of the "estimation". In some cases, we can attain higher efficiency. See Freedman (2008a,b), Lin (2013), and Negi and Wooldridge (2021) for relevant discussion.

on the robustness of a regression.

4.1.1 Total variation bound

The bound of the first form is called the "total variation bound," since the total variance distance is used as a measure of covariate imbalance. Define

$$m^{\mathrm{TV}} \equiv \|f(x,0) - l(x,0)\|_{L^\infty(\mu)} \text{ and } \tag{8}$$

$$c^{\mathrm{TV}} \equiv \int |g^1(x) - g^0(x)|\mu(dx), \tag{9}$$

where m^{TV} is the essential supremum of $f(\cdot,0) - l(\cdot,0)$ with respect to μ, and c^{TV} is the total variation distance between G^1 and G^0. $m^{\mathrm{TV}} = 0$ if and only if $f(\cdot,0) = l(\cdot,0)$ μ-almost surely, and $c^{\mathrm{TV}} = 0$ if and only if D and X are independent.

Corollary 1. *Suppose that the conditions of Proposition 1 are satisfied. Then,*

$$|\beta - \tau| \leq m^{\mathrm{TV}} c^{\mathrm{TV}}. \tag{10}$$

Example 1 (Cont') We illustrate the result of Corollary 1 using our toy example. Here, we calculate m^{TV} for each specification and show the specific form of equation (10) for each case. First, note that $c^{\mathrm{TV}} = \int_{\mathcal{X}^0} |g^1(x) - g^0(x)|\mu(dx) = |g^1(1) - g^0(1)| + |g^1(0) - g^0(0)| = 2|1-4p|$. In terms of Specification A, $m_A^{\mathrm{TV}} = \|f(\cdot,0) - l_A(\cdot,0)\|_{L^\infty(\mu)} = \|\cdot - 2p\|_{L^\infty(\mu)} = 2p \vee (1-2p)$, and thus equation (10) holds in the form of

$$|1-4p| = |\beta_A - \tau| \leq m_A^{\mathrm{TV}} c^{\mathrm{TV}} = \underbrace{(2p \vee (1-2p))}_{\geq 1/2} \times 2|1-4p|,$$

where the bounds are tight at $p = 1/4$. For Specification B, $m_B^{\mathrm{TV}} = \|f(\cdot,0) - l_B(\cdot,0)\|_{L^\infty(\mu)} = \|\cdot - (-p + (3/2)\cdot)\|_{L^\infty(\mu)} = p \vee (1/2 - p)$, and thus equation (10) takes the form of

$$(1/2)|1-4p| = |\beta_B - \tau| \leq m_B^{\mathrm{TV}} c^{\mathrm{TV}} = \underbrace{(p \vee (1/2-p))}_{\geq 1/4} \times 2|1-4p|.$$

4.1.2 Kolmogorov-Smirnov bound

The bound of the second form is called the "Kolmogorov-Smirnov bound," since it uses the Kolmogorov-Smirnov distance to assess the degree of covariate imbalance. To streamline the discussion, here we assume that X is one-dimensional, that is, $p = 1$.

Assumption 3. $\mathbf{E}[Y|X = x, D = 0]$ *is continuous and bounded on the support of* F^0*.*

Continuity and boundedness are required for our result. Let $\mathcal{H}$ be the class of all continuous and bounded extensions h on $\mathbb{R}$ such that $h(\cdot) = f(\cdot,0) - l(\cdot,0)$ on $\mathcal{X}^0$, which is non-empty by Assumption 3. Then, define

$$m^{\mathrm{KS}} \equiv \inf_{h \in \mathcal{H}} V_{-\infty}^{\infty}(h) \text{ and} \tag{11}$$

$$c^{\mathrm{KS}} \equiv \sup_{x \in \mathbb{R}} |G^1(x) - G^0(x)|, \tag{12}$$

where $V_{-\infty}^{\infty}(\cdot)$ is the total variation of the argument on $\mathbb{R}$. m^{KS} captures the total variation of $f(\cdot,0) - l(\cdot,0)$ on $\mathcal{X}^0$, and c^{KS} is the Kolmogorov-Smirnov distance between G^1 and G^0.

Corollary 2. *Suppose that the conditions of Proposition 1 are satisfied. Also, suppose that Assumption 2–3 holds. Then,*

$$|\beta - \tau| \leq m^{\mathrm{KS}} c^{\mathrm{KS}}. \tag{13}$$

Example 1 (Cont') We show how equation (13) works in our toy example. Since $\mathcal{X}^0 = \{0,1\}$, $c^{\mathrm{KS}} = |G^1(0) - G^0(0)| = |1-4p|$. For Specification A, $m_A^{\mathrm{KS}} = |(f(1,0) - l_A(1,0)) - (f(0,0) - l_A(0,0))| = |(1-2p) - (0-2p)| = 1$, and similarly for Specification B, $m_B^{\mathrm{KS}} =$

$|(1-(-p+3/2))-(0-(-p))|=1/2$. Then, equation (13) becomes equality in the form of

$$|1-4p| = |\beta_A - \tau| = m_A^{\mathrm{KS}} c^{\mathrm{KS}} = 1 \times |1-4p| \text{ and}$$
$$(1/2)|1-4p| = |\beta_B - \tau| = m_B^{\mathrm{KS}} c^{\mathrm{KS}} = (1/2) \times |1-4p|.$$

We note that the equality is specific to this particular example.

4.1.3 Density ratio bound

m^{TV} and m^{KS} may be overly stringent measures for assessing misspecification, as they basically pick up only the most extreme discrepancy between $f(\cdot,0)$ and $l(\cdot,0)$ irrespective of how likely it can occur. For instance, suppose that p is close to zero, and consider the case of Specification A in our example. Then, since $G^0(\{0\})$ is close to one, it could be argued that $l_A(x,0)=2p$ is close enough to $f(x,0)=x$, that is, the regression model of Specification A is "almost" correctly specified. Nevertheless, this aspect cannot be reflected in m_A^{TV}, for example, given that $m_A^{\mathrm{TV}} \geq 1/2$. The bound now we provide uses an alternative measure for misspecification, which circumvents this issue.

The bound of the third form is called the "density ratio bound," as it uses the moments of g^1/g^0 to assess the overlap in covariate distributions. Define

$$\begin{aligned} m^{\mathrm{DR}} &\equiv \|f(x,0)-l(x,0)\|_{L^2(G^0)} \text{ and} \\ c^{\mathrm{DR}} &\equiv \|g^1(x)/g^0(x)-1\|_{L^2(G^0)}. \end{aligned} \tag{14}$$

$m^{\mathrm{DR}}=0$ if the regression model is correctly specified, but the reverse is not necessarily true. m^{DR} is a more flexible measure of misspecification compared to m^{TV}. In the previous setting where p is close to zero, it can be shown that $m_A^{\mathrm{DR}} = \mathbf{E}[(X-2p)^2|D=0]^{1/2} = (2p(1-2p))^{1/2}$ is close to zero, according with our intuition that $l_A(x,0)=2p$ and $f(x,0)=x$ are virtually the same. $c^{\mathrm{DR}}=0$ if and only if X and D are independent. m^{DR} and c^{DR} are finite in cases where Y has a second moment and the propensity scores are bounded away from one.

Corollary 3. *Suppose that the conditions of Proposition 1 are satisfied. Then,*

$$|\beta - \tau| \leq m^{\mathrm{DR}} c^{\mathrm{DR}}. \tag{15}$$

Example 1 (Cont') We revisit the previous example to illustrate the result of Corollary 3. First, note that $c^{\mathrm{DR}} = |1-4p|/(2p(1-2p))^{1/2}$, which equals zero if and only if $p = 1/4$. For Specification A and B, equation (15) holds as an equality in the form of

$$|1-4p| = |\beta_A - \tau| = m_A^{\mathrm{DR}} c^{\mathrm{DR}} = (2p(1-2p))^{1/2} \times |1-4p|/(2p(1-2p))^{1/2} \text{ and}$$
$$(1/2)|1-4p| = |\beta_B - \tau| = m_B^{\mathrm{DR}} c^{\mathrm{DR}} = (p(1/2-p))^{1/2} \times |1-4p|/(2p(1-2p))^{1/2}.$$

The equality here is also specific to our example setup.

4.1.4 Comparison of the bounds

The three forms of bound do not subsume one another in the sense that the orders of m's and c's can be reversed, that is, $m^{\mathrm{DR}} < m^{\mathrm{TV}} < m^{\mathrm{KS}}$ and $c^{\mathrm{KS}} < c^{\mathrm{TV}} < c^{\mathrm{DR}}$. There is no general order among the products of m's and c's. In our example, $m_A^{\mathrm{KS}} c^{\mathrm{KS}} = |1-4p| = m^{\mathrm{DR}} c^{\mathrm{DR}}$, while $m_A^{\mathrm{TV}} c^{\mathrm{TV}} \geq |1-4p|$.

4.2 Robust inference to misspecification

The bound analyses in the previous subsection can be used for defining the role of covariate balance when conducting inference with regression. We show that covariate balance extends the range of possible misspecification under which a researcher can maintain her statistical decision. Here, we focus on the t-test.

Suppose that we perform inference for τ using a t-statistic

$$t \equiv \frac{\hat{\beta} - \tau_0}{se_\beta(\hat{\beta})}, \tag{16}$$

where $\hat{\beta}$ is an estimator for β, $se_\beta(\hat{\beta})$ denotes the standard error of $\hat{\beta}$, and τ_0 is the value of τ set at the null. The subscript β on *se* indicates that the formula for the standard error of $\hat{\beta}$ can vary depending on the estimand, i.e., β.

In general, β is different from τ due to misspecification, and thus the size of the t-test converges to one as the sample size increases. Hence, to have a more meaningful discussion about the size distortion of a t-test, we employ local asymptotics.

Assumption 4. *$mc = v/\sqrt{N}$ for some $v > 0$.*

The constant v captures the overall deviation from the ideal scenario that is free of either model misspecification or covariate imbalance.

The assumption below is a high-level condition, which ensures the existence of the limiting distribution of $\hat{\beta}$.

Assumption 5. *$(\hat{\beta} - \beta)/se_\beta(\hat{\beta})$ converges in distribution to the standard normal. Also, the probability limit σ_β of $\sqrt{N} se_\beta(\hat{\beta})$ exists.*

We then have the following result:

Corollary 4. *Suppose that the conditions of either Corollary 1, 2, or 3 are satisfied. Also, suppose that Assumptions 4–5 hold. Then, under the null $H_0 : \tau = \tau_0$,*

$$\Phi(z - v/\sigma_\beta) \leq \liminf_{N\to\infty} \mathbf{Pr}[t \leq z] \leq \limsup_{N\to\infty} \mathbf{Pr}[t \leq z] \leq \Phi(z + v/\sigma_\beta), \tag{17}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal and $z \in \mathbb{R}$.

The asymptotic size distortion of a t-test is a decreasing function of v, and thus mc. This implies that to make a valid inference for τ, it is not necessary for both m and c to be small. For instance, when analyzing experimental data, misspecification could be allowed for to a great extent, considering that c would likely be very small.

5 Role of the design phase in a linear regression

The role of covariate balance studied in the previous section has been based on the "population." In practice, however, it is infeasible to select or manipulate the population to possess better covariate balance. Instead, researchers typically construct a subsample whose empirical distribution mimics the ideal situation where X and D are independent; they select the one with better overlap in covariate distributions between the treated and untreated units. Imbens and Rubin (2015, Section 12.5) specifically refer to this prior stage as the "design phase," a concept encompassing any groundwork to construct a subsample that is "more suitable for estimating causal estimands, in the sense of being better balanced in terms of covariate distributions."

In this section, we provide a formal justification for this viewpoint, using our results in Section 4. Our justification unfolds in two steps, each corresponding to a separate subsection. In the first subsection, we characterize the design phase as a process of subsample selection, through which a researcher adjusts the estimand of her regression. In the second subsection, we then establish the validity of covariate balance as a criterion for the selection process. Specifically, we show that for a given sample, covariate balance informs the upper bound of misspecification that a researcher using the subsample can compromise, when she wants her estimand β to be close to τ within a target tolerance. In the last subsection, we address how we conduct inference for the adjusted estimand.

5.1 Estimand adjustment

In this subsection, we demonstrate that the design phase is basically a process of adjusting the estimand of a regression, i.e., β, via selecting a subsample. Our result applies to any methods for constructing subsamples, such as matching or trimming, as long as they implement a prespecified rule where the units are selected based on $\mathbb{X} \equiv (X_i)_{i=1}^N$ and $\mathbb{D} \equiv (D_i)_{i=1}^N$.

Let $\mathcal{S}^* \subseteq \mathcal{S}$ be a constructed subsample in the design phase. Let $\hat{\theta}^* \equiv (\hat{\alpha}^*, \hat{\beta}^*, \hat{\gamma}^*)$ be the least squares estimator

$$\hat{\theta}^* \equiv \left(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i Z_i'\right)^{-1}\left(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i Y_i\right) \tag{18}$$

obtained by regressing Y_i on Z_i in the subsample $\mathcal{S}^*$, where $|\mathcal{S}^*|$ denotes its cardinality.

We adopt a more compact notation. Let $(\tilde{G}, \tilde{F})$ be a population, and define

$$\theta_{\tilde{G},\tilde{F}} \equiv (\mathbf{E}_{\tilde{G}}[ZZ'])^{-1}\,\mathbf{E}_{\tilde{G}}[Z\,\mathbf{E}_{\tilde{F}}[Y|X,D]], \tag{19}$$

where $\theta_{\tilde{G},\tilde{F}} \equiv (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ equals the linear projection coefficient obtained by regressing Y on Z in that population. Let $\mathbb{G}^*$ and $\mathbb{F}^*$ denote the empirical joint distribution of (X, D) and the empirical conditional distribution of Y given (X, D), respectively, in the subsample $\mathcal{S}^*$. Then, for instance, we can replace $\hat{\theta}^*$ with $\theta_{\mathbb{G}^*,\mathbb{F}^*}$, where the dependence on each component of the relevant population becomes more apparent.

Now we make the following assumptions.

Assumption 6. *$\mathcal{S}^*$ is a function of $\mathbb{X}$ and $\mathbb{D}$.*

Assumption 7. *Let $\lambda_{\min}(M)$ denote the smallest eigenvalue of a square matrix M. There exists a random variable λ^* such that*

$$\lambda_{\min}(\mathbf{E}_{\mathbb{G}^*}[ZZ']) = \lambda_{\min}\left(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i Z_i'\right) \geq \lambda^*, \tag{20}$$

where $\lambda^* \to_p \underline{\lambda}$ for some positive constant $\underline{\lambda} > 0$ as $|\mathcal{S}^*|$ tends to infinity.

