Chapter 12 Unconfounded Treatment Assignment

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

August 19, 2021

Introduction

Now we leave the conceptually straightforward world of ideal randomized experiments.

Although in simple situations with observational data we can still directly apply the tools from randomized experiments and exploit the exact results that accompany them, quickly we will be forced to make approximations in our inferences.

No longer will estimators be exactly unbiased as in Chapter 6, nor will we be able to calculate exact p-values of the type considered in Chapter 5.

The first step towards addressing observational studies is to relax the assumption that the probability of treatment assignment is a known function.

Introduction

We maintain the *unconfoundedness* assumption that states that assignment is free from dependence on the potential outcomes.

We will continue to assume that the assignment mechanism is *individualistic*, so that the probability for unit i is essentially a function of the pre-treatment variables for unit i only.

We also maintain the assumption that the assignment mechanism is *probabilistic*.

The implication of these assumptions is that the assignment mechanism can be interpreted as if, within subpopulations of units with the same value for the covariates, a CRE was conducted, although an experiment with unknown assignment probabilities for the units.

Introduction

Although we do not know *a priori* the assignment probabilities for each of these units, we know these probabilities are identical because their covariate values are identical.

Hence, conditional on the number of treated and control units comprising such a subpopulation, the probability of receiving the treatment is equal to

$N_t(x)/(N_c(x) + N_t(x))$ for all units with $X_i = x$;

where $N_t(x)$ and $N_c(x)$ are the number of units in the control and treatment groups respectively with pretreatment value $X_i = x$.

In practice, this is of limited value, as typically $N_c(x)$ or $N_t(x)$ will be equal to zero in some strata.

However, this insight has an important implication that suggests feasible alternatives

Introduction

We start by discussing general aspects of the unconfoundedness assumption (UA), the broad strategies we recommend in settings where unconfoundedness is viewed as an appropriate assumption.

We then explore a particular implication of unconfoundedness related to the *propensity score*. Even if a large set of covariates is used to ensure unconfoundedness, it is generally sufficient, in a certain sense, to adjust for a scalar function of the covariates, namely the propensity score.

Next, we outline broad strategies for inference under regular assignment mechanisms.

Finally, we will discuss preliminary analyses not involving the outcome data, that we recommend as part of the *design* stage and outline how, in some settings, one can do additional analyses that help the researcher assess the plausibility of the UA.

Regular Assignment Mechanisms

As discussed in Chapter 3, a regular assignment mechanism satisfies three conditions.

(1) *Probabilistic*: $0 < p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1, \text{for } i = 1, \ldots, N.$

(2) *Individualistic*: (i) $p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q(X_i, Y_i(0), Y_i(1)), \text{for } i = 1, \ldots, N$, and (ii)

$$\Pr(\mathbf{W} \,|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} q(X_i, Y_i(0), Y_i(1))^{W_i} \cdot (1 - q(X_i, Y_i(0), Y_i(1)))^{1-W_i},$$

for some constant c, for $\mathbf{W} \in \mathbb{W}^+$, and zero elsewhere.

(3) *Unconfounded*: all the assignment probabilities $\Pr(\mathbf{W} \,|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ are free from dependence on the potential outcomes.

Regular Assignment Mechanisms

In combination with individualistic assignment, (3) implies that we can write the assignment mechanism as

$$\Pr\left(\mathbf{W} \,|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right) = c \cdot \prod_{i=1}^{N} e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i},$$

where $e(x)$ is the *propensity score*.

Under these assumptions we can give a causal interpretation to the comparison of observed outcomes for treated and control units within subpopulations defined by values of the pretreatment variables, $X_i = x$.

For example, the difference in average observed outcomes is unbiased for the average effect of the treatment at $X_i = x$.

Regular Assignment Mechanisms

Consider the case with a single binary covariate, e.g., sex, so that $X_i \in \{f, m\}$. Within the subsamples of women and men, the average finite sample treatment effects are:

$$\tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:X_i=f} \Big(Y_i(1) - Y_i(0) \Big), \qquad \text{and} \quad \tau_{\text{fs}}(m) = \frac{1}{N(m)} \sum_{i:X_i=m} \Big(Y_i(1) - Y_i(0) \Big),$$

where $N(f)$ and $N(m)$ are the number of women and men, respectively, in the sample.

Within each of these subsamples, estimation and inference are entirely standard.

The fact that we do not know *a priori* the probability of assignment to the treatment is irrelevant. We can use the results for the analysis of CRE by conditioning on the number of treated women and treated men.

Regular Assignment Mechanisms

If we instead are interested in the overall average effect

$$\tau_{\text{fs}} = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m),$$

we can simply use the methods for SRE.

This approach of partitioning the population into strata by values of the pre-treatment variables extends, in principle, to all settings with discrete-valued pre-treatment variables.

However, with pre-treatment variables taking on many distinct values in the sample, there may be a substantial number of strata with only treated or with only control units. For such strata, we cannot estimate the stratum-specific treatment effects.

Regular Assignment Mechanisms

This setting is of great practical relevance, and indeed of much of the theoretical literature on estimation of, and inference for, causal effects in statistics and related disciplines.

In this case, we compare outcomes for treated and control units with "similar' values for the pre-treatment variables.

For such comparisons to be appropriate, we require smoothness and modeling assumptions, and decisions regarding tradeoffs between differences in one covariate versus another.

These 'modeling challenges', are central topics in Parts III and IV of the book. Beyond depending on substantive thematic insights related to the assessments of the UA, evaluating the various approaches to estimation and inference also requires statistical expertise.

A Super-population Perspective

For the purpose of discussing properties of various frequentist approaches it is useful to take a super-population perspective.

Moreover, it is helpful to view the covariates X_i as having been randomly drawn from an approximately continuous distribution.

If, instead, we view the covariates as having a discrete distribution with finite support there will be, in large samples, both treated and control units with the exact same values of the covariates.