Assumption 8. *For $d \in \{0, 1\}$, $\mathbf{E}_F[\|ZU\|^2|X = x, D = d]$ is uniformly bounded on $\mathcal{X}^d$.*

Proposition 2. *Under Assumptions 6–8, as $|\mathcal{S}^*|$ tends to infinity[3],*

$$\hat{\theta}^* - \theta_{\mathbb{G}^*,F} \to_p 0. \tag{21}$$

[3] Formally, this can be stated as "if $\operatorname{plim}_{N\to\infty} |S^*(\mathbb{X}, \mathbb{D})| = \infty$, as N tends to infinity."

Proposition 2 is related to Theorem 1 of Abadie et al. (2014), where the authors show equation (21) in the case where $\mathcal{S}^* = \mathcal{S}$.[4] It extends their full-sample result to arbitrary subsamples, which provides us with a useful interpretation of subsample construction. Based on this result, selecting another subsample with different $\mathbb{G}^*$ can be considered as adjusting the estimand, i.e., $\beta_{\mathbb{G}^*,F}$, of a regression, possibly expecting the adjusted one to be much closer to the parameter of our interest, i.e., $\tau_{\mathbb{G}^*,F}$, which may be given a causal interpretation.

5.2 Estimand assessment via covariate balance

In the design phase, researchers compare subsamples based on their balances in covariate distributions. That is, covariate balance is utilized as a criterion for selecting the subsample on which they run regressions. Given our previous formalization of the design phase, this use of covariate balance is justifiable if it reveals desirable attributes of the adjusted estimand. In this subsection, we demonstrate that covariate balance informs the robustness of an estimand to possible misspecification of the regression model.

Suppose that $|\mathcal{S}^*|^{-1} \sum_{i \in \mathcal{S}^*} Z_i Z_i$ is positive definite. Then, if the conditions of Corollary 1–3 are satisfied for the population (G, F), they are satisfied almost surely for the population $(\mathbb{G}^*, F)$ as well, and thus

$$|\beta_{\mathbb{G}^*,F} - \tau_{\mathbb{G}^*,F}| \leq m_{\mathbb{G}^*,F} c_{\mathbb{G}^*} \text{ holds almost surely,} \tag{22}$$

where the right-hand side could be either the Kolmogorov-Smirnov bound, total variation bound, or density ratio bound. The crucial part of equation (22) is that $c_{\mathbb{G}^*}$ depends solely on $\mathbb{G}^*$ and can be directly calculated from a subsample. In other words, covariate balance is a feasible criterion a researcher can employ when assessing the estimand $\beta_{\mathbb{G}^*,F}$ of a subsample. Indeed, $1/c_{\mathbb{G}^*}$ is the order of the maximum misspecification up to which a researcher can compromise when aiming to control the bias within a pre-specified tolerance.

[4] However, it should be noted that their result was not confined to linear regression.

Example 2 We illustrate our point using a simple setup. Let $Y_i = X_i + U_i$, where X_i is conditionally normal given D such that $\mathcal{L}_{X|D=0} = \mathcal{N}(0,1)$ and $\mathcal{L}_{X|D=1} = \mathcal{N}(1,1)$, and U_i is any random variable that has a finite first moment and is independent of D_i and X_i. Suppose that we are given a sample $\mathcal{S}$ with 24 observations, where the half of them, T1, . . . , T12, are treated and the other half, U1, . . . , U12, are not. Let the empirical conditional distribution, $\mathbb{G}^d$, of X given $D = d$ be given as in Table 1.

	U1	U2	T1	T2	T3	T4	U3	U4	U5	U6	U7	U8
$\mathbb{G}^1$			-0.95	-0.90	-0.29	-0.23						
$\mathbb{G}^0$	-1.26	-1.00					-0.07	0.21	0.31	0.35	0.39	0.39

	T5	T6	U9	U10	T7	T8	U11	T9	T10	T11	T12	U12
$\mathbb{G}^1$	0.48	0.59			0.87	1.46		1.78	2.03	2.07	2.31	
$\mathbb{G}^0$			0.60	0.79			1.48					2.89

Table 1: Empirical conditional distributions, $\mathcal{L}_{X|D=d}$.

Suppose that we run the regression $Y_i = \alpha + \beta D_i + E_i$, where the least squares estimator is then the difference in means of Y_i between the treated and untreated.

Suppose that we use the full sample to estimate τ, which in this setting is zero. Since $\mathbf{E}[Y_i|X_i, D_i] = X_i$, we can calculate from the table that the full sample estimand is $\beta_{\mathbb{G},F} = 0.34$ and thus the bias is $|\beta_{\mathbb{G},F} - \tau_{\mathbb{G},F}| = 0.34$. The Kolmogorov-Smirnov distance between $\mathbb{G}^1$ and $\mathbb{G}^0$ is $c^{\mathrm{KS}}_{\mathbb{G}} = 0.3$, which shows that if we want to be confident of $\tau_{\mathbb{G},F}$ at the precision of 0.5, for example, the degree of misspecification, i.e., $m^{\mathrm{KS}}_{\mathbb{G}^*,F}$, should be less than 5/3,[5] which may be too stringent considering the variation of X_i in this sample.

Now, suppose we construct a subsample $\mathcal{S}^*$ consisting of U2 and T1, T4 and U3, T6 and U9, U10 and T7, and T8 and U11. Then, $\beta_{\mathbb{G}^*,F} = -0.01$, and the bias is $|\beta_{\mathbb{G}^*,F} - \tau_{\mathbb{G}^*,F}| = 0.01$. Since $c^{\mathrm{KS}}_{\mathbb{G}^*} = 0.2$, for the same precision, the upper limit of misspecification, i.e., $m^{\mathrm{KS}}_{\mathbb{G}^*}$, now becomes 5/2, which is larger than 5/3 before.

[5] Indeed, in this case, since $f_F(x,0) - l_{\mathbb{G},F}(x,0) = x - \alpha_{\mathbb{G},F} = x - 0.676$, $m^{\mathrm{KS}} = |-4.51 - 0.676| + |3.98 - 0.676|$, which is definitely bigger than 2.

In practice, however, researchers usually use summary statistics when assessing balance of subsamples; for instance, the difference in the covariates means (Imbens and Rubin, 2015, Chapter 14). In our framework, for the formal justification of this convention, those statistics should has to do with $c_{\mathbb{G}^*}$ of the subsample. We provide the assumption for this requirement.

Assumption 9. *There exists a signed measure J^* on $\mathbb{R}^p$ that depends on $\iota(\mathbb{G}^{*1}, \mathbb{G}^{*0})$, where ι is a functional that maps $(\mathbb{G}^{*1}, \mathbb{G}^{*0})$, a pair of the empirical conditional distributions in a subsample $\mathcal{S}^*$, to a vector, such that as $|\mathcal{S}^*|$ tends to infinity,*

$$\int (f(x,0) - l_{\mathbb{G}^*,F}(x,0))(\mathbb{G}^{*1} - \mathbb{G}^{*0})(dx) - \int (f(x,0) - l_{\mathbb{G}^*,F}(x,0))J^*(dx) \to_p 0. \tag{23}$$

Example 2 (Cont') The averages of X_i among the treated and untreated units in the subsample $\mathcal{S}^*$ are 0.35 and 0.36, respectively. We may consider $J^*(dx) = (\delta_{0.35}(dx) - \delta_{0.36}(dx))$, which is a signed measure that depends on the first moments of $\mathbb{G}^{*1}$ and $\mathbb{G}^{*0}$. Then, since $f(x,0) = x$ and $l_{\mathbb{G}^*,F}(x,0) = \alpha_{\mathbb{G}^*,F}$, for each $d \in \{0,1\}$, $\int (f(x,0) - l_{\mathbb{G}^*,F}(x,0))\, \mathbb{G}^{*d}(dx)$ equals

$$\mathbf{E}_{\mathbb{G}^{*d}}[X] - \alpha_{\mathbb{G}^*,F} = f(\mathbf{E}_{\mathbb{G}^{*d}}[X], 0) - l_{\mathbb{G}^*,F}(\mathbf{E}_{\mathbb{G}^{*d}}[X], 0) = \int (f(x,0) - l_{\mathbb{G}^*,F}(x,0))\delta_{\mathbf{E}_{\mathbb{G}^{*d}}[X]}(dx).$$

Thus, the left-hand side of equation (23) is zero, and

$$\beta_{\mathbb{G}^*,F} - \tau_{\mathbb{G}^*,F} = \int (f(x,0) - l_{\mathbb{G}^*,F}(x,0))(\delta_{\mathbf{E}_{\mathbb{G}^{*1}}[X]} - \delta_{\mathbf{E}_{\mathbb{G}^{*0}}[X]})(dx) \text{ holds.}$$

$c^{\mathrm{KS}}_{\mathbb{G}^*,F}$ based on $\delta_{\mathbf{E}_{\mathbb{G}^{*1}}[X]} - \delta_{\mathbf{E}_{\mathbb{G}^{*0}}[X]}$ is $|0.35 - 0.36| = 0.01$. This is smaller than that based on $\mathbb{G}^{*1} - \mathbb{G}^{*0}$, which was 0.2. Thus, for the precision of 0.5, the degree of misspecification, i.e., $m^{\mathrm{KS}}_{\mathbb{G},F}$, could have been in fact up to 0.5/0.01, which is the case where we may not be much concerned about the specification of our regression model.

This example demonstrates the case in which using summary statistics, here, the means, is not only valid, but also improves the maximum misspecification that we can allow for our model. In fact, summary statistics can be useful especially when covariates are continuous.

Since the supports of $\mathbb{G}^{1*}$ and $\mathbb{G}^{0*}$ do not intersect with each other in a finite sample, $c^{\mathrm{TV}}_{\mathbb{G}^*}$ and $c^{\mathrm{DR}}_{\mathbb{G}^*}$ almost surely equals one, which is uninformative. Here, as alternatives, we may use $c^{\mathrm{KS}}_{\mathbb{G}^*}$ or coarsen X to make $c^{\mathrm{TV}}_{\mathbb{G}^*}$ and $c^{\mathrm{DR}}_{\mathbb{G}^*}$ usable. However, $m^{\mathrm{KS}}_{\mathbb{G}^*,F}$ could overstate misspecification, and coarsening could affect the interpretation of τ. If a researcher is concerned about these issues while dealing with continuous X, comparing summary statistics between the treated and untreated could be better.

One problem of using them could be that we in general do not know the form of J^* and thus it is difficult to quantify the exact maximum misspecification allowed for us.

5.3 Inference for the adjusted estimand

In the previous subsections, the construction of a subsample in the design phase is characterized as an estimand adjustment through subsample selection. We further justified the use of covariate balance during this selection process, demonstrating that it gauges the robustness of an estimand to possible misspecification. However, an estimand is still a population quantity that is observed with sampling error. Hence, to make use of the adjusted estimand, we need to quantify the associated uncertainty as well.

In this subsection, we establish the asymptotic normality of $\hat{\theta}^* - \theta_{\mathbb{G}^*,F}$, and propose an estimator for its asymptotic variance.

We make the following assumptions.

Assumption 10. *$\mathbf{E}_{\mathbb{G}^*}[ZZ'] = |\mathcal{S}^*|^{-1}\sum_{i\in\mathcal{S}^*} Z_iZ_i' \to_p \Gamma^*$, as $|\mathcal{S}^*|$ tends to infinity, where Γ^* is a positive definite matrix.*

Assumption 11. *For $d \in \{0,1\}$, $\mathbf{E}_F[\|ZU\|^{2+\delta}|X = x, D = d]$ is uniformly bounded on $\mathcal{X}^d$, where $\delta > 0$.*

Assumption 12. *$\mathbf{E}_{\mathbb{G}^*,F}[ZZ'U^2] = |\mathcal{S}^*|^{-1}\sum_{i\in\mathcal{S}^*} Z_iZ_i'\,\mathbf{E}[U_i^2|X_i, D_i] \to_p \Delta^*$, as $|\mathcal{S}^*|$ tends to infinity, where Δ^* is a positive semi-definite matrix.*

Proposition 3. *Under Assumptions 6 and 10–12, as $|\mathcal{S}^*|$ tends to infinity,*

$$\sqrt{|\mathcal{S}^*|}(\hat{\theta}^* - \theta_{\mathbb{G}^*,F}) \to_d \mathcal{N}(0, (\Gamma^*)^{-1}\Delta^*(\Gamma^*)^{-1}), \tag{24}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean μ and covariance Σ.

Proposition 3 extends the previous full-sample result by Abadie et al. (2014, Theorem 2) to arbitrary subsamples.

As is addressed in Abadie et al. (2014), the conditional variance of U, i.e., $\mathbf{E}[U^2|X, D]$ complicates the estimation of the asymptotic variance, especially in case where X is continuous. Hence, they propose alternative estimators, which do not involve the estimation of the conditional variance. Here, we show that one of their estimators is also valid when using subsamples.

Assumption 13. *For some constant θ^*, as $|\mathcal{S}^*|$ tends to infinity, $\theta_{\mathbb{G}^*,F} \to_p \theta^*$.*

Assumption 14. *For $d \in \{0, 1\}$, $\mathcal{X}^d$ is compact with respect to a metric ρ.*

Assumption 15. *For $d \in \{0, 1\}$, $\mathbf{E}[Z_{j,i}^{r_j} Z_{k,i}^{r_k}(Y - Z'\theta^*)^{r_j+r_k}|X = x, D = d]$ is Lipschitz on $\mathcal{X}^d$ with respect to the metric ρ, where $Z_{j,i}$ and $Z_{k,i}$ denote the j^{th} and k^{th} components of Z_i, respectively, and r_j and r_k are non-negative integers that are no larger than 2.*

Assumption 13 does not require $\theta_{\mathbb{G}^*,F}$ to converge to $\theta_{G^*,F}$ for some G^* under which X and D are independent, which is the case explored in Abadie and Spiess (2022).

Now, for each $i \in \mathcal{S}^*$, define

$$l_{X,D}(i) \equiv \arg\min_{j \in \mathcal{S}^*} \|(X_i, D_i) - (X_j, D_j)\| \tag{25}$$

to be the index of the closest unit in $\mathcal{S}^*$ to i, where $\|(x_i, d_i) - (x_j, d_j)\| \equiv \rho(x_i, x_j) + \omega|d_i - d_j|$ for some $\omega > 0$. Let $\hat{E}_i^* \equiv Y_i - Z_i'\hat{\theta}^*$. Then, our proposed estimator for Δ^* is

$$\hat{\Delta}^* \equiv \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} (Z_i \hat{E}_i^* - Z_{l_{X,D}(i)} \hat{E}_{l_{X,D}(i)}^*)(Z_i \hat{E}_i^* - Z_{l_{X,D}(i)} \hat{E}_{l_{X,D}(i)}^*)'.$$

Proposition 4. *Under Assumptions 10, 12, and 13–15, as $|\mathcal{S}^*|$ tends to infinity,*

$$(\hat{\Gamma}^*)^{-1}\hat{\Delta}^*(\hat{\Gamma}^*)^{-1} \to_p (\Gamma^*)^{-1}\Delta^*(\Gamma^*)^{-1}. \tag{26}$$

6 Simulation

In this section, we conduct a simulation exercise that concretizes our discussion thus far; we demonstrate that the design phase adjusts the estimand of a regression, making it more robust to possible misspecification of the regression model, which ultimately leads to a reduced bias.

We adopt the simulation setup of Abadie and Imbens (2012, Section 6.1), who use the Boston U.S. Home Mortgage Disclosure Act (HMDA) dataset. This expanded version of the 1990 HMDA dataset, enriched by the Federal Reserve Bank of Boston, includes an additional 38 variables tied to minority status and critical to the mortgage lending decision, such as credit histories. It has been utilized in the literature to investigate racial discrimination in the mortgage market (Munnell et al., 1996).

Following the authors, the sample is restricted to male applicants purchasing single-family residences, who are either black or white, not self-employed, and were approved for private mortgage insurance. Furthermore, they must have no public record of default. This results in a sample of 148 black and 1336 white applicants.

For each unit i in this sample, let Y_i be the indicator that takes value 1 if i's mortgage application is denied, and 0 if approved. In addition, let D_i be the indicator that takes value 1 if i is black, and 0 if white. For simplicity, we use consumer credit history as a sole control variable, denoted as X_i, which takes on 6 values; each value indicates the frequency of prior delinquencies.

Our simulations proceed as follows. First, we run the logistic regression of Y_i on (X_i, D_i) and use the estimated model as the population conditional distribution of Y given (X, D), which we denote by F. Second, we construct a simulation sample consisting of N_1 blacks, i.e., $D = 1$, for whom the X values are generated from the empirical conditional distribution of X_i given $D_i = 1$, and $N_0 (\geq N_1)$ whites, i.e., $D = 0$, for whom the X values are generated from that given $D_i = 0$. For reference, we denote by $\mathbb{G}$ the empirical joint distribution of X and D in this simulation sample. Then, third, based on $\mathbb{G}$ and F, we compute the full-sample quantities, i.e., $\beta_{\mathbb{G},F}$, $\tau_{\mathbb{G},F}$, $m_{\mathbb{G},F}$, and $c_{\mathbb{G}}$. Fourth, we apply nearest-neighbor matching to the constructed simulation sample, where each of the N_1 black applicants is matched to a single white applicant with similar X; the remaining $N_0 - N_1$ whites are discarded. This step corresponds to the design phase in observational studies, though in practice researchers can employ alternative balancing techniques such as propensity score matching or trimming. We denote by $\mathbb{G}^*$ the empirical joint distribution of X and D in this matched subsample. Fifth, based on $\mathbb{G}^*$ and F, we compute the subsample quantities, i.e., $\beta_{\mathbb{G}^*,F}$, $\tau_{\mathbb{G}^*,F}$, $m_{\mathbb{G}^*,F}$, and $c_{\mathbb{G}^*}$. Finally, we repeat steps 2 to 5 for R times.

We consider the following three regression models:

Specification A: $Y_i = \alpha_A + \beta_A D_i + E_i$,

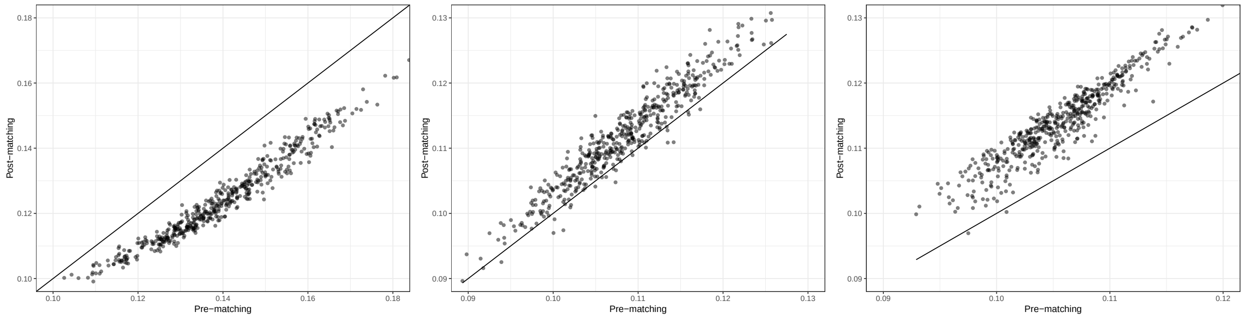
Specification B: $Y_i = \alpha_B + \beta_B D_i + \gamma_B \mathbf{1}\{X_i \geq 4\} + E_i$, and

Specification C: $Y_i = \alpha_C + \beta_C D_i + \gamma_C X_i + E_i$.

To underscore the utility of the design phase, even in cases where there are relatively few untreated units, we examine three scenarios $(N_1, N_0) \in \{(50, 75), (50, 100), (50, 125)\}$. However, only the second scenario is presented in the main text for conciseness. We set $R = 500$.

Figure 1 displays the pairs of the pre- and post-matching estimands, i.e., $\beta_{\mathbb{G},F}$ and $\beta_{\mathbb{G}^*,F}$, showing how nearest-neighbor matching adjusts the estimands of the regressions. It decreases the estimands for Specification A, while it generally results in an increase for Specifications B and C.

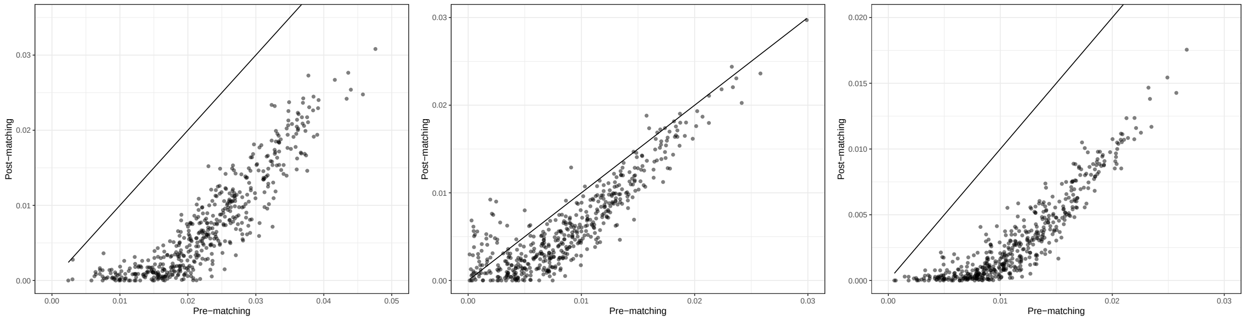
Figure 2 plots the pairs of the absolute sizes of the pre- and post-matching biases, i.e., $|\beta_{\mathbb{G},F} - \tau_{\mathbb{G},F}|$ and $|\beta_{\mathbb{G}^*,F} - \tau_{\mathbb{G}^*,F}|$. It illustrates that the estimand adjustment overall reduces the biases of the regressions. In light of our justification of the design phase, this reduction



Specification A Specification B Specification C

Figure 1: Pre- and Post-estimands ($N_1 = 50$, $N_0 = 100$)

should be driven by the extended range of the degree of misspecification that is allowed for the regression models. The next figures confirm that this is indeed the case.



Specification A Specification B Specification C

Figure 2: Pre- and Post-biases ($N_1 = 50$, $N_0 = 100$)

The points in Figure 3–5 denote the pairs of the values of model misspecification and covariate imbalance for the bounds provided in Section 4. The green ones are based on full samples, while the red ones are based on the subsamples constructed by nearest-neighbor matching. The curves denote the target levels of precision; if a point is situated within the curve labeled 0.005, it indicates that the multiplication of m and c, and therefore the bias, is of less than 0.005.

An immediate observation we can make is that the design phase generates subsamples that demonstrate better covariate balance when compared to the full sample; the distribution of the red points are shifted towards the left in comparison to the green points. This results in more red points being located inside each curve, which explains the overall reduction in bias after the design phase.

7 Conclusion

In this article, we investigate the role of the design phase in the context of running linear regression, offering a formal justification for its implementation. Our justification is twofold. First, we conceptualize the design phase as a subsample selection process, where a researcher adjusts the estimand of her regression. Then, we demonstrate that covariate balance is a valid criterion for this selection process, in that it quantifies the maximum misspecification that can be compromised for each subsample. As a result, the design phase can be understood as a means to identify a "better" estimand that is more robust to bias due to misspecification of the regression model.

References

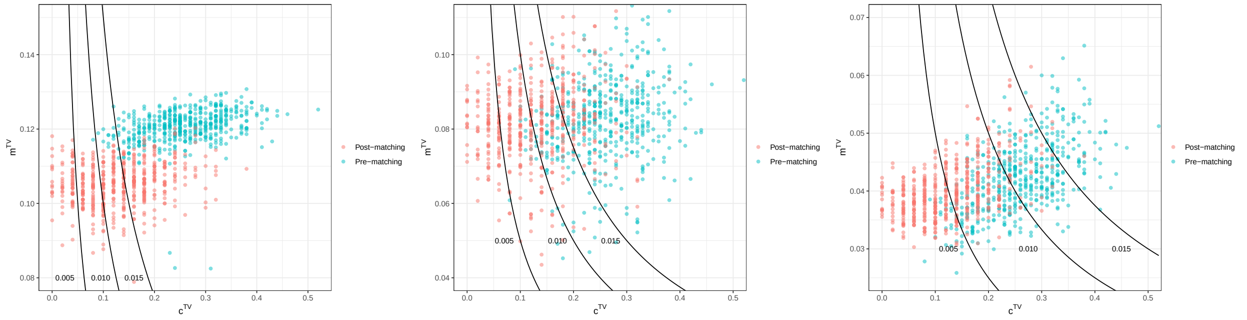
Abadie, A. and G. W. Imbens (2008). Estimation of the conditional variance in paired experiments. *Annales d'économie et de statistique* (91/92), 175–187.

Abadie, A. and G. W. Imbens (2012). Martingale representation for matching estimators. *Journal of the American Statistical Association 107*(498), 833–843.

Abadie, A., G. W. Imbens, and F. Zheng (2014). Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association 109*(508), 1601–1614.

Abadie, A. and J. Spiess (2022). Robust post-matching inference. *Journal of the American Statistical Association 117*(538), 983–995.

Billingsley, P. (2011). *Probability and Measure.* Wiley series in probability and statistics.

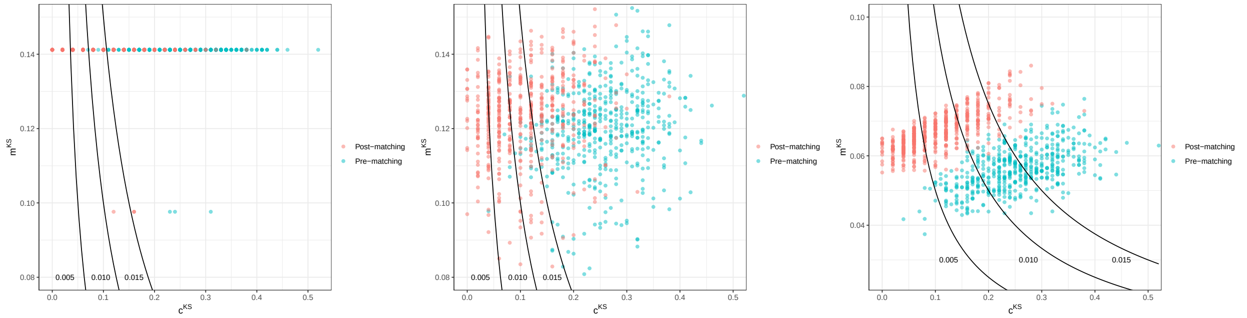


Specification A

Specification B

Specification C

Figure 3: Total variation bound ($N_1 = 50$, $N_0 = 100$)

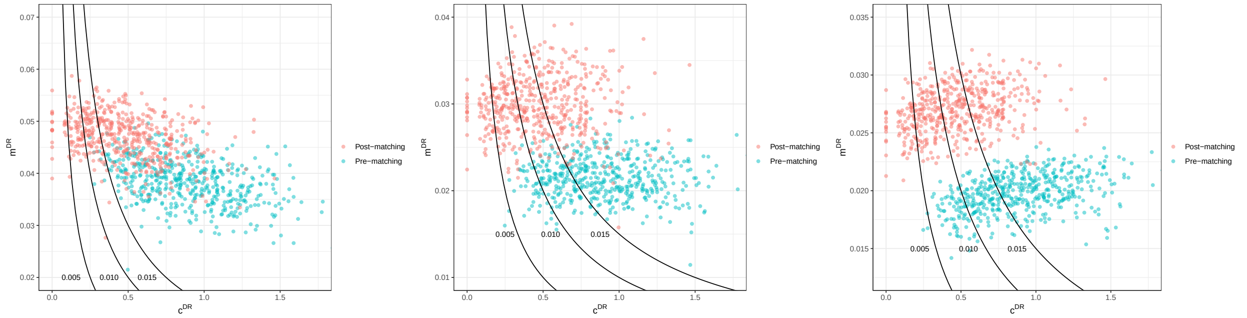


Specification A

Specification B

Specification C

Figure 4: Kolmogorov-Smirnov bound ($N_1 = 50$, $N_0 = 100$)



Specification A

Specification B

Specification C

Figure 5: Density ratio bound ($N_1 = 50$, $N_0 = 100$)

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

Freedman, D. A. (2008a). On Regression Adjustments in Experiments with Several Treatments. *The Annals of Applied Statistics 2*(1), 176–196.

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

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

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Kolmogorov, A. N. and S. V. Fomin (1975). *Introductory Real Analysis.* Dover Publications.

Lin, W. (2013). Agonistic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *The Annals of Applied Statistics 7*(1), 295–318.

Munnell, A. H., G. M. B. Tootell, L. E. Browne, and J. McEneaney (1996). Mortgage Lending in Boston: Interpreting HMDA Data. *The American Economic Review 86*(1), 25–53.

Negi, A. and J. M. Wooldridge (2021). Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects. *Econometric Reviews 40*(5), 504–534.

Protter, M. H. and C. B. Morrey (1991). *A First Course in Real Analysis.* Springer-Verlag New York.

A Proofs

A.1 Proof of Proposition 1

We use the following observation for our result.