In this way we can immediately remove all biases under the UA arising from differences between covariates, and many adjustment methods will give similar, or even identical, answers.

Regular Assignment Mechanisms

However, in practice it is not feasible to stratify fully on all covariates, because too many strata would have only a single unit.

The differences between various adjustment methods arise precisely in such settings where it is not feasible to stratify on all values of the covariates.

The differences in properties are most easily analyzed in settings with random samples from large populations using effectively continuous distributions for the covariates.

In the super-population,unconfoundedness implies a restriction on the joint distribution of $(Y_i(0), Y_i(1), W_i, X_i)$, namely

$$\Pr(W_i = 1|Y_i(0), Y_i(1), X_i) = \Pr(W_i = 1|X_i) = e(X_i), \quad (1)$$

Regular Assignment Mechanisms

or, in the Dawid (1979) conditional independence notation,

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, X_i,$$

Probabilistic assignment now requires that

$$0 < e(x) < 1,$$

for all x in the support of X_i, where we ignore measure-theoretic details.

Unconfoundedness is not Testable

A key feature of the UA is that it has no directly testable implications.

There is no information in the data that can tell us that UA does not hold.

To gain further insight into this feature of the UA, it is useful to look at this assumption in a setting with a large sample, where we can estimate the joint distribution of $(Y_i^{\text{obs}}, W_i, X_i)$.

Unconfoundedness is not Testable

Theorem

(Super-population Unconfoundedness) *Super-population unconfoundedness implies two restrictions on the conditional distributions of the potential outcomes. First,*

$$\left(Y_i(0) \,\middle|\, W_i = 1, X_i\right) \sim \left(Y_i(0) \,\middle|\, W_i = 0, X_i\right), \qquad \text{for } i = 1, \ldots, N, \tag{2}$$

and, second,

$$\left(Y_i(1) \,\middle|\, W_i = 0, X_i\right) \sim \left(Y_i(1) \,\middle|\, W_i = 1, X_i\right), \qquad \text{for } i = 1, \ldots, N. \tag{3}$$

(Here "$\sim$" denotes equality in distribution.)

Proof: By super-population unconfoundedness, defined in Chapter 3, Section 10, W_i is independent of $(Y_i(0), Y_i(1))$ given X_i. Hence $Y_i(0)$ is independent of W_i given X_i, implying the first claim in Theorem 1. The second claim follows by an analogous argument. □

Unconfoundedness is not Testable

It is useful to restate (2) and (3), in terms of missing and observed outcomes:

$$\left(Y_i^{\text{mis}} \;\middle|\; W_i = 1, X_i\right) \;\sim\; \left(Y_i^{\text{obs}} \;\middle|\; W_i = 0, X_i\right), \qquad \text{for } i = 1, \ldots, N.$$

Thus, the UA implies the equality of the distribution of Y_i^{mis} to the distribution of Y_i^{obs}.

In large samples we can infer the conditional distribution of Y_i^{obs} given W_i and X_i, but no amount of observable data will allow us to infer the conditional distribution of Y_i^{mis}.

Although UA is not testable, there are in some cases analyses one may be able carry out that assist the researcher when assessing the plausibility of this critical assumption (to be discussed later).

Why is Unconfoundedness an Important Assumption?

Of the three assumptions required for regularity of the assignment mechanism, probabilistic assignment is easiest to motivate.

If a particular subpopulation has zero probability of being in one of the treatment groups, then estimates of treatment effects for this subpopulation must, by necessity, rely on extrapolation.

For example, suppose we are interested in evaluating a new drug, and suppose the sample studied contains both women and men, $X_i \in \{f, m\}$.

However, suppose that the treatment group contains only women. In that case it would clearly require strong assumptions to estimate the effect for the entire population.

It would appear more reasonable to estimate the effect for women, and then separately discuss the plausibility of extrapolating that estimate for women to men.

Why is Unconfoundedness an Important Assumption?

Even more prevalent is the case where the probabilistic assumption is close to being violated, without the probabilities being exactly equal to zero or one, which can severely impact our ability to obtain precise estimates of the causal estimands.

This raises a number of issues, which is discuss in detail in Chapters 15 and 16.

The second assumption, individualistic assignment, is rarely controversial. Although formally it is possible that there is dependence in the assignment indicators beyond that allowed through, for example, stratification on covariates, there are no practical examples we are aware of, other than sequential assignment mechanisms, where this is plausibly violated.

Why is Unconfoundedness an Important Assumption?

The typically most controversial component is the UA.

The assumption is extremely widely used. By a wide margin, most analyses involving observational studies fundamentally rely on unconfoundedness, often implicitly, and often in combination with other assumptions, in order to estimate causal effects.

In many empirical studies in social sciences, causal effects are estimated through linear regression, where, typically it is implicitly assumed that in the super-population,

$$\mathbb{E}\left[\, Y_i(w) | \, X_i \right] = \alpha + \tau \cdot w + X_i \beta,$$

for some values of the three unknown parameters α, τ, and β.

Why is Unconfoundedness an Important Assumption?

Defining $\varepsilon_i = Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta$, so that we can write

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon_i, \tag{4}$$

it is then assumed that

$$\varepsilon_i \perp\!\!\!\perp W_i, X_i.$$

This assumption is often referred to as *exogeneity* of the treatment (and the pretreatment variables) in the econometrics literature.

The regression function (4) is interpreted as a causal relation, in our sense of the term "causal", namely that if we manipulate the treatment W_i, then the outcome would change in expectation by an amount τ.

Why is Unconfoundedness an Important Assumption?

Hence, in the potential outcome formulation, we have

$$Y_i(0) = \alpha + X_i\beta + \varepsilon_i, \qquad \text{and} \quad Y_i(1) = Y_i(0) + \tau.$$

Then, because ε_i is a function of $Y_i(0)$ and X_i given the parameters,

$$\Pr(W_i = 1 | Y_i(0), Y_i(1), X_i) = \Pr(W_i | \varepsilon_i, X_i),$$

and by exogeneity of the treatment indicator, we have

$$\Pr(W_i | \varepsilon_i, X_i) = \Pr(W_i | X_i),$$

and thus unconfoundedness holds.

The exogeneity assumption combines unconfoundedness with functional form and constant treatment effect assumptions that are quite strong, and arguably unnecessary. Therefore we focus here on the cleaner, functional-form-free UA.

Why is Unconfoundedness an Important Assumption?

A second motivation for the UA is based on a comparison with alternative assumptions.

Unconfoundedness implies that one should compare units similar in terms of pretreatment variables, that is, one should compare "like with like".

This has intuitive appeal, and underlies many informal, as well as formal, causal inferences.

If we do not believe in the UA in a study one could still conduct a sensitivity analysis, or, in an extreme version, calculate ranges of values for the causal estimands consistent with the data (see Chapter 22).

However, any alternative approach that would provide specific guidance on which treated units to compare with which control units would have to compare units that differ in terms of observed pretreatment variables.

Why is Unconfoundedness an Important Assumption?

Suppose we are interested in the causal effect of a job training program (JTP).

Now suppose there is a 40-year old man who has been unemployed for 6 months, and who was continuously employed for 18 months prior to that in the automobile industry, with a high school education, who is going through the JTP.

The UA implies that in order to estimate the causal effect of this program for him, we should look for a man with the same pre-training characteristics, who did not go through the JTP.

Any plausible alternative strategy would involve 'looking' for a person, or combination of persons, who did not go through the JTP.

In other words, an alternative to the UA involve 'looking' for a comparison person who is systematically *different* in terms of observed pretreatment variables.

Why is Unconfoundedness an Important Assumption?

Suppose that a researcher is concerned that the UA may be violated, because typically individuals who enrolled in this JTP may be more interested in finding jobs, that is, more motivated, than the individuals who did not enroll.

Suppose that motivation is a permanent characteristic of individuals.

It is plausible that more highly motivated individuals are, typically, better at finding employment conditional on their observed treatment status.

Unconfoundedness may in this case be a reasonable assumption *if* motivation were observed.

If motivation is not observed, however, the implication is that the potential outcomes would be correlated with the treatment indicator, and thus the UA would be violated.

Why is Unconfoundedness an Important Assumption?

However, it is not clear that, in such a scenario, using a control person who *differs* in terms of observed pretreatment characteristics as the comparison would improve the credibility of the causal interpretation.

In order to improve the comparison, one would have to be able to trade off observed pretreatment characteristics against the unobserved motivation, without direct information on the latter. It would appear often difficult to do so in a credible manner.

Thus, the claim is *not* that UA is plausible *per sé*. The claim is that allowing for systematic differences in such pretreatment characteristics is unlikely to improve comparisons in general practice.

Why is Unconfoundedness an Important Assumption?

A third aspect concerns the interpretation of assignment processes that lead to differences in treatment levels for units who are identical in terms of observed pretreatment characteristics.

In observational studies it is, in contrast to the experiment, in general not clear why similar units receive different treatment assignments.

Especially in settings where the units are individuals, and the assignment is based on individual choices, one might be concerned that individuals who look *ex ante* identical, but who make different choices, must be different in unobserved ways (e.g. motivation) that invalidates a causal interpretation of differences in their outcomes.

Why is Unconfoundedness an Important Assumption?

Examples of such settings include those where individuals choose to enroll in JTP and those where medical treatment decisions are made by physicians, in consultation with patients.

However, if the unobserved differences that led the individuals to make different choices, are independent of the potential outcomes, conditional on observed covariates, unconfoundedness still holds.

This may arise, for example, in settings where unobserved differences in terms of the costs or benefits associated with exposure to the treatment are unrelated to the potential outcomes.

Why is Unconfoundedness an Important Assumption?

For instance, suppose two patients with a particular medical condition have identical symptoms and that they share the same physician.

In consultation with these patients, the physician faces the choice between two treatments, say drug A and drug B. Suppose A is more expensive than B.

Furthermore, suppose that as a result of differing health insurance plans, the incremental cost of taking A relative to B is higher for one patient than for the other.

This cost difference may well affect the choice of drug, and as a result one may have data on individuals with similar medical conditions exposed to different treatments without violating the UA (assuming that the choice of insurance plan is not related to outcomes given exposure to drug A or drug B, especially after conditioning on observed covariates such as sex, age *et cetera*).

Selecting Pre-treatment Variables for Conditioning

Variables that are possibly affected by the treatment should not be included in the set of pre-treatment variables, and correctly adjusting for differences in such variables is generally difficult.

Given this set of proper pre-treatment variables, one generally wants to control for as many as possible, or all of them.

For instance, in the evaluation of a JTP on individuals disadvantaged in the labor market, one would like to include detailed labor market histories and individual characteristics of the individuals to eliminate such characteristics as alternative explanations for differences in outcomes between trainees and control individuals.

Selecting Pre-treatment Variables for Conditioning

There are some exceptions to this general advice. In some cases there is additional prior information regarding the dependence of potential outcomes on pre-treatment variables that suggests alternative estimation strategies that do not remove differences in all observed pre-treatment variables. An important case is *instrumental variables* discussed in more detail in Chapters 23-25.

In practice, however, such cases are typically easy to recognize and rarely lead to confusion. Variables that are truly instrumental variables are relatively rare, and when they exist, it is even more rare that they are mistakenly used as covariates to be adjusted for.

Balancing Scores and the Propensity Score

Under unconfoundedness, we can remove all biases in comparisons between treated and control units by adjusting for differences in observed covariates.

Although feasible in principle, in practice this will be difficult to implement with a large number of covariates.