Lemma 1. *Suppose that Assumption 1 holds. Then,*

$$\mathbf{E}[Y|D=d] = \alpha + \beta d + \mathbf{E}[s(X)'\gamma|D=d]. \tag{27}$$

Proof. Since the right-hand side of

$$Y - s(X)'\gamma = \alpha + \beta D + E \tag{28}$$

is saturated-in-D, $\mathbf{E}[E] = \mathbf{E}[DE] = 0$, and $\mathbf{E}[ZZ']$ is positive definite,

$$\mathbf{E}[Y - s(X)'\gamma|D] = \alpha + \beta D, \tag{29}$$

where the left-hand side is well-defined since Y and $s(X)$ have their first moments. Then, expanding the left-hand side yields the desired result. □

Now, we have

$$\begin{aligned}
\tau &= \mathbf{E}[Y|D=1] - \mathbf{E}[\mathbf{E}[Y|X, D=0]|D=1] \\
&= \mathbf{E}[Y|D=1] - \left(\int \mathbf{E}[Y|X=x, D=0]G^0(dx) + \int \mathbf{E}[Y|X=x, D=0](G^1 - G^0)(dx)\right) \\
&= \beta + (\mathbf{E}[s(X)'\gamma|D=1] - \mathbf{E}[s(X)'\gamma|D=0]) - \int \mathbf{E}[Y|X=x, D=0](G^1 - G^0)(dx) \\
&= \beta - \int (\mathbf{E}[Y|X=x, D=0] - s(x)'\gamma)(G^1 - G^0)(dx) \\
&= \beta - \int (\mathbf{E}[Y|X=x, D=0] - (\alpha + s(x)'\gamma))(G^1 - G^0)(dx),
\end{aligned}$$

where Lemma 1 is used in the third equality.

A.2 Proof of Corollary 1

Using the Jensen's inequality,

$$\begin{aligned}
&\left|\int (f(x,0)-l(x,0))(g^1(x)-g^0(x))\mu(dx)\right| \\
&\leq \int |f(x,0)-l(x,0)||g^1(x)-g^0(x)|\mu(dx) \\
&\leq \|f(x,0)-l(x,0)\|_{L^\infty(\mu)} \int |g^1(x)-g^0(x)|\mu(dx).
\end{aligned}$$

A.3 Proof of Corollary 2

By Assumption 3, $f(\cdot,0)-l(\cdot,0)$ is continuous and bounded on $\mathcal{X}^0$, which is closed by definition. By the Tietze extension theorem, there exists a continuous and bounded extension on $\mathbb{R}$. Thus, $\mathcal{H}$ is non-empty. Take any $h \in \mathcal{H}$. Then,

$$\begin{aligned}
\int_{[a,b]} (f(x,0)-l(x,0))G^d(dx) &= \int_{[a,b]\cap\mathcal{X}^0} (f(x,0)-l(x,0))G^d(dx) \\
&= \int_{[a,b]\cap\mathcal{X}^0} h(x)G^d(dx) = \int_{[a,b]} h(x)G^d(dx),
\end{aligned}$$

where Assumption 2 is used in the first and last equalities when $d=1$. Since h is continuous and G^d is nondecreasing, Riemann-Stieltjes integral $\int_a^b h(x)dG^d(x)$ exists and coincides with $\int_{[a,b]} h(x)G^d(dx)$. (Kolmogorov and Fomin, 1975, p.368) Integrating by parts,

$$\int_{[a,b]} h(x)G^d(dx) = \int_a^b h(x)dG^d(x) = h(b)G^d(b) - h(a)G^d(a) - \int_a^b G^d(x)dh(x).$$

(Protter and Morrey, 1991, p.320) Combining the previous two observations, we have

$$\int_{[a,b]} (f(x,0)-l(x,0))(G^1-G^0)(dx) = h(b)(G^1-G^0)(b) - h(a)(G^1-G^0)(a) - \int_a^b (G^1-G^0)(x)dh(x).$$

The third term of the right-hand side is bounded by

$$\left|\int_a^b (G^1 - G^0)(x)dh(x)\right| \leq \sup_{x\in[a,b]} |G^1(x) - G^0(x)| V_a^b(h) \leq \sup_{x\in[-\infty,\infty]} |G^1(x) - G^0(x)| V_{-\infty}^{\infty}(h).$$

Taking $a, b \to \pm\infty$, by the dominated convergence theorem, and since $\lim_{b\to\infty} h(b)(G^1 - G^0)(b) = \lim_{a\to-\infty} h(a)(G^1 - G^0)(a) = 0$, where we use the boundedness of h,

$$\lim_{a,b\to\pm\infty}\left|\int_a^b (G^1 - G^0)(x)dh(x)\right| = \left|\int_{[-\infty,\infty]} (f(x,0) - l(x,0))(G^1 - G^0)(dx)\right|.$$

This term does not depend on h, and thus we have

$$\left|\int (f(x,0) - l(x,0))(G^1 - G^0)(dx)\right| \leq \sup_{x\in[-\infty,\infty]} |G^1(x) - G^0(x)| \inf_{h\in\mathcal{H}} V_{-\infty}^{\infty}(h).$$

A.4 Proof of Corollary 3

In cases where either m^{DR} or c^{DR} is infinity, the inequality trivially holds. Thus, we assume that both are finite. Then, by the Cauchy-Schwartz inequality,

$$\begin{aligned} m^{\mathrm{DR}} c^{\mathrm{DR}} &\geq \int |(f(x,0) - l(x,0))(g^1(x)/g^0(x) - 1)| G^0(dx) && (30) \\ &= \int |(f(x,0) - l(x,0))(g^1(x)/g^0(x) - 1)| g^0(x)\mu(dx) && (31) \\ &\geq \left|\int (f(x,0) - l(x,0))(g^1(x) - g^0(x))\mu(dx)\right|. && (32) \end{aligned}$$

A.5 Proof of Corollary 4

It follows from Corollary 1, 2, or 3 that $|\beta - \tau_0| \leq mc$. Note that

$$\frac{\hat{\beta} - \beta}{se_\beta(\hat{\beta})} - \frac{mc}{se_\beta(\hat{\beta})} \leq t = \frac{\hat{\beta} - \tau_0}{se_\beta(\hat{\beta})} = \frac{\hat{\beta} - \beta}{se_\beta(\hat{\beta})} + \frac{\beta - \tau_0}{se_\beta(\hat{\beta})} \leq \frac{\hat{\beta} - \beta}{se_\beta(\hat{\beta})} + \frac{mc}{se_\beta(\hat{\beta})},$$

where the left- and right-hand sides converge in distribution to $\mathcal{N}(-v/\sigma_\beta, 1)$ and $\mathcal{N}(v/\sigma_\beta, 1)$ by Assumption 5 and the Slutsky theorem. Thus, as N tends to infinity,

$$\begin{aligned}
\mathbf{Pr}[t \le z] &\ge \mathbf{Pr}\left[\frac{\hat\beta - \beta}{se_\beta(\hat\beta)} + \frac{mc}{se_\beta(\hat\beta)} \le z\right] \to \Phi(z - v/\sigma_\beta) \text{ and} \\
\mathbf{Pr}[t \le z] &\le \mathbf{Pr}\left[\frac{\hat\beta - \beta}{se_\beta(\hat\beta)} - \frac{mc}{se_\beta(\hat\beta)} \le z\right] \to \Phi(z + v/\sigma_\beta).
\end{aligned}$$

A.6 Proof of Proposition 2

We first show that $\mathbf{E}_{\mathbb{G}^*}[Z(\mathbf{E}_F[Y|X, D] - \mathbf{E}_{\mathbb{F}^*}[Y|X, D])]$ converges to zero in probability. For any joint distribution $\tilde G$ of D and X,

$$\begin{aligned}
&\mathbf{E}_{\tilde G}[Z(\mathbf{E}_F[Y|X, D] - \mathbf{E}_{\mathbb{F}^*}[Y|X, D])] \\
&= \mathbf{E}_{\tilde G}\left[Z\left(\mathbf{E}_F[Y|X, D] - \sum_{i\in\mathcal{S}^*} \frac{\mathbf{1}\{X_i = X, D_i = D\}}{\sum_{j\in\mathcal{S}^*} \mathbf{1}\{X_j = X, D_j = D\}} Y_i\right)\right] \\
&= \mathbf{E}_{\tilde G}\left[Z\left(\mathbf{E}_F[Y|X, D] - \sum_{i\in\mathcal{S}^*} \frac{\mathbf{1}\{X_i = X, D_i = D\}}{\sum_{j\in\mathcal{S}^*} \mathbf{1}\{X_j = X, D_j = D\}} (\mathbf{E}_F[Y_i|X_i, D_i] + U_i)\right)\right] \\
&= \mathbf{E}_{\tilde G}\left[Z\left(-\sum_{i\in\mathcal{S}^*} \frac{\mathbf{1}\{X_i = X, D_i = D\}}{\sum_{j\in\mathcal{S}^*} \mathbf{1}\{X_j = X, D_j = D\}} U_i\right)\right] \\
&= -\sum_{i\in\mathcal{S}^*} \mathbf{E}_{\tilde G}\left[Z \frac{\mathbf{1}\{X_i = X, D_i = D\}}{\sum_{j\in\mathcal{S}^*} \mathbf{1}\{X_j = X, D_j = D\}}\right] U_i.
\end{aligned}$$

Since

$$\begin{aligned}
\mathbf{E}_{\mathbb{G}^*}\left[Z \frac{\mathbf{1}\{X_i = X, D_i = D\}}{\sum_{j\in\mathcal{S}^*} \mathbf{1}\{X_j = X, D_j = D\}}\right] &= \frac{1}{|\mathcal{S}^*|} \sum_{k\in\mathcal{S}^*} Z_k \frac{\mathbf{1}\{X_i = X_k, D_i = D_k\}}{\sum_{j\in\mathcal{S}^*} \mathbf{1}\{X_j = X_k, D_j = D_k\}} \\
&= Z_i \frac{1}{|\mathcal{S}^*|} \underbrace{\sum_{k\in\mathcal{S}^*} \frac{\mathbf{1}\{X_i = X_k, D_i = D_k\}}{\sum_{j\in\mathcal{S}^*} \mathbf{1}\{X_j = X_k, D_j = D_k\}}}_{=1 \text{ if } j=k \text{ and } =0 \text{ otherwise}} = \frac{1}{|\mathcal{S}^*|} Z_i,
\end{aligned}$$

we then have

$$\mathbf{E}_{\mathbb{G}^*}[Z(\mathbf{E}_F[Y|X,D] - \mathbf{E}_{\mathbb{F}^*}[Y|X,D])] = -\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i U_i.$$

Note that

$$\begin{aligned}\mathbf{Var}\left[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i U_i \,\middle|\, \mathbb{X},\mathbb{D}\right] &= \mathbf{Var}\left[\frac{1}{|\mathcal{S}^*|}\sum_{i=1}^{N} \mathbf{1}\{i\in\mathcal{S}^*\} Z_i U_i \,\middle|\, \mathbb{X},\mathbb{D}\right]\\ &= \frac{1}{|\mathcal{S}^*|^2}\sum_{i,j}\mathbf{Cov}[\mathbf{1}\{i\in\mathcal{S}^*\}Z_iU_i, \mathbf{1}\{j\in\mathcal{S}^*\}Z_jU_j|\mathbb{X},\mathbb{D}]\\ &= \frac{1}{|\mathcal{S}^*|^2}\sum_{i,j}\mathbf{1}\{i\in\mathcal{S}^*\}\mathbf{1}\{j\in\mathcal{S}^*\}\,\mathbf{Cov}[Z_iU_i, Z_jU_j|\mathbb{X},\mathbb{D}]\\ &= \frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\mathbf{Var}[Z_iU_i|\mathbb{X},\mathbb{D}]\\ &= \frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\mathbf{E}[Z_iU_i(Z_iU_i)'|X_i,D_i],\end{aligned}$$

where the second and third to the last equation hold by

$$\begin{aligned}\mathbf{Cov}[Z_iU_i, Z_jU_j|\mathbb{X},\mathbb{D}] &= \mathbf{E}[Z_iZ_j'U_iU_j|\mathbb{X},\mathbb{D}] - \mathbf{E}[Z_iU_i|\mathbb{X},\mathbb{D}]\,\mathbf{E}[Z_jU_j|\mathbb{X},\mathbb{D}]'\\ &= Z_iZ_j'\,\mathbf{E}[U_iU_j|\mathbb{X},\mathbb{D}] - Z_iZ_j'\,\mathbf{E}[U_i|\mathbb{X},\mathbb{D}]\,\mathbf{E}[U_j|\mathbb{X},\mathbb{D}]' = 0\end{aligned}$$

and Assumption 6, respectively. Thus, by Assumption 8, it follows that

$$\left\|\mathbf{Var}\left[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i U_i \,\middle|\, \mathbb{X},\mathbb{D}\right]\right\| \le \frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\mathbf{E}[\|Z_iU_i\|^2|X_i,D_i] \le \frac{M}{|\mathcal{S}^*|}. \tag{33}$$

for some $M > 0$. Since

$$\mathbf{Var}\left[\mathbf{E}\left[-\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i U_i \,\middle|\, \mathbb{X},\mathbb{D}\right]\right] = \mathbf{Var}\left[-\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i \underbrace{\mathbf{E}[U_i|X_i,D_i]}_{=0}\right] = 0,$$

by the law of total variance and equation (33), we have

$$\left\|\mathbf{Var}\left[-\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_iU_i\right]\right\| = \left\|\mathbf{E}\left[\mathbf{Var}\left[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_iU_i \;\middle|\; \mathbb{X},\mathbb{D}\right]\right]\right\| \le \mathbf{E}\left[\frac{M}{\mathcal{S}^*}\right] \to 0,$$

as N tends to infinity. The inequality holds, because $1/\mathcal{S}^*$ is bounded by one and converges to zero in probability.

Now, if $\lambda^* > \underline{\lambda}/2 > 0$,

$$\begin{aligned}
&\|(\mathbf{E}_{\mathbb{G}^*}[ZZ'])^{-1}\,\mathbf{E}_{\mathbb{G}^*}[Z(\mathbf{E}_F[Y|X,D] - \mathbf{E}_{\mathbb{F}^*}[Y|X,D])]\| \\
&\le \|(\mathbf{E}_{\mathbb{G}^*}[ZZ'])^{-1}\|\|\,\mathbf{E}_{\mathbb{G}^*}[Z(\mathbf{E}_F[Y|X,D] - \mathbf{E}_{\mathbb{F}^*}[Y|X,D])]\| \\
&\lesssim \|\,\mathbf{E}_{\mathbb{G}^*}[Z(\mathbf{E}_F[Y|X,D] - \mathbf{E}_{\mathbb{F}^*}[Y|X,D])]\|.
\end{aligned}$$

Thus, by Assumption 7, for $\eta > 0$,

$$\begin{aligned}
&\mathbf{Pr}[\|(\mathbf{E}_{\mathbb{G}^*}[ZZ'])^{-1}\,\mathbf{E}_{\mathbb{G}^*}[Z(\mathbf{E}_F[Y|X,D] - \mathbf{E}_{\mathbb{F}^*}[Y|X,D])]\| > \eta] \\
&\le \mathbf{Pr}[\lambda^* \le \underline{\lambda}/2] + \mathbf{Pr}[\|\,\mathbf{E}_{\mathbb{G}^*}[Z(\mathbf{E}_F[Y|X,D] - \mathbf{E}_{\mathbb{F}^*}[Y|X,D])]\| > \eta] \to_p 0.
\end{aligned}$$

A.7 Proof of Proposition 3

We first show that $(1/\sqrt{|\mathcal{S}^*|})\sum_{i\in\mathcal{S}^*} Z_iU_i \to_d \mathcal{N}(0,\Delta^*)$.