The idea of balancing scores is to find lower-dimensional functions of the covariates that suffice for removing the bias associated with differences in the pretreatment variables.

Balancing Scores and the Propensity Score

Definition

(Balancing Scores)
A balancing score $b(x)$ is a function of the covariates such that

$$W_i \perp\!\!\!\perp X_i \,\Big|\, b(X_i).$$

Balancing scores are not unique. By definition, the vector of covariates X_i itself is a balancing score, and any one-to-one function of a balancing score is also a balancing score. We are most interested in low dimensional balancing scores.

Balancing Scores and the Propensity Score

One scalar balancing scores is the propensity score (or any one-to-one transformation of the propensity score, such as the linearized propensity score or log odds ratio, $\ell(x) = \ln(e(x)/(1 - e(x)))$.

Lemma

(Balancing Property of the Propensity Score)
The propensity score is a balancing score.

Balancing Scores and the Propensity Score

Proof: We show that

$$W_i \perp\!\!\!\perp X_i \,\Big|\, e(X_i),$$

or, equivalently,

$$\Pr(W_i = 1|X_i, e(X_i)) = \Pr(W_i = 1|e(X_i)),$$

implying that W_i is independent of X_i given $e(X_i)$. First, consider the left hand side:

$$\Pr(W_i = 1|X_i, e(X_i)) = \Pr(W_i = 1|X_i) = e(X_i),$$

where the first equality follows because $e(X_i)$ is a function of X_i and the second is by the definition of $e(X_i)$. Second, consider the right hand side. By the definition of probability and iterated expectations,

$$\Pr(W_i = 1|e(X_i)) = \mathbb{E}[W_i|e(X_i)] = \mathbb{E}\left[\mathbb{E}[W_i|X_i, e(X_i)]|e(X_i)\right] = \mathbb{E}[e(X_i)|e(X_i)] = e(X_i).$$

□

Balancing Scores and the Propensity Score

Balancing scores have an important property: if assignment to treatment is unconfounded given the full set of covariates, then assignment is also unconfounded conditioning only on a balancing score:

Lemma

(Unconfoundedness given a Balancing Score) *Suppose assignment to treatment is unconfounded. Then assignment is unconfounded given any balancing score:*

$$W_i \perp\!\!\!\perp Y_i(0), Y_i(1) \;\Big|\; b(X_i).$$

Balancing Scores and the Propensity Score

Proof: We show that

$$\Pr(W_i = 1|Y_i(0), Y_i(1), b(X_i)) = \Pr(W_i = 1|b(X_i)),$$

which is equivalent to the statement in the lemma. By iterated expectations we can write

$$\Pr\left(W_i = 1|Y_i(0), Y_i(1), b(X_i)\right) = \mathbb{E}\left[W_i \,|\, Y_i(0), Y_i(1), b(X_i)\right]$$

$$= \mathbb{E}\Big[\mathbb{E}\left[W_i \,|\, Y_i(0), Y_i(1), X_i, b(X_i)\right]\Big| Y_i(0), Y_i(1), b(X_i)\Big].$$

By unconfoundedness, the inner expectation is equal to $\mathbb{E}\left[W_i \,|X_i, b(X_i)\right]$ and by the definition of balancing scores, this is equal to $\mathbb{E}[W_i|b(X_i)]$. Hence the last expression is equal to

$$\mathbb{E}\Big[\mathbb{E}_W[W_i|b(X_i)]\Big| Y_i(0), Y_i(1), b(X_i)\Big] = \mathbb{E}[W_i|b(X_i)] = \Pr(W_i = 1|b(X_i),$$

which is equal to the righthand side. □

Balancing Scores and the Propensity Score

The first implication of the Lemma is that given a vector of covariates that ensure unconfoundedness, adjustment for balancing scores suffices for removing all biases.

Hence, even if a covariate is associated with the potential outcomes, differences in covariates between treated and control units do not lead to bias because they cancel out by averaging over all units with the same value for the balancing score.

The situation is analogous to that in a CRE where the distribution of covariates is the same in both treatment arms.

Balancing Scores and the Propensity Score

Even though the covariates may differ between specific treated and control units with the same value for the balancing score, they have the same *distribution* of values in the treatment and control group.

Because the propensity score is a balancing score, the Lemma implies that conditional on the propensity score, assignment to treatment is unconfounded. But within the class of balancing scores, the propensity score has a special place

Balancing Scores and the Propensity Score

Lemma

(Coarseness of Balancing Scores)
The propensity score is the coarsest balancing score. That is, the propensity score is a function of every balancing score.

Proof: Let $b(x)$ be a balancing score. Suppose that we can *not* write the propensity score as a function of the balancing score. Then it must be the case that for two values x and x' we have $b(x) = b(x')$, and at the same time $e(x) \neq e(x')$. Then, $\Pr(W_i = 1|X_i = x) = e(x) \neq e(x') = \Pr(W_i = 1|X_i = x')$, and so W_i and X_i are not independent given $b(X_i) = b(x)$, which violates the definition of a balancing score. □

Balancing Scores and the Propensity Score

Because the propensity score is the coarsest possible balancing score, it provides the biggest benefit in terms of reducing the number of variables we need to adjust for.

An important difficulty though arises from the complication that we do not know the value of the propensity score for all units, and thus we cannot directly exploit this result.

Estimation and Inference

Before discussing some of the specific approaches to estimation, it is useful to examine how well these methods can work.

An important tool for this purpose is the *semiparametric efficiency bound*. This is a generalization of the Cramér-Rao sampling variance bound for unbiased estimators.

Let

$$\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}\left[\tau_{\text{sp}}(X_i)\right],$$

where

$$\tau_{\text{sp}}(x) = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x] = \mu_t(x) - \mu_c(x).$$

Thus,

$$\mu_c(x) = \mathbb{E}_{\text{sp}}\left[Y_i(0)|X_i = x\right], \qquad \mu_t(x) = \mathbb{E}_{\text{sp}}\left[Y_i(1)|X_i = x\right].$$

Estimation and Inference

Under unconfoundedness and probabilistic assignment, and without additional functional form restrictions beyond smoothness, the sampling variance bound for estimators for τ_{sp}, normalized by the sample size, is,

$$\mathbb{V}_{\text{sp}}^{\text{eff}} = \mathbb{E}_{\text{sp}}\left[\frac{\sigma_c^2(X_i)}{1-e(X_i)} + \frac{\sigma_t^2(X_i)}{e(X_i)} + \left(\tau_{\text{sp}}(X_i) - \tau_{\text{sp}}\right)^2\right], \tag{5}$$

where

$$\sigma_c^2(x) = \mathbb{V}_{\text{sp}}\left(Y_i(0) \middle| X_i = x\right) \qquad \text{and} \quad \sigma_t^2(x) = \mathbb{V}_{\text{sp}}\left(Y_i(1) \middle| X_i = x\right).$$

Estimation and Inference

This result implies that for any *regular* estimator, its asymptotic sampling variance, after normalizing by the square root of the sample size, cannot be smaller than $\mathbb{V}_{\text{sp}}^{\text{eff}}$.

It is useful to distinguish τ_{sp} from two other average treatment effects,

The *finite sample average treatment effect:* $\tau_{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(1) - Y_i(0) \Big)$

The *conditional average treatment effect:* $\tau_{\text{cond}} = \frac{1}{N} \sum_{i=1}^{N} \tau_{\text{sp}}(X_i)$

Estimation and Inference

The first term and second term in (5) shows that it is more difficult to estimate the average treatment effect if there is a substantial number of units with propensity score values close to one or zero.

The third term in (5) is the variance of the treatment effect conditional on the pretreatment variables. This term is zero if the treatment effect is constant.

Overall the variance expression (5) shows that if the population distribution of covariates is unbalanced between treated and control units, the sampling variance of any estimator will be large. This will be important for analyses (more on this in Chapters 15 and 16).

Estimation and Inference

If instead we focus on τ_{cond}, the efficiency bound changes to

$$\mathbb{V}_{\text{cond}}^{\text{eff}} = \mathbb{E}_{\text{sp}}\left[\frac{\sigma_c^2(X_i)}{1-e(X_i)} + \frac{\sigma_t^2(X_i)}{e(X_i)}\right].$$

We can, at least in principle, estimate τ_{cond} more accurately than τ_{sp} because the latter also reflects the difference between the distribution of the covariates in the sample and the population.

The intuition for this is easily presented in terms of a simple example.

Estimation and Inference

Suppose there is a single binary covariate, with unknown marginal distribution in the super population, $X_i \in \{f, m\}$, with $\Pr(X_i = f) = p$ unknown.

Suppose we can estimate the average effects $\tau_{\text{sp}}(f)$ and $\tau_{\text{sp}}(m)$ accurately for both subpopulations separately because the conditional variances are small, and suppose these average effects differ substantially.

Then it follows that we can estimate τ_{cond} accurately because it is a known function of $\tau_{\text{sp}}(f)$ and $\tau_{\text{sp}}(m)$. However, because p is unknown, we would not be able to estimate τ_{sp} as accurately.

The implication is that it is important for inference to be precise about the estimand.

If we focus on τ_{fs} or τ_{cond}, we need to use a different estimator for the sampling variance than if we focus on τ_{sp}.

Strategies for Estimation

We discuss five strategies for estimation, with some overlap between them.

The first four strategies are model-based imputation, weighting, blocking, and matching methods. These four classes differ in their focus on the unknown components of the joint distribution of the potential outcomes,assignment process and covariates.

Below we briefly describe these four general approaches, as well as a fifth class of estimators that combines aspects of some of these strategies.

Variations of all five of these strategies have been used extensively in empirical work, although we do not recommend all of them.

Strategies for Estimation

In Chapters 17 and 18 we discuss in more detail the implementation for two specific strategies that we view as particularly attractive in practice.

These two strategies are blocking (*i.e.,* subclassification) on the propensity score, in combination with covariance adjustment within the blocks (Chapter 17), and matching, again in combination with covariance adjustment, possibly within the matched pairs (Chapter 18).

We view these two approaches as relatively attractive because of the robustness properties that stem from the combination of methods that ensure approximate comparability, either through blocking or matching, with additional bias removal and precision increases through covariance adjustment.

Strategies for Estimation

Although all four strategies aim at estimating the same treatment effects, there are fundamental differences among them.

One important difference between the model-based imputations and the other three (weighting, blocking, and matching methods), is that the first requires building models for the potential outcomes, whereas for the other three, all decisions regarding the implementation of the estimators can be made before seeing any outcome data.

An important difference because not having outcome data prevents the researcher from adapting the model to make it fit prior notions about the treatment effects of interest.

Although the researcher does have to make a number of important decisions when using weighting, blocking and matching methods, these cannot be implemented in a way to bias directly the estimates for treatment effects, and so have arguably more credibility.

Model-based Imputation

Following the exposition from Chapter 8, we need a model for

$$\mathbf{Y}^{\text{mis}} \,\Big|\, \mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W}. \tag{6}$$

Suppose we specify a model for the joint distribution of the two vectors of potential outcomes given the covariates, now explicitly in terms of an unknown parameter θ:

$$\mathbf{Y}(0), \mathbf{Y}(1) \,\Big|\, \mathbf{X}, \theta. \tag{7}$$

Because of unconfoundedness, $\mathbf{W}$ is independent of $(\mathbf{Y}(0), \mathbf{Y}(1))$ given $\mathbf{X}$, and the specification of (7) implies the distribution

$$\mathbf{Y}(0), \mathbf{Y}(1) \,\Big|\, \mathbf{W}, \mathbf{X}, \theta, \tag{8}$$

which in turns allows us to derive the conditional distribution (6).