Fix $\lambda \in \mathbb{R}$. Let $\xi_i \equiv (1/\sqrt{|\mathcal{S}^*|})\mathbf{1}\{i \in \mathcal{S}^*\}Z_iU_i$ so that

$$\sum_{i=1}^N \xi_i = \frac{1}{\sqrt{|\mathcal{S}^*|}}\sum_{i\in\mathcal{S}^*} Z_iU_i.$$

Consider the filtration $\mathcal{F}_k \equiv \sigma(D_1,\ldots,D_N,X_1,\ldots,X_N,\{U_i\}_{i\in\mathcal{S}^*,1\le i\le k})$, where $\sigma(\cdot)$ denotes the σ-field generated by the arguments. Then, $\sum_{i=1}^k \xi_i'\lambda$ is a martingale with respect to $\mathcal{F}_k$, because (i) $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$, (ii) $\sigma(\sum_{i=1}^k \xi_i'\lambda) \subseteq \mathcal{F}_k$, (iii) by Assumption 11,

$$\mathbf{E}\left[\left|\sum_{i=1}^k \xi_i'\lambda\right|\right] \leq \left(\sum_{i=1}^k \mathbf{E}[\|\xi_i\|_2]\right)\|\lambda\|_2 \leq \left(\sum_{i=1}^k (\mathbf{E}[\|Z_iU_i\|^2])^{1/2}\right)\|\lambda\|_2 < \infty, \tag{34}$$

and (iv)

$$\mathbf{E}\left[\sum_{i=1}^k \xi_i'\lambda \;\middle|\; \mathcal{F}_{k-1}\right] = \begin{cases} \sum_{i=1}^{k-1} \xi_i'\lambda + Z_k\,\mathbf{E}[U_k|X_k, D_k]'\lambda = \sum_{i=1}^{k-1} \xi_i'\lambda & \text{if } k \in \mathcal{S}^*, \\ \mathbf{E}[\sum_{i=1}^{k-1} \xi_i'\lambda|\mathcal{F}_{k-1}] = \sum_{i=1}^{k-1} \xi_i'\lambda & \text{if } k \notin \mathcal{S}^*. \end{cases}$$

Note that by Assumptions 6 and 12,

$$\sum_{i=1}^N \mathbf{E}[(\xi_i'\lambda)^2|\mathcal{F}_{i-1}] = \lambda'\sum_{i=1}^N \mathbf{E}[\xi_i\xi_i'|\mathcal{F}_{i-1}]\lambda = \lambda'\left(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_iZ_i'\,\mathbf{E}[U_i^2|X_i, D_i]\right)\lambda \to_p \lambda'\Delta^*\lambda,$$

and for each $\epsilon > 0$, by Assumption 11,

$$\begin{aligned}
\sum_{i=1}^N \mathbf{E}[|\xi_i'\lambda|^2\mathbf{1}\{\xi_i'\lambda > \epsilon\}] &\leq \sum_{i=1}^N \mathbf{E}\left[|\xi_i'\lambda|^{2+\delta}\frac{1}{\epsilon^\delta}\right] \\
&\leq \frac{\|\lambda|^{2+\delta}}{\epsilon^\delta}\sum_{i=1}^N \mathbf{E}[\|\xi_i\|^{2+\delta}] \\
&\leq \frac{\|\lambda|^{2+\delta}}{\epsilon^\delta}\mathbf{E}\left[\sum_{i=1}^N \mathbf{E}[\|\xi_i\|^{2+\delta}|\mathbb{X}, \mathbb{D}]\right] \\
&\leq \frac{\|\lambda|^{2+\delta}}{\epsilon^\delta}\mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^{1+\delta/2}}\sum_{i\in\mathcal{S}^*} \mathbf{E}[\|Z_iU_i\|^{2+\delta}|X_i, D_i]\right] \\
&\leq \frac{1}{\epsilon^\delta}\|\lambda\|_2^{2+\delta}\,\mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^{\delta/2}}\right]c \to 0,
\end{aligned}$$

where c is some constant. Then, by the Lindeberg-Feller martingale central limit theorem (Billingsley, 2011, Theorem 35.12), $\sum_{i=1}^n \xi_i'\lambda \to_d \mathcal{N}(0, \lambda'\Delta^*\lambda)$, and the assertion follows from the Cramér-Wold device.

Now, by Slutsky's lemma and Assumption 10, we have

$$\sqrt{|\mathcal{S}^*|}(\theta_{\mathbb{G}^*,\mathbb{F}^*} - \theta_{\mathbb{G}^*,F}) = \left(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} Z_i Z_i'\right)^{-1}\left(\frac{1}{\sqrt{|\mathcal{S}^*|}}\sum_{i\in\mathcal{S}^*} Z_i U_i\right) \to_d \mathcal{N}(0, (\Gamma^*)^{-1}\Delta^*(\Gamma^*)^{-1}).$$

A.8 Proof of Proposition 4

Lemma 4 is required for our result, and for that, we show Lemma 2–3. The proofs of Lemma 2–4 closely follow those of Abadie and Imbens (2008, Lemma 1) and Abadie et al. (2014, Lemma A.2, Lemma A.3), respectively.

Lemma 2. *Let $W_1, \ldots, W_N$ be random variables whose support $\mathcal{W} \subseteq \mathbb{R}^p$ is bounded with respect to a metric $\|\|$. Let S^* be a function of $\mathbb{W} \equiv (W_i)_{i=1}^N$. Define $l_W(i) \equiv \arg\min_{j\in\mathcal{S}^*, j\neq i} \|W_i - W_j\|$ to be the index of the closest unit in $\mathcal{S}^*$ to i. Then, if $\operatorname{plim}_{N\to\infty} |S^*(\mathbb{W})| = \infty$, as N tends to infinity,*

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \|W_i - W_{l_W(i)}\| \to_p 0. \tag{35}$$

Proof. Let $\epsilon > 0$ be given. Suppose that there are M i's such that $\|W_i - W_{l_W(i)}\| > 2\epsilon$. Define $B_\epsilon(w) \equiv \{w' \in \mathcal{W} : \|w' - w\| < \epsilon\}$. By the definition of $l_W(i)$, then $B_\epsilon(W_i) \cap B_\epsilon(W_j) = \emptyset$ for all $j \in \mathcal{S}^*$ such that $j \neq i$.

Since $\mathcal{W}$ is bounded, there exists a closed ball $\mathcal{C}$ with radius no larger than $\operatorname{diam}(\mathcal{W})$ such that $\mathcal{W} \subseteq \mathcal{C}$, where $\operatorname{diam}(\mathcal{W})$ denotes the diameter of $\mathcal{W}$. Then,

$$\begin{aligned} M\frac{\pi^{p/2}\epsilon^p}{\Gamma(\frac{p}{2}+1)} &= \sum_{i\in\mathcal{S}^*:\|W_i - W_{l_W(i)}\|>2\epsilon} \operatorname{Vol}[B_\epsilon(W_i)] \\ &= \operatorname{Vol}\left[\bigcup_{i\in\mathcal{S}^*:\|W_i - W_{l_W(i)}\|>2\epsilon} B_\epsilon(W_i)\right] \\ &< \operatorname{Vol}[\mathcal{C}^\epsilon] = \frac{\pi^{p/2}(\operatorname{diam}(\mathcal{W})(1+\epsilon))^p}{\Gamma(\frac{p}{2}+1)}, \end{aligned}$$

where $\mathcal{C}^\epsilon$ denotes ϵ-enlargement of $\mathcal{C}$. Hence, $M < \operatorname{diam}(\mathcal{W})^p((1+\epsilon)/\epsilon)^p$. Assuming $|\mathcal{S}^*| >$

M, we then have

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|W_i - W_{l_W(i)}\| \le \frac{1}{|\mathcal{S}^*|}\Big(M\,\mathrm{diam}(\mathcal{W}) + 2\epsilon(|\mathcal{S}^*| - M)\Big)$$
$$< \mathrm{diam}(\mathcal{W})\frac{M}{|\mathcal{S}^*|} + 2\epsilon < \mathrm{diam}(\mathcal{W})^{p+1}\frac{(\frac{1+\epsilon}{\epsilon})^p}{|\mathcal{S}^*|} + 2\epsilon.$$

Now, let $\epsilon = 1/|\mathcal{S}^*|^{1/(p+1)}$. Then, the last equation converges to zero in probability, because $\epsilon \to_p 0$ and for any $\eta > 0$,

$$\mathbf{Pr}\bigg[\mathrm{diam}(\mathcal{W})^{p+1}\frac{(\frac{1+\epsilon}{\epsilon})^p}{|\mathcal{S}^*|} > \eta\bigg] \le \mathbf{Pr}\bigg[\Big(\frac{1+\epsilon}{\epsilon}\Big)^p \gtrsim |\mathcal{S}^*|\bigg] = \mathbf{Pr}[|\mathcal{S}^*|^{-1/(p+1)} \gtrsim 1] \to 0.$$

Combining results, for any $\eta > 0$,

$$\begin{aligned}&\mathbf{Pr}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|W_i - W_{l_W(i)}\| > \eta\bigg]\\ &\le \mathbf{Pr}[M \ge |\mathcal{S}^*|] + \mathbf{Pr}\bigg[|\mathcal{S}^*| > M, \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|W_i - W_{l_W(i)}\| > \eta\bigg]\\ &\le \mathbf{Pr}\bigg[\Big(\frac{1+\epsilon}{\epsilon}\Big)^p \gtrsim |\mathcal{S}^*|\bigg] + \mathbf{Pr}\bigg[\mathrm{diam}(\mathcal{W})^{p+1}\frac{(\frac{1+\epsilon}{\epsilon})^p}{|\mathcal{S}^*|} + 2\epsilon > \eta\bigg] \to 0.\end{aligned}$$

□

Lemma 3. *Suppose that $(V_1, W_1), \ldots, (V_N, W_N)$ are i.i.d., where V_i is scalar and the support of W_i, which we denote by $\mathcal{W}$, is compact. Let $S^* \subseteq \{1, \ldots, N\}$ be a function of $\mathbb{W} \equiv (W_i)_{i=1}^N$. Assume that $\mu_r(w) \equiv \mathbf{E}[V_i^r|W_i = w]$ is Lipschitz in w with a constant L_r, where $r = 1, \ldots, R$. Then, for all nonnegative integers k and m such that $k \vee m \le R/2$, as $|\mathcal{S}^*|$ tends to infinity,*

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_i^k V_{l_W(i)}^m - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i] \to_p 0. \qquad (36)$$

Proof. We first show that

$$\mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_i^k V_{l_W(i)}^m - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\bigg] = o(1). \tag{37}$$

Note that V_i and $V_{l_W(i)}$ is independent conditional on $\mathbb{W}$. Hence,

$$\begin{aligned}
\mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_i^k V_{l_W(i)}^m\bigg] &= \mathbf{E}\bigg[\mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_i^k V_{l_W(i)}^m \,\bigg|\, \mathbb{W}\bigg]\bigg] \\
&= \mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_{l_W(i)}^m|W_{l_W(i)}]\bigg] \\
&= \mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \mu_k(W_i)\mu_m(W_i)\bigg] + \mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \mu_k(W_i)(\mu_m(W_{l_W(i)}) - \mu_m(W_i))\bigg].
\end{aligned}$$

The second term in the last equation is bounded by

$$\begin{aligned}
&\bigg|\mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \mu_k(W_i)(\mu_m(W_{l_W(i)}) - \mu_m(W_i))\bigg]\bigg| \\
&\leq \mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} |\mu_k(W_i)||\mu_m(W_{l_W(i)}) - \mu_m(W_i)|\bigg] \\
&\leq c_k \cdot \mathbf{E}\bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \|W_{l_W(i)} - W_i\|\bigg] \cdot L_m = o(1),
\end{aligned}$$

where $c_l \equiv \sup_{w\in\mathcal{W}} \mu_l(w)$, which is finite by the compactness of $\mathcal{W}$ and (Lipschitz) continuity of μ_k. The second inequality follows from $\mu_r(\cdot)$ being Lipschitz, and the last equality holds as $|\mathcal{S}^*|^{-1}\sum_{i\in\mathcal{S}^*} \|W_{l_W(i)} - W_i\|$ is bounded and $o_p(1)$ by Lemma 3. Equation (37) then follows.

Next, we show that

$$\mathbf{E}\bigg[\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_i^k V_{l_W(i)}^m - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} \mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\bigg)^2\bigg] = o(1). \tag{38}$$

We expand the left-hand side of equation (38):

$$\underbrace{\mathbf{E}\Bigg[\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}V_i^kV_{l_W(i)}^m\bigg)^2\Bigg]}_{\equiv A}+\underbrace{\mathbf{E}\Bigg[\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\bigg)^2\Bigg]}_{\equiv B}$$
$$-2\underbrace{\mathbf{E}\Bigg[\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}V_i^kV_{l_W(i)}^m\bigg)\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\bigg)\Bigg]}_{\equiv C}.$$

We start by investigating the term C. Note that

$$C=\mathbf{E}\Bigg[\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\bigg)^2\Bigg]$$
$$+\mathbf{E}\Bigg[\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}V_i^kV_{l_W(i)}^m-\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\bigg)\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\bigg)\Bigg],$$

where the second term is bounded by

$$\Bigg|\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}V_i^kV_{l_W(i)}^m-\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\Bigg]\Bigg|\cdot c_kc_m=o(1)$$

where we use equation (37) in the last equality. Thus, we obtain $C=B+o(1)$.