Model-based Imputation

Given exchangeability of the units and an appeal to DiFinetti's Theorem, all we need to specify is the joint distribution of

$$(Y_i(0), Y_i(1)) \Big| X_i, \theta,$$

Given such a distribution, we can follow the same approach as in Chapter 8.

With few covariates, it is relatively easy to specify a flexible functional form for this conditional distribution. If there are many covariates, however, such a specification is more difficult, and the results can be sensitive to poor choices.

This situation is qualitatively different from Chapter 8 where such sensitivity was often minor because the covariate distributions in treatment and control groups were similar.

Model-based Imputation

There is no generally need to specify a parametric model for the conditional distribution of the treatment indicator given the covariates, the super-population assignment mechanism,

$$p(\mathbf{W}|\mathbf{X};\phi),$$

because if ϕ and θ are distinct parameters, inference for causal effects is not affected by the functional form of the specification of this assignment mechanism. However, it is important for this argument that ϕ and θ are distinct parameters.

The Concern with Regression Estimators

Suppose we model the potential outcome distributions as :

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| X_i, \theta \sim \mathcal{N}\left(\left(\begin{array}{c} X_i\beta_c \\ X_i\beta_t \end{array} \right), \left(\begin{array}{cc} \sigma_c^2 & \sigma_c \cdot \sigma_t \\ \sigma_c \cdot \sigma_t & \sigma_t^2 \end{array} \right) \right),$$

where $\theta = (\beta_c, \beta_t, \sigma_c^2, \sigma_t^2)$. (Note: X_i is assumed to include a constant term.)

Then we can estimate β_c and β_t by least squares methods:

$$\hat{\beta}_c = \arg\min_{\beta} \sum_{i:W_i=0} (Y_i - X_i\beta)^2, \qquad \text{and} \quad \hat{\beta}_t = \arg\min_{\beta} \sum_{i:W_i=1} (Y_i - X_i\beta)^2.$$

The Concern with Regression Estimators

The population and sample average treatment effects are then be estimated as

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left(W_i \cdot (Y_i^{\text{obs}} - X_i \hat{\beta}_c) + (1 - W_i) \cdot (X_i \hat{\beta}_t - Y_i^{\text{obs}}) \right).$$

The concern is that, in many situations, it can rely heavily on extrapolation.

To see this, it is useful to rewrite the estimator as

$$\hat{\tau} = \frac{N_t}{N_t + N_c} \cdot \hat{\tau}_t + \frac{N_c}{N_t + N_c} \cdot \hat{\tau}_c,$$

where

$$\hat{\tau}_c = \frac{1}{N_c} \sum_{i:W_i=0} \left(X_i \hat{\beta}_t - Y_i^{\text{obs}} \right), \qquad \text{and} \quad \hat{\tau}_t = \frac{1}{N_t} \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - X_i \hat{\beta}_c \right).$$

are estimators for the population average effects.

The Concern with Regression Estimators

Furthermore, because of the presence of a constant term in X_i, we can write $\hat{\tau}_t$ as

$$\hat{\tau}_t = \overline{Y}_t^{\text{obs}} - \overline{X}_t\hat{\beta}_c = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - (\overline{X}_t - \overline{X}_c)\hat{\beta}_c, \tag{9}$$

and similarly

$$\hat{\tau}_c = \overline{X}_c\hat{\beta}_t - \overline{Y}_c^{\text{obs}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - (\overline{X}_t - \overline{X}_c)\hat{\beta}_t. \tag{10}$$

$(\overline{X}_t - \overline{X}_c)\hat{\beta}_c$ and $(\overline{X}_t - \overline{X}_c)\hat{\beta}_t$, are at the core of the concern.

If the two covariate distributions are substantially apart, the difference $\overline{X}_t - \overline{X}_c$ is substantial. Then the "adjustment" terms $(\overline{X}_t - \overline{X}_c)\hat{\beta}_c$ and $(\overline{X}_t - \overline{X}_c)\hat{\beta}_t$ will be sensitive to details of the specification of the regression function.

The Concern with Regression Estimators

In the context of CRE this was less of an issue, because the randomization ensured that in expectation, the covariate distributions were balanced, e.g. $\mathbb{E}_W\left[\overline{X}_t - \overline{X}_c\right] = 0$.

Here, in contrast, the covariate distributions can potentially be far apart even under unconfoundedness. Prior to using regression methods or other modeling approaches, therefore, one has to ensure that there is substantial balance in the covariate distributions.

We return to this issue in short (see Chapters 14-15 for more details).

Weighting Estimators That Use the Propensity Score

Given knowledge of the propensity score, one can directly use some of the strategies that apply to the analysis of randomized experiments with variation in assignment probabilities.

Such possible strategies include weighting, subclassification (similar to stratification in the case of randomized experiments), and matching.

The key difference with the imputation strategy is that these three model, and estimate, the propensity score, whereas an imputation strategy models the conditional outcome distributions.

Weighting Estimators That Use the Propensity Score

The issues in implementing any of these three methods therefore are related to estimation of the propensity score (PS).

One approach is to treat the estimation of the PS as a standard problem of estimating an unknown regression function with a binary outcome.

An alternative approach, more widely used in the evaluation literature, focuses on the essential property of the PS, that of balancing the covariates between the two groups.

Thus, a specification is sought for the PS such that, within blocks with similar values of the PS, the first few (cross) moments of the covariates are balanced between treatment groups.