Then, we argue that $A=B+o(1)$. The first term in the right-hand side of

$$\mathbf{E}\Bigg[\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}V_i^kV_{l_W(i)}^m\bigg)^2\Bigg]=\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}V_i^{2k}V_{l_W(i)}^{2m}\Bigg]+\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i}V_i^kV_{l_W(i)}^mV_j^kV_{l_W(j)}^m\Bigg]$$

is $o(1)$, because

$$\begin{aligned}\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}V_i^{2k}V_{l_W(i)}^{2m}\Bigg]&=\mathbf{E}\Bigg[\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}V_i^{2k}V_{l_W(i)}^{2m}\;\Bigg|\;\mathbb{W}\Bigg]\Bigg]\\&=\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_i^{2k}|W_i]\,\mathbf{E}[V_{l_W(i)}^{2m}|W_{l_W(i)}]\Bigg]\\&\le\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|}\Bigg]c_{2k}c_{2m}=o(1),\end{aligned}$$

where $2k, 2m \leq R$. Thus, we focus on the second term.

It can be decomposed into two terms:

$$\begin{aligned}&\mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i: l_W(j)\neq i, l_W(i)\neq j, l_W(j)\neq l_W(i)} V_i^k V_{l_W(i)}^m V_j^k V_{l_W(j)}^m\right]\\ &+\mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i: l_W(j)=i \text{ or } l_W(i)=j \text{ or } l_W(j)=l_W(i)} V_i^k V_{l_W(i)}^m V_j^k V_{l_W(j)}^m\right].\end{aligned} \tag{39}$$

Note that for each $i \in \mathcal{S}^*$, $|\{j \in \mathcal{S}^* : l_W(j) = i\}| \leq \overline{K}(p)$, where $\overline{K}(p)$ denotes the "kissing number," i.e., the maximum number of times that each unit can be used as a match in p-dimensions. Since $|\{j \in \mathcal{S}^* : l_W(i) = j\}| = 1$, $|\{j \neq i : l_W(j) = i \text{ or } l_W(i) = j \text{ or } l_W(j) = l_W(i)\}| \leq 2\overline{K}(p) + 1$. Then, the second term of equation (39) is bounded by

$$\begin{aligned}&\left|\mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i: l_W(j)=i \text{ or } l_W(i)=j \text{ or } l_W(j)=l_W(i)} V_i^k V_{l_W(i)}^m V_j^k V_{l_W(j)}^m\right]\right|\\ &=\mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i: l_W(j)=i \text{ or } l_W(i)=j \text{ or } l_W(j)=l_W(i)} |\,\mathbf{E}[V_i^k V_{l_W(i)}^m V_j^k V_{l_W(j)}^m|\mathbb{W}]|\right] \lesssim \mathbf{E}\left[\frac{1}{|\mathcal{S}^*|}\right] = o(1),\end{aligned}$$

where the inequality results from

$$\sum_{j\neq i: l_W(j)=i \text{ or } l_W(i)=j \text{ or } l_W(j)=l_W(i)} |\,\mathbf{E}[V_i^k V_{l_W(i)}^m V_j^k V_{l_W(j)}^m|\mathbb{W}]| \leq (2\overline{K}(p)+1)\cdot(c_{k+m}^2 \vee c_{k+m}c_k c_m \vee c_{2m}c_k^2).$$

Hence, it suffices to show that the first term of equation (39) converges to B.

This is done by three steps. First, note that

$$\begin{aligned}&\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i: l_W(j)\neq i, l_W(i)\neq j, l_W(j)\neq l_W(i)} V_i^k V_{l_W(i)}^m V_j^k V_{l_W(j)}^m\\ &\quad -\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i: l_W(j)\neq i, l_W(i)\neq j, l_W(j)\neq l_W(i)} \mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\,\mathbf{E}[V_j^k|W_j]\,\mathbf{E}[V_j^m|W_j]\Bigg]\\ &=\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i: l_W(j)\neq i, l_W(i)\neq j, l_W(j)\neq l_W(i)} \mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_j^k|W_j]\\ &\quad\times(\mathbf{E}[V_{l_W(i)}^m|W_{l_W(i)}]\,\mathbf{E}[V_{l_W(j)}^m|W_{l_W(j)}] - \mathbf{E}[V_i^m|W_i]\,\mathbf{E}[V_j^m|W_j])\Bigg]\end{aligned}$$

$$\le \mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i:l_W(j)\neq i,l_W(i)\neq j,l_W(j)\neq l_W(i)}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_j^k|W_j]$$
$$\times\frac{\mathbf{E}[V_{l_W(i)}^m|W_{l_W(i)}]+\mathbf{E}[V_i^m|W_i]}{2}(\mathbf{E}[V_{l_W(j)}^m|W_{l_W(j)}]-\mathbf{E}[V_j^m|W_j]])\Bigg]$$
$$+\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i:l_W(j)\neq i,l_W(i)\neq j,l_W(j)\neq l_W(i)}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_j^k|W_j]$$
$$\times\frac{\mathbf{E}[V_{l_W(j)}^m|W_{l_W(j)}]+\mathbf{E}[V_j^m|W_j]}{2}(\mathbf{E}[V_{l_W(i)}^m|W_{l_W(i)}]-\mathbf{E}[V_i^m|W_i]])\Bigg].$$

We bound each term in the last equation. The first term is bounded by

$$\Bigg|\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i:l_W(j)\neq i,l_W(i)\neq j,l_W(j)\neq l_W(i)}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_j^k|W_j]$$
$$\times\frac{\mathbf{E}[V_{l_W(i)}^m|W_{l_W(i)}]+\mathbf{E}[V_i^m|W_i]}{2}(\mathbf{E}[V_{l_W(j)}^m|W_{l_W(j)}]-\mathbf{E}[V_j^m|W_j]])\Bigg]\Bigg|$$
$$\le \mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i:l_W(j)\neq i,l_W(i)\neq j,l_W(j)\neq l_W(i)}|\,\mathbf{E}[V_i^k|W_i]||\,\mathbf{E}[V_j^k|W_j]|$$
$$\times\frac{|\,\mathbf{E}[V_{l_W(i)}^m|W_{l_W(i)}]|+|\,\mathbf{E}[V_i^m|W_i]|}{2}|\,\mathbf{E}[V_{l_W(j)}^m|W_{l_W(j)}]-\mathbf{E}[V_j^m|W_j]]|\Bigg]\Bigg|$$
$$\le \mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\in\mathcal{S}^*}c_k^2c_mL_m\|W_{l_W(j)}-W_j\|\Bigg]=c_k^2c_mL_m\,\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|}\sum_{j\in\mathcal{S}^*}\|W_{l_W(j)}-W_j\|\Bigg]=o(1),$$

where we use Lemma 2 in the last equation. Likewise, the second term can be bounded by $o(1)$. Next,

$$\Bigg|\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i:l_W(j)\neq i,l_W(i)\neq j,l_W(j)\neq l_W(i)}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\,\mathbf{E}[V_j^k|W_j]\,\mathbf{E}[V_j^m|W_j]$$
$$-\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i}\mathbf{E}[V_i^k|W_i]\,\mathbf{E}[V_i^m|W_i]\,\mathbf{E}[V_j^k|W_j]\,\mathbf{E}[V_j^m|W_j]\Bigg]\Bigg|$$
$$=\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|^2}\sum_{i\in\mathcal{S}^*}\sum_{j\neq i:l_W(j)=i\text{ or }l_W(i)=j\text{ or }l_W(j)=l_W(i)}\underbrace{|\,\mathbf{E}[V_i^k|W_i]||\,\mathbf{E}[V_i^m|W_i]||\,\mathbf{E}[V_j^k|W_j]||\,\mathbf{E}[V_j^m|W_j]|}_{\le c_k^2c_m^2}\Bigg]$$
$$\le c_k^2c_m^2(2\overline{K}(p)+1)\,\mathbf{E}\Bigg[\frac{1}{|\mathcal{S}^*|}\Bigg]=o(1).$$

Lastly, combining the previous two results,

$$\begin{aligned}
&\left| B - \mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^2} \sum_{i\in\mathcal{S}^*} \sum_{j\neq i: l_W(j)\neq i, l_W(i)\neq j, l_W(j)\neq l_W(i)} V_i^k V_{l_W(i)}^m V_j^k V_{l_W(j)}^m \right] \right| \\
&\leq \left| B - \mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^2} \sum_{i\in\mathcal{S}^*} \sum_{j\neq i} \mathbf{E}[V_i^k|W_i]\, \mathbf{E}[V_i^m|W_i]\, \mathbf{E}[V_j^k|W_j]\, \mathbf{E}[V_j^m|W_j] \right] \right| + o(1) \\
&= \mathbf{E}\left[\frac{1}{|\mathcal{S}^*|^2} \sum_{i\in\mathcal{S}^*} (\mathbf{E}[V_i^k|W_i]\, \mathbf{E}[V_i^m|W_i])^2 \right] \leq c_k^2 c_m^2\, \mathbf{E}\left[\frac{1}{|\mathcal{S}^*|} \right] = o(1)
\end{aligned}$$

Now, $A + B - 2C = (B + o(1)) + B - 2(B + o(1)) = o(1)$.

□

Lemma 4. *Suppose that $(V_1', W_1), \ldots, (V_N', W_N)$ are i.i.d. where $V_i \equiv (V_{1,i}, \ldots, V_{\dim V, i})$ is a vector and the support of W_i is compact. Moreover, assume that, for any $j, k = 1, \ldots, \dim V$, $\mathbf{E}[V_{j,i}^{r_j} V_{k,i}^{r_k} | W_i = w]$ is Lipschitz with a constant L_{r_j, r_k}, where $0 \leq r_1, r_2 \leq 2$. Define:*

$$\hat{\mathbb{V}} \equiv \frac{1}{2|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} (V_i - V_{l_W(i)})(V_i - V_{l_W(i)})'. \tag{40}$$

Then, as $|\mathcal{S}^*|$ tends to infinity,

$$\hat{\mathbb{V}} - \frac{1}{|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} \mathbf{Var}[V_i|W_i] \to 0. \tag{41}$$

Proof. Let $\hat{\mathbb{V}}_{j,k}$ be the $(j,k)^{\text{th}}$ element of $\hat{\mathbb{V}}$. Then,

$$\begin{aligned}
\hat{\mathbb{V}}_{j,k} &= \frac{1}{2|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} (V_{j,i} - V_{j,l_W(i)})(V_{k,i} - V_{k,l_W(i)}) \\
&= \frac{1}{2|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} V_{j,i} V_{k,i} + \frac{1}{2|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} V_{j,l_W(i)} V_{k,l_W(i)} - \left(\frac{1}{2|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} V_{j,i} V_{k,l_W(i)} + \frac{1}{2|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} V_{j,l_W(i)} V_{k,i} \right)
\end{aligned}$$

By applying Lemma 3 to $V_{j,i}V_{k,i}$ with $(k,m) = (1,0)$ and $(k,m) = (0,1)$, we have

$$\frac{1}{|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} V_{j,i} V_{k,i} - \frac{1}{|\mathcal{S}^*|} \sum_{i\in\mathcal{S}^*} \mathbf{E}[V_{j,i} V_{k,i} | W_i] \to 0 \text{ and} \tag{42}$$

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_{j,l_W(i)}V_{k,l_W(i)} - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_{j,i}V_{k,i}|W_i] \to 0. \tag{43}$$

Note that

$$\begin{aligned}&\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_{j,i}V_{k,l_W(i)} + \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_{j,l_W(i)}V_{k,i}\\ &= \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(V_{j,i}+V_{k,i})(V_{j,l_W(i)}+V_{k,l_W(i)}) - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_{j,i}V_{j,l_W(i)} - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_{k,i}V_{k,l_W(i)}.\end{aligned}$$

Lemma 3 applied to $V_{j,i}+V_{k,i}$, $V_{j,i}$, and $V_{k,i}$ with $(k,m)=(1,1)$ yields

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(V_{j,i}+V_{k,i})(V_{j,l_W(i)}+V_{k,l_W(i)}) - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(\mathbf{E}[V_{j,i}+V_{k,i}|W_i])^2 \to_p 0, \tag{44}$$

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_{j,i}V_{j,l_W(i)} - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(\mathbf{E}[V_{j,i}|W_i])^2 \to_p 0, \text{ and} \tag{45}$$

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*} V_{k,i}V_{k,l_W(i)} - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(\mathbf{E}[V_{k,i}|W_i])^2 \to_p 0. \tag{46}$$

Combining results,

$$\begin{aligned}&\hat{\mathbb{V}}_{j,k} - \Bigg(\frac{1}{2|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_{j,i}V_{k,i}|W_i] + \frac{1}{2|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{E}[V_{j,i}V_{k,i}|W_i]\\ &\qquad - \bigg(\frac{1}{2|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(\mathbf{E}[V_{j,i}+V_{k,i}|W_i])^2 - \frac{1}{2|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(\mathbf{E}[V_{j,i}|W_i])^2 - \frac{1}{2|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(\mathbf{E}[V_{k,i}|W_i])^2\bigg)\Bigg)\\ &= \hat{\mathbb{V}}_{j,k} - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(\mathbf{E}[V_{j,i}V_{k,i}|W_i] - \mathbf{E}[V_{j,i}|W_i]\,\mathbf{E}[V_{k,i}|W_i]) \to_p 0.\end{aligned}$$

□

Let $\mathcal{W} \equiv (\mathcal{X}^1\times\{1\})\cup(\mathcal{X}^0\times\{0\}) \subseteq \mathbb{R}^p\times\{0,1\}$. First, we show that $\mathcal{W}$ is (sequentially) compact with respect to the metric $\|\|$. Let (x_r,d_r) be a sequence in $\mathcal{W}$. Without loss of generality, we assume that $\{r\in\mathbb{N} : d_r = 1\}$ is infinite. Then, consider the subsequence $(x_r,d_r)_{d_r=1}$ of (x_r,d_r). Since $(x_r)_{d_r=1}$ is a sequence in $\mathcal{X}^1$, which is (sequentially) compact by Assumption 14, there exists a subsequence $(x_l)_{d_l=1}$ of $(x_r)_{d_r=1}$ such that $\rho(x_l,x^*) =$

$\|(x_l, d_l) - (x^*, 1)\| \to 0$ for some $x^* \in \mathcal{X}^1$. Thus, (x_r, d_r) has a convergent subsequence in $\mathcal{W}$.

Let $V_i \equiv Z_i(Y_i - Z_i'\theta^*)$ and $W_i \equiv (X_i, D_i)$. First, by Assumptions 14–15, $\mathbf{E}[V_{j,i}^{r_j} V_{k,i}^{r_k} | W_i = w]$ is Lipschitz with respect to $\|\|$. Then, by the compactness of $\mathcal{W}$, Lemma 4 yield

$$\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(V_i - V_{l_W(i)})(V_i - V_{l_W(i)})' - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\mathbf{Var}[V_i|W_i] \to_p 0. \tag{47}$$

Now, let $V_i(\theta) \equiv Z_i(Y_i - Z_i'\theta)$. Since $V_i(\theta) - V_i = Z_iZ_i'(\theta^* - \theta)$,

$$\begin{aligned}
&\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(V_i(\theta) - V_{l_W(i)}(\theta))(V_i(\theta) - V_{l_W(i)}(\theta))' - \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(V_i - V_{l_W(i)})(V_i - V_{l_W(i)})' \\
&= \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(Z_iZ_i' - Z_{l_W(i)}Z_{l_W(i)}')(\theta^* - \theta)(V_i - V_{l_W(i)})' \\
&\quad + \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(V_i - V_{l_W(i)})(\theta^* - \theta)'(Z_iZ_i' - Z_{l_W(i)}Z_{l_W(i)}')' \\
&\quad + \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}(Z_iZ_i' - Z_{l_W(i)}Z_{l_W(i)}')(\theta^* - \theta)(\theta^* - \theta)'(Z_iZ_i' - Z_{l_W(i)}Z_{l_W(i)}')',
\end{aligned}$$

The first two terms are bounded by

$$\begin{aligned}
&\|\theta^* - \theta\|\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_iZ_i' - Z_{l_W(i)}Z_{l_W(i)}'\|\|V_i - V_{l_W(i)}\| \\
&\le \|\theta^* - \theta\|\Bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_iZ_i'\|\|V_i\| + \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_{l_W(i)}Z_{l_W(i)}'\|\|V_i\| \\
&\qquad\qquad + \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_iZ_i'\|\|V_{l_W(i)}\| + \frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_{l_W(i)}Z_{l_W(i)}'\|\|V_{l_W(i)}\|\Bigg) \\
&\le \|\theta^* - \theta\|\Bigg(\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|V_i\|^2\bigg)^{\frac{1}{2}} + \bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|V_{l_W(i)}\|^2\bigg)^{\frac{1}{2}}\Bigg) \\
&\qquad\qquad \times \Bigg(\bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_iZ_i'\|^2\bigg)^{\frac{1}{2}} + \bigg(\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_{l_W(i)}Z_{l_W(i)}'\|^2\bigg)^{\frac{1}{2}}\Bigg) = \|\theta^* - \theta\|O_p(1),
\end{aligned}$$

where we use Lemma 3 in the last inequality. Likewise, the third term is bounded by

$$\|\theta^* - \theta\|^2\frac{1}{|\mathcal{S}^*|}\sum_{i\in\mathcal{S}^*}\|Z_iZ_i' - Z_{l_W(i)}Z_{l_W(i)}'\|^2$$

$$\leq \|\theta^* - \theta\|^2 \bigg(\frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} \|Z_i Z_i'\| \|Z_i Z_i'\| + \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} \|Z_{l_W(i)} Z_{l_W(i)}'\| \|Z_i Z_i'\|$$

$$+ \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} \|Z_i Z_i'\| \|Z_{l_W(i)} Z_{l_W(i)}'\| + \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} \|Z_{l_W(i)} Z_{l_W(i)}'\| \|Z_{l_W(i)} Z_{l_W(i)}'\| \bigg) = \|\theta^* - \theta\|^2 O_p(1).$$

Hence,

$$\bigg\| \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} (V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))(V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))' - \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} (V_i - V_{l_W(i)})(V_i - V_{l_W(i)})' \bigg\|$$

$$\leq \|\theta^* - \hat{\theta}^*\| O_p(1) = o_p(1),$$

where in the last equation we use $\hat{\theta}^* - \theta^* = (\hat{\theta}^* - \theta_{\mathbb{G}^*,F}) + (\theta_{\mathbb{G}^*,F} - \theta^*) = o_p(1)$, which holds by Proposition 2 and Assumption 13. Combined with equation (47), we then have

$$\frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} (V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))(V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))' - \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} \mathbf{Var}[V_i | W_i] \to_p 0.$$

Because $\mathbf{Var}[V_i|W_i] = Z_i \, \mathbf{Var}[Y_i - Z_i'\theta^* | X_i, D_i] Z_i' = Z_i Z_i' \, \mathbf{E}[U_i^2 | X_i, D_i]$, it follows that

$$\bigg\| \bigg(\frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \bigg)^{-1} \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} (V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))(V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))' \bigg(\frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \bigg)^{-1}$$

$$- \bigg(\frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \bigg)^{-1} \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \, \mathbf{E}[U_i^2 | X_i, D_i] \bigg(\frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \bigg)^{-1} \bigg\|$$

$$\leq \bigg\| \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \bigg\|^{-2} \bigg\| \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} (V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))(V_i(\hat{\theta}^*) - V_{l_W(i)}(\hat{\theta}^*))'$$

$$- \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} Z_i Z_i' \, \mathbf{E}[U_i^2 | X_i, D_i] \bigg\| = O_p(1) o_p(1)$$

where we use Assumption 10 in the last equation.

B Supplementary materials

B.1 Extension of Proposition 1

Let L be a class of functions with arguments (x, d), where $x \in \mathbb{R}^p$ and $d \in \{0, 1\}$.

Assumption 16. *There exists $l \in L$ such that $l(X, D)$ has a finite first moment. Furthermore, there exists a random variable E with a finite first moment such that $\mathbf{E}[DE] = 0$ and $Y = l(X, D) + E$. Finally, $\mathbf{Pr}[D = 1] \in (0, 1)$.*

For notational convenience, let $\Delta l(X, \cdot) \equiv l(X, 1) - l(X, 0)$ and $\Delta\,\mathbf{E}[l(X, 0)|D = \cdot] \equiv \mathbf{E}[l(X, 0)|D = 1] - \mathbf{E}[l(X, 0)|D = 0]$.

Proposition 5. *Suppose that Assumption 16 holds. Then,*

$$\mathbf{E}[\Delta l(X, \cdot)|D = 1] - \tau = \int (f(x, 0) - l(x, 0))(g^1(x) - g^0(x))\mu(dx). \tag{48}$$

Consider an extension of (3)

$$Y = \underbrace{\alpha + \beta D + s(X)'\gamma + D(s(X) - \mathbf{E}[s(X)|D = 1])'\delta}_{l(X,D)} + E, \tag{49}$$

which additionally includes the interaction terms between D and $s(X)$. Note that $\Delta l(X, \cdot) = \beta + (s(X) - \mathbf{E}[s(X)|D = 1])$, and thus $\mathbf{E}[\Delta l(X, \cdot)|D = 1] = \beta$.

B.1.1 Proof of Proposition 5

Lemma 5. *Suppose that Assumption 16 holds. Then,*

$$\mathbf{E}[Y|D] = \mathbf{E}[l(X, 0)|D = 0] + (\mathbf{E}[\Delta l(X, \cdot)|D = 1] + \Delta\,\mathbf{E}[l(X, 0)|D = \cdot])D. \tag{50}$$

Proof. Let $a \equiv \mathbf{E}[l(X, D)|D = 0]$ and $b \equiv \mathbf{E}[l(X, D)|D = 1] - \mathbf{E}[l(X, D)|D = 0]$, both of which exist and finite. Define $\tilde{E} \equiv l(X, D) - (a + bD)$. Then, $\mathbf{E}[D\tilde{E}] = \mathbf{E}[Dl(X, D)] - (a +$

b) $\mathbf{Pr}[D=1]=0$. Thus, $\mathbf{E}[l(X,D)|D]=a+bD$. Also, since $\mathbf{E}[DE]=0$, $\mathbf{E}[E|D]=0$.

Combining results,

$$\begin{aligned}\mathbf{E}[Y|D] &= \mathbf{E}[l(X,D)|D]+\mathbf{E}[E|D]\\ &= \mathbf{E}[l(X,D)|D=0]+(\mathbf{E}[l(X,D)|D=1]-\mathbf{E}[l(X,D)|D=0])D,\end{aligned}$$

and by telescoping $\mathbf{E}[l(X,0)|D=1]$ in the coefficient of D, we obtain the desired result. □

Now,

$$\begin{aligned}\tau &= \mathbf{E}[Y|D=1]-\int \mathbf{E}[Y|X=x,D=0]G^1(dx)\\ &= \mathbf{E}[Y|D=1]-\mathbf{E}[Y|D=0]-\int \mathbf{E}[Y|X=x,D=0](G^1-G^0)(dx)\\ &= \mathbf{E}[\Delta l(X,\cdot)|D=1]+\Delta\,\mathbf{E}[l(X,0)|D=\cdot]-\int \mathbf{E}[Y|X=x,D=0](G^1-G^0)(dx)\\ &= \mathbf{E}[\Delta l(X,\cdot)|D=1]-\int (\mathbf{E}[Y|X=x,D=0]-l(x,0))(G^1-G^0)(dx),\end{aligned}$$

where we use Lemma 5 in the third equality.

B.2 Proofs of selected equations

B.2.1 Proof of equation (5)

Let $(a,b)\equiv \operatorname{argmin}_{(\tilde{a},\tilde{b})}\mathbf{E}[(D-(\tilde{a}+s(X)'\tilde{b}))^2]$ be the linear projection coefficient of $(1,s(X)')$ obtained by regressing D on $(1,s(X))$ in the population. Define $\tilde{D}\equiv D-(a+s(X)'b)$. We reformulate the linear regression model as

$$\begin{aligned}Y &= \alpha+\beta D+s(X)'\gamma+E \qquad (51)\\ &= \alpha+\beta(a+s(X)'b+\tilde{D})+s(X)'\gamma+E \qquad (52)\\ &= \beta\tilde{D}+\underbrace{(\alpha+\beta a)+s(X)'(\beta b+\gamma)+E}_{\equiv\tilde{E}}. \qquad (53)\end{aligned}$$

Suppose that $\mathbf{E}[\tilde{D}^2] = 0$. Then, $D = a + s(X)'b$ almost surely, and thus

$$Z = \begin{pmatrix} 1 \\ D \\ s(X) \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ a & b' \\ 0 & I \end{pmatrix}}_{M} \begin{pmatrix} 1 \\ s(X) \end{pmatrix} \text{ almost surely.} \tag{54}$$

Here, the rank of M is $\dim \theta - 1$. Hence, the rank of $\mathbf{E}[ZZ']$ cannot be larger than $\dim \theta - 1$, which violates Assumption 1. From contraposition, $\mathbf{E}[\tilde{D}^2] > 0$ then follows.

Furthermore,

$$\mathbf{E}[\tilde{D}\tilde{E}] = \beta \underbrace{\mathbf{E}[\tilde{D}(1 \ s(X)')]}_{=0 \text{ by the definition of linear projection}} \begin{pmatrix} \alpha + \beta a \\ \beta b + \gamma \end{pmatrix} + \beta(-a, 1 - b') \underbrace{\mathbf{E}[ZE]}_{=0 \text{ by Assumption 1}} = 0. \tag{55}$$

Combining results, $(\mathbf{E}[\tilde{D}^2])^{-1} \mathbf{E}[\tilde{D}Y] = \beta + (\mathbf{E}[\tilde{D}^2])^{-1} \mathbf{E}[\tilde{D}\tilde{E}] = \beta$.

B.2.2 Proof of equation (22)

By Assumption 1, $\mathbf{E}_{G,F}[\|ZY\|] \leq \mathbf{E}_G[\|Z\|^2]\|\theta\| + \mathbf{E}_{G,F}[\|ZE\|] < \infty$. Thus, $\mathbf{E}_F[ZY|X, D]$ is finite almost surely (a.s.), and so is $|\mathcal{S}^*|^{-1} \sum_{i \in \mathcal{S}^*} \mathbf{E}_F[ZY|X = X_i, D = D_i] = \mathbf{E}_{\mathbb{G}^*}[\mathbf{E}_F[ZY|X, D]]$. Then, since $\mathbf{E}_{\mathbb{G}^*}[ZZ'] = |\mathcal{S}^*|^{-1} \sum_{i \in \mathcal{S}^*} Z_i Z_i'$ is positive definite, $\theta_{\mathbb{G}^*,F}$ is finite a.s.

Let $\tilde{E}_i \equiv Y_i - Z_i' \theta_{\mathbb{G}^*,F}$. We show that Z and $\tilde{E}$ satisfy Assumption 1 for the population $(\mathbb{G}^*, F)$ a.s. First, since $\mathcal{S}^*$ is finite, so is $\mathbf{E}_{\mathbb{G}^*}[\|Z\|^2] = |\mathcal{S}^*|^{-1} \sum_{i \in \mathcal{S}^*} \|Z_i\|^2$. Second,

$$\begin{aligned}\mathbf{E}_{\mathbb{G}^*,F}[\|Z\tilde{E}\|] &\leq \mathbf{E}_{\mathbb{G}^*}[\mathbf{E}_F[\|ZY\||X, D]] + \mathbf{E}_{\mathbb{G}^*}[\|Z\|^2]\|\theta_{\mathbb{G}^*,F}\| \\ &= \frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} \mathbf{E}_F[\|ZY\||X = X_i, D = D_i] + \left(\frac{1}{|\mathcal{S}^*|} \sum_{i \in \mathcal{S}^*} \|Z_i\|^2 \right) \|\theta_{\mathbb{G}^*,F}\|\end{aligned}$$

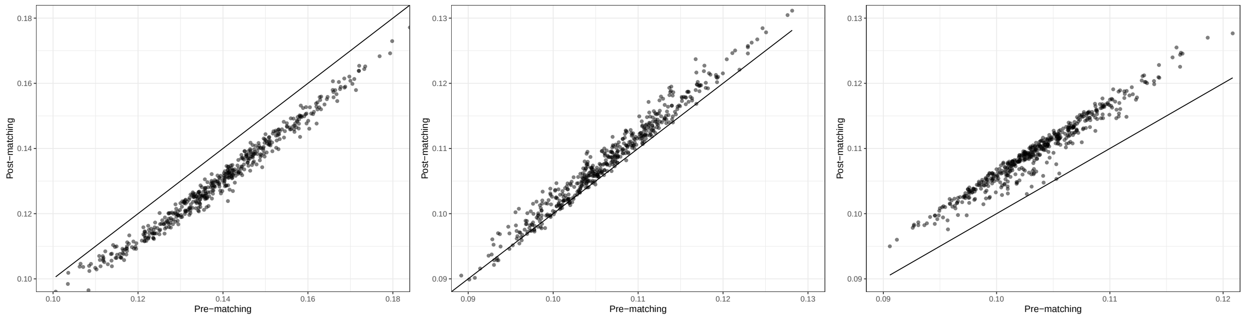
is finite a.s., by combining previous results. Third, $\mathbf{E}_{\mathbb{G}^*}[ZZ']$ is positive definite by assumption. Lastly, $\mathbf{E}_{\mathbb{G}^*,F}[ZE] = \mathbf{E}_{\mathbb{G}^*,F}[ZY] - \mathbf{E}_{\mathbb{G}^*}[ZZ'](\mathbf{E}_{\mathbb{G}^*}[ZZ'])^{-1} \mathbf{E}_{\mathbb{G}^*,F}[ZY] = 0$ a.s.