Weighting Estimators That Use the Propensity Score

The first method involving the PS is weighting. Weighting exploits the fact that

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right], \qquad \text{and} \quad \mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right].$$

These equalities follow by taking iterated expectations, and exploiting unconfoundedness, e.g.,

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] &= \mathbb{E}_{\text{sp}}\left[\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\middle| X_i\right]\right] \\
&= \mathbb{E}_{\text{sp}}\left[\mathbb{E}\left[\frac{Y_i(1) \cdot W_i}{e(X_i)}\middle| X_i\right]\right] \\
&= \mathbb{E}_{\text{sp}}\left[\frac{\mathbb{E}_{\text{sp}}[Y_i(1)|X_i] \cdot \mathbb{E}_W[W_i|X_i]}{e(X_i)}\right] \\
&= \mathbb{E}_{\text{sp}}\left[\mathbb{E}_{\text{sp}}[Y_i(1)|X_i]\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right],
\end{aligned}$$

Weighting Estimators That Use the Propensity Score

and similarly for the second equality.

One can exploit these equalities by estimating the average treatment effect as

$$\begin{aligned}\hat{\tau}^{\text{ht}} &= \frac{1}{N}\sum_{i=1}^{N}\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)\cdot Y_i^{\text{obs}}}{1-e(X_i)} \\ &= \frac{1}{N}\sum_{i:W_i=1}\lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N}\sum_{i:W_i=0}\lambda_i \cdot Y_i^{\text{obs}},\end{aligned}$$

where

$$\lambda_i = \frac{1}{e(X_i)^{W_i}\cdot(1-e(X_i))^{1-W_i}} = \begin{cases} 1/(1-e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

Weighting Estimators That Use the Propensity Score

In practice typically we would not know the true population PS and one would have to use an estimate of the PS, $\hat{e}(x)$ in place of $e(x)$.

Instead of using these weights directly, one would further adjust the weights, so that they add up to the sample size for each treatment group, that is, use $\hat{\lambda}_i$, where

$$\hat{\lambda}_i = \begin{cases} N \cdot (1 - \hat{e}(X_i))^{-1} / \sum_{j:W_j=0} (1 - \hat{e}(X_i)^{-1} & \text{if } W_i = 0, \\ N \cdot \hat{e}(X_i)^{-1} / \sum_{j:W_j=1} \hat{e}(X_i)^{-1} & \text{if } W_i = 1. \end{cases}$$

Just like we do not recommend the simple regression estimator, we do not recommend this type of estimator in settings with a substantial difference in the covariate distributions by treatment status.

Weighting Estimators That Use the Propensity Score

In a CRE, the PS would be constant, and even when the PS is estimated, the weights are likely to be similar for all treated and for all control units.

In contrast, when the covariate distributions are far apart, the PS will be close to zero or one for some units, and the weights, proportional to $1/e(X_i)$ or $1/(1-e(X_i))$, can be large. As a result, in such settings estimators can be sensitive to minor changes in the specification of the model for the PS.

Blocking Estimators That Use the Propensity Score

In this third approach, the sample is partitioned into subclasses, based on the value of the estimated PS.

Within each subclass, the data can be analyzed as if they arose from a CRE.

Let b_j, $j = 0, 1, \ldots, J$ denote the subclass boundaries, with $b_0 = 0$ and $b_J = 1$, and let $B_i(j)$ be a binary indicator, equal to 1 if $b_{j-1} \leq \hat{e}(X_i) < b_j$, and zero otherwise.

Then we estimate the finite sample average effect in subclass j, $\tau_{\text{fs}}(j)$, by:

$$\hat{\tau}^{\text{dif}}(j) = \frac{\sum_{i:B_i(j)=1} Y_i \cdot W_i}{\sum_{i:B_i(j)=1} W_i} - \frac{\sum_{i:B_i(j)=1} Y_i \cdot (1 - W_i)}{\sum_{i:B_i(j)=1} (1 - W_i)}.$$

Blocking Estimators That Use the Propensity Score

To estimate the overall finite sample average effect of the treatment, τ_{fs}, we use

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \hat{\tau}^{\text{dif}}(j),$$

where $N(j) = \sum_{i=1}^{N} B_i(j)$, and using the label "$\text{strat}$" to stress the connection with the estimators used in the SRE.

Although this method is more robust than the weighting estimator to the presence of units with extreme values of the estimated PS, we still do not recommend it.

In particular, we recommend reducing the bias and increasing the precision further by using covariate adjustment within the subclasses (more on this in Chapter 17).

Matching Estimators

Unlike model-based imputation and weighting and blocking methods, the fourth approach, matching, does not rely on estimating an unknown function.

Suppose we wish to to assess the effect of a JTP on income for, say a 30 year old woman with 2 children under the age of 6, with a high school (HS) education and 4 months of work experience in the last 12 months, who went through this JTP.

In the matching approach we look for a 30 year old woman with 2 children under the age of 6, with a HS education and 4 months of work experience in the last 12 months, who did *not* attend the JTP.

If exact matches can be found, this is a particularly attractive, intuitive and simple strategy. If no exact matches can be found this approach becomes more unwieldy.

Matching Estimators

In that case one needs to assess the trade-offs of different violations of exact matching.

Whom should we use as a match for the 30 year old HS educated woman with 2 children and 4 months of work experiments who went through the JTP?

One possibility may be a HS educated woman from the control group who is 4 years older, with 4 months more work experience.

Assessing the relative merits of such matches requires careful inspection of the joint distribution of the covariates and substantive knowledge.

Clearly, as soon as such compromises need to be made, matching is more difficult to implement.

Matching Estimators

Difficulties in dealing with many covariates shows up here in a different form than in the model-based imputation methods, but they do not disappear.

With many covariates, the quality of the matching, measured by some metric of the typical distance between covariates of units and the covariates of their matches, decreases.

To implement the matching approach, one needs to be able to assess the trade-offs in choosing between different controls, and thus requires a distance metric.

Chapter 18 discusses some of the choices that have been used in the literature

Mixed Estimators

Using regression, not globally, but only within blocks with similar covariate distributions for treated and control units, for example defined by the estimated PS, may combine attractive properties of regression adjustment in relatively well-balanced samples with the robustness of subclassification methods across different distributions.

Similarly one can combine matching with regression, again exploiting the strengths of both methods.

Subclassification with covariate adjustment within subclasses, and matching with covariance adjustment, are two of the most attractive methods for estimating treatment effects, especially when flexibly implemented.

These approaches, and specific methods methods for implementing them, is discussed in more detail in Chapters 17 and 18.

Design Phase

Prior to any analysis, it is important to conduct what we call the *design phase* of an observational study.

In this stage, one should investigate the extent of overlap in the covariate distributions.

This, in turn, may lead to the construction of a subsample more suitable for estimating causal estimands, in the sense of being better balanced in terms of covariate distributions.

There is one important feature of this initial analysis: this stage does not involve the outcome data, which need not be available at this stage, or even collected yet.

As a result, this analysis cannot be "contaminated" by knowledge of estimated outcome distributions, or by preferences, conscious or unconcious, for particular results.

Assessing Balance

Examine the difference in average covariate values by treatment status, scaled by their sample standard deviation.

As a rule-of-thumb, when treatment groups have important covariates that are more than one quarter or one half of a standard deviation apart, OLS estimators are unreliable for removing biases associated with differences in covariates.

In addition, examine the distributions of the PS.

If the super-population covariate distributions are identical in the two treatment groups, then the true PS must be constant, and *vice versa*.

Variation in the estimated PS is therefore a simple way to assess differences between two multivariate distributions.

Assessing Balance

Estimating the PS involves choosing a specification for the PS and estimating the unknown parameters of that specification (to be discussed in Chapter 13).

In Chapter 14 the specific methods for comparing covariate distributions and assessing balance is discussed in detail.

Subsample Selection Using Matching on the Propensity Score

Chapter 15 provide details for one method of constructing balanced subsamples.

The method relies on having a relatively large number of controls, and is appropriate for settings where we are interested in the effect of the treatment on the subpopulation of treated units.

The proposed procedure consists of two steps. First we estimate the PS. Then we sequentially match each treated unit to the closest control unit in terms of the estimated PS.

Within this matched sample, we apply some of the adjustment methods introduced previously, including those that allow for estimation of more general causal estimands than average effects.

Subsample Selection Through Trimming Using the Propensity Score

Chapter 16 discuss in more detail a second method for constructing balanced samples that also uses the estimated PS.

The idea here is that for units with covariate values such that the PS is close to zero or one, it is difficult to obtain precise estimates of the typical effect of the treatment because, for such units, there are few controls relative to the number of treated units, or the other way around.

We therefore propose putting aside such units and focusing on estimating causal effects in the subpopulation of units with PS values bounded away from zero and one.

More precisely, we discard all units with estimated PS outside an interval, and we propose a specific way to chose the interval.

Assessing Unconfoundedness

Chapter 20 discusses methods for assessing the UA.

The term "assess" is used rather than "test," because unconfoundedness has no directly testable implications.

Nevertheless, there are a number of statistical analyses that we can conduct that can shed light on its plausibility.

Some of these assessments, like the analyses assessing balance, do not involve the outcome data, and so are part of the design stage.

The conclusion from such assessments can be that one may deem the UA implausible or plausible and then pursue with the causal analysis.

Here we briefly introduce three of these assessments.

Estimating the Effect of the Treatment on an Unaffected Outcome

The first set of assessments focuses on estimating the causal effect of the treatment on a variable that is known *a priori* not to be affected by the treatment, typically because its value is determined prior to the treatment itself.

Such a variable can be a time-invariant covariate, but the most interesting case is where this is a lagged outcome.

In this case, one uses all the covariates except the single covariate that is being assessed, say the lagged outcome.

One estimates the pseudo treatment effects on the lagged outcome.

Estimating the Effect of the Treatment on an Unaffected Outcome

If these estimated effects are near zero, it is more plausible that the UA holds than if the estimated effects are large.

Of course the assessment is not directly testing the UA, and so no matter what the p-value of the null hypothesis of no effect, it does not directly reflect on the assumption of interest, unconfoundedness.

Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the assessment has arguably more force than if the variables are unrelated to the outcome of interest.

For these analyses, it is clearly helpful to have a number of lagged outcomes.

This approach is a *design* approach, not using any outcome data.

Estimating the Effect of a Pseudo Treatment on the Outcome

The second set of assessments focuses on estimating the causal effect of a different treatment on the original outcome, and in particular a "pseudo" treatment that is known *a priori* not to have an effect.

This approach relies on the presence of multiple control groups and uses actual outcome data, but only for the control units.

One interpretation of the assessment is that one compares estimated treatment effects calculated using one control with average treatment effects calculated using the other control group.

With two control groups, the assessment would be implement, using data on the control groups only, by estimating an average treatment effect with the treatment indicator redefined as an indicator for one of the two control groups.

Estimating the Effect of a Pseudo Treatment on the Outcome

In that case, the pseudo treatment effect is known to be zero, and statistical evidence of a non-zero estimated treatment effect implies that, for at least one of the control groups, the UA is violated.

Again, failure to reject this "test" does not mean the UA is valid because it could be that both control groups have similar biases, but non-rejection in the case where the two control groups are *a priori* likely to have different biases makes it more plausible that the UA holds.

The key for the value of this assessment is to have control groups that are likely to have different biases, if at all.

One may use different geographic control groups, for example on either side of the treatment group. This approach is a *semi-design* approach, using only outcome data for the control units.

Assessing Sensitivity of Estimates to the Choice of Pretreatment Variables

The last approach to assessing the UA uses outcome data for all units.

The idea is to partition the covariates again into two parts.

Now the assessment involves comparing estimates for treatment effects using only a subset of the covariates to those for the full set of covariates.

Substantial differences suggest that either unconfoundedness relies critically on all covariates, or it does not hold.

Because this approach uses outcome data for all units, it is not a (semi-)design approach.