By Assumption 2, the support of $\mathbb{G}^{1*}$, which is a subset of $\mathcal{X}^1$, is contained in that of

F^0. Assumption 3 is automatically satisfied for $(\mathbb{G}^*, F)$.

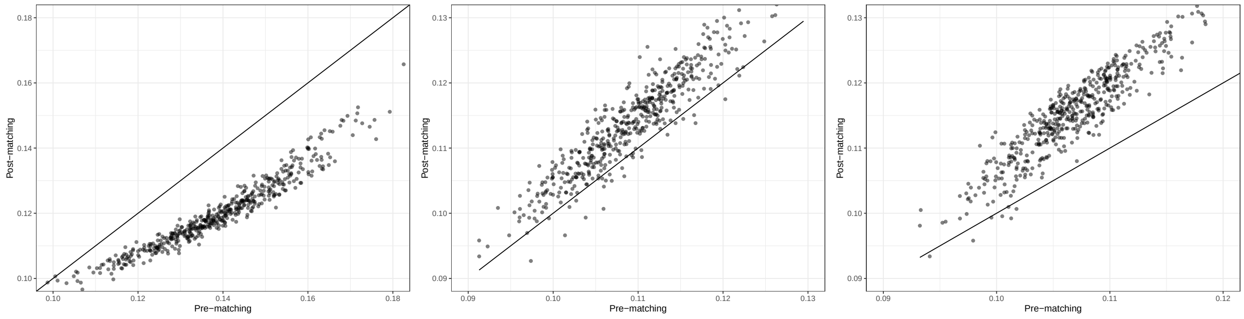
B.3 Omitted simulation results

In this subsection, we present the simulation results for the omitted cases where $(N_1, N_0) \in \{(50, 75), (50, 125)\}$. These results demonstrate that an increase in the ratio N_0/N_1 leads to improved covariate balance, consequently reducing the bias of regressions.



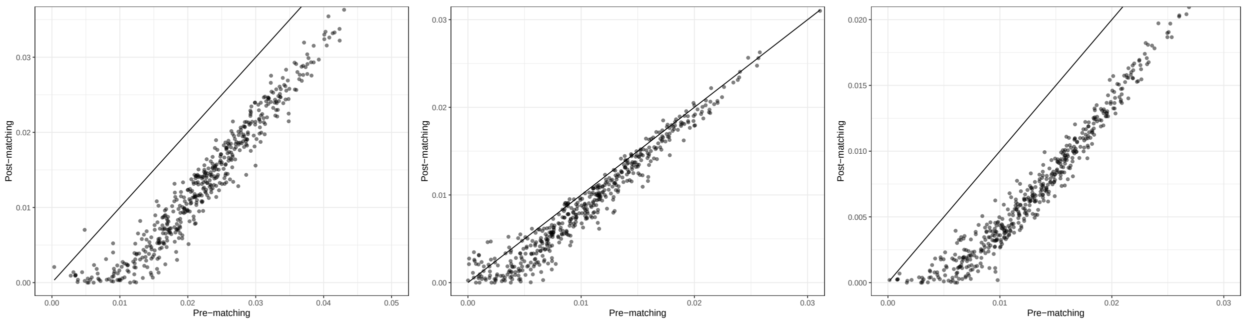
Specification A Specification B Specification C

Figure 6: Pre- and Post-estimands ($N_1 = 50$, $N_0 = 75$)



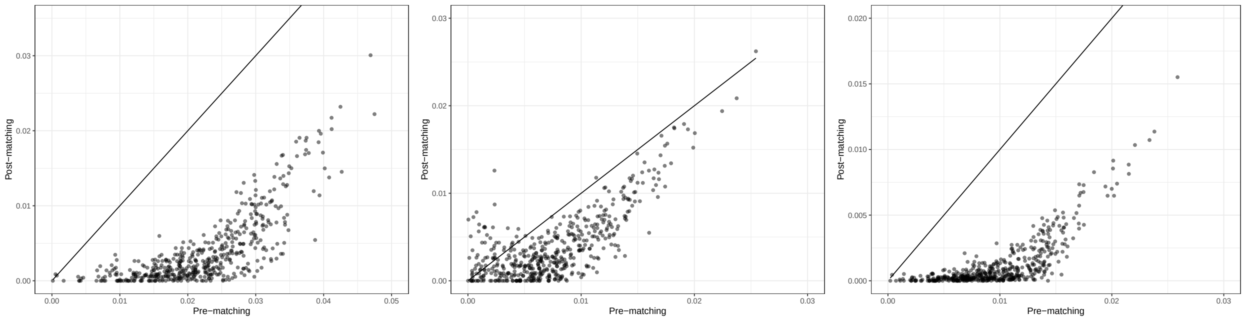
Specification A Specification B Specification C

Figure 7: Pre- and Post-estimands ($N_1 = 50$, $N_0 = 125$)



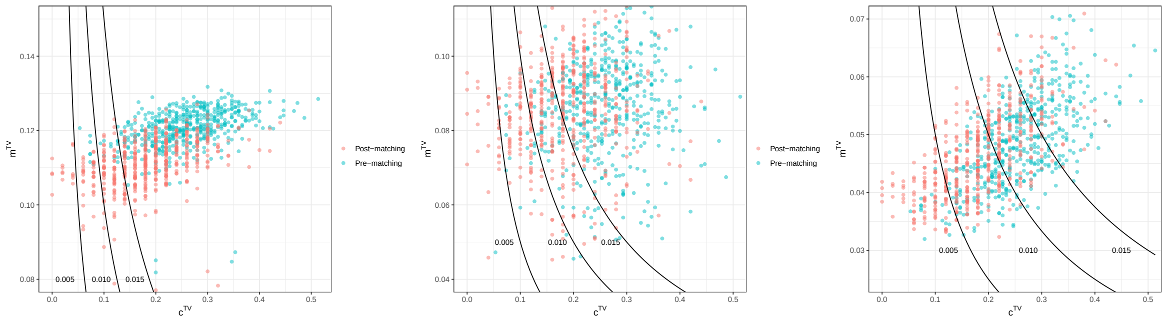
Specification A Specification B Specification C

Figure 8: Pre- and Post-biases ($N_1 = 50$, $N_0 = 75$)



Specification A Specification B Specification C

Figure 9: Pre- and Post-biases ($N_1 = 50$, $N_0 = 125$)

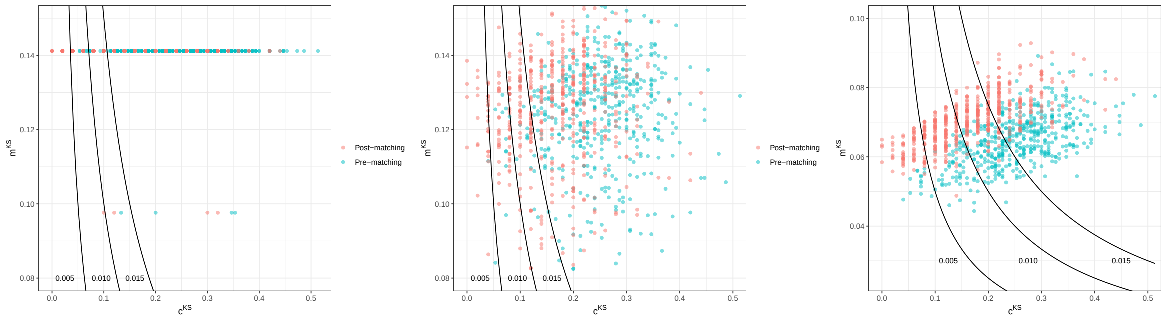


Specification A

Specification B

Specification C

Figure 10: Total variation bound ($N_1 = 50$, $N_0 = 75$)

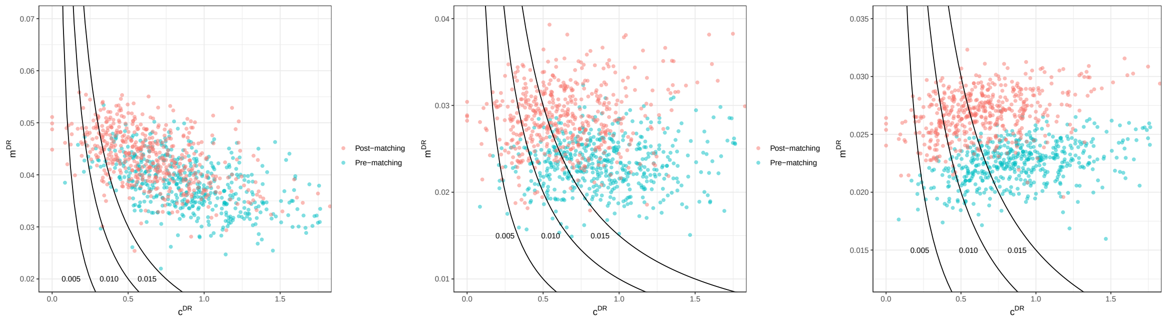


Specification A

Specification B

Specification C

Figure 11: Kolmogorov-Smirnov bound ($N_1 = 50$, $N_0 = 75$)

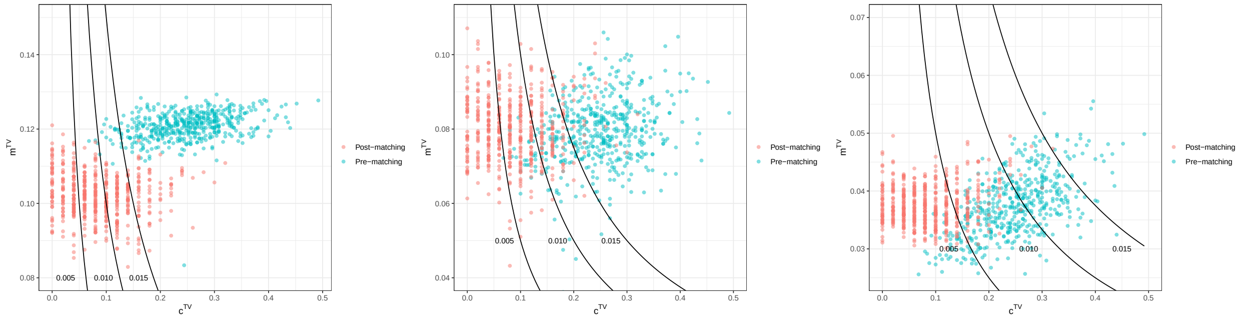


Specification A

Specification B

Specification C

Figure 12: Density ratio bound ($N_1 = 50$, $N_0 = 75$)

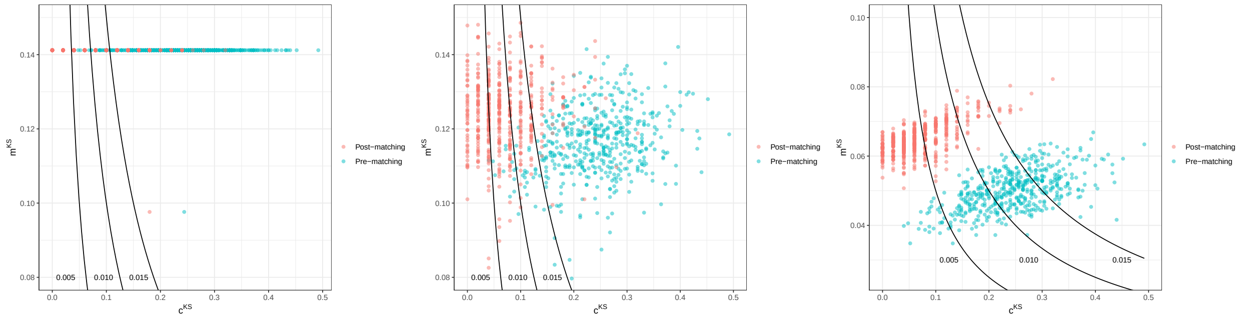


Specification A

Specification B

Specification C

Figure 13: Total variation bound ($N_1 = 50$, $N_0 = 125$)

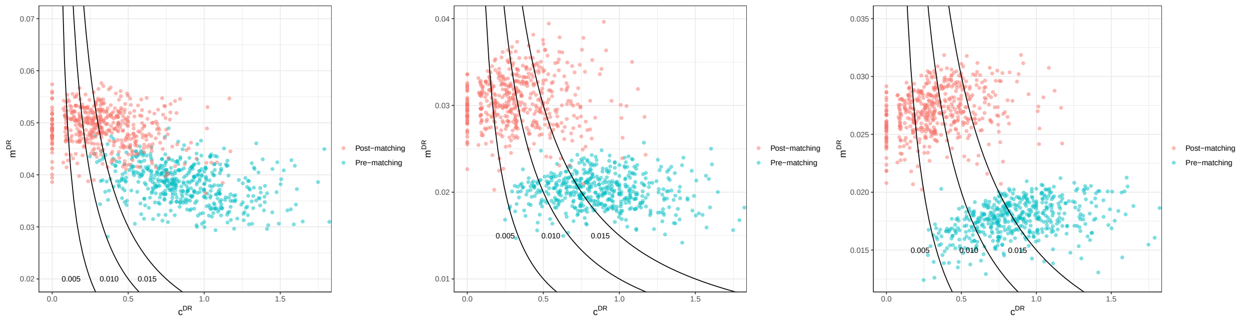


Specification A

Specification B

Specification C

Figure 14: Kolmogorov-Smirnov bound ($N_1 = 50$, $N_0 = 125$)



Specification A

Specification B

Specification C

Figure 15: Density ratio bound ($N_1 = 50$, $N_0 = 125$)

국문 초록

“설계 단계”란 연구자가 추론의 강건성과 신뢰성을 위해 통제 변수 분포가 처치 집단별로 균형 잡힌 하위 표본을 구축하는 관측 데이터 연구의 한 단계를 일컫는다. 본 논문은 선형회귀분석에서의 이러한 설계 단계의 역할을 탐구하고 그것을 정당화한다. 이를 위해 우선 설계 단계는 추정 대상을 조정하는 하위 표본의 선택 과정으로서 규정된다. 그 다음, 통제 변수 분포의 균형은 그 선택 과정의 타당한 기준으로 정당화되는데, 이는 주어진 하위 표본에 대해 통제 변수 분포의 균형이 추정 대상과 관심 모수 간 차이를 목표 정밀도 이내로 제한하고자 할 때 최대 허용될 수 있는 모형 설정 오류의 정도를 말해주기 때문이다. 따라서, 통제 변수 분포가 균형 잡힌 하위 표본을 구축하는 작업은 모형 설정 오류로 인한 편향에 강건한 추정 대상을 탐색하는 과정으로서 이해될 수 있다.