Regression Adjustment for Causal Inference: A Primer with Examples

P. Richard Hahn and Andrew Herren

May 16, 2025

1 Introduction

This monograph is a primer on regression adjustment for causal inference, particularly questions of variable and feature selection in that context. The special case of no adjustment is also considered, as would be appropriate for data from a randomized experiment.

To begin, three distinct formalisms for causal inference are presented: potential outcomes [Imbens and Rubin, 2015], causal diagrams [Pearl, 2009a], and structural equations [Heckman and Vytlacil, 2007]. It is shown (Section 2) that the key condition licensing valid causal inference from observational data can be expressed equivalently in each of the three distinct frameworks: conditional unconfoundedness, the back-door criterion, and additive errors that are independent of treatment assignment. While this equivalence is know to experts, it seems to be not well-known among rank-and-file data analysts and is rarely spelled out in any detail in expository texts; we do so here. The result is established in terms of a generic function of observed covariates, meaning that it covers not only variable selection, but "feature selection" as might be applied in the setting of "machine learning" with regression trees or neural networks, methods which implicitly introduce multivariate transformations of the observed covariates.

Note that the "examples" from the title are not empirical examples, but are rather specific causal structures and data generating processes meant to illustrate particular technical points.

2 Formal frameworks for causal inference

Let Y be the outcome/response of interest, Z be a binary treatment assignment, and X be a vector of covariates drawn from covariate space $\mathcal{X}$, all denoted here as random variables. For a sample of size n, observations are assumed to be drawn independently as triples (X_i, Y_i, Z_i), for $i = 1, \ldots, n$. The goal of causal effect estimation is to understand how the response variable Y changes according to hypothetical manipulations of the treatment assignment variable, Z.

For simplicity, we will refer to our observational units as "individuals", although of course in applications that need not be the case, they could be schools or businesses or teams or countries or what have you.

The essential challenge to causal estimation is that only one of the two possible treatment assignments can be observed; as a consequence, if individuals who happen to receive the treatment differ systematically from those who do not, either in terms of their likely response value or in terms of how they respond to treatment, naive comparisons between the treated and untreated units will not simply reflect the causal impact of the treatment — the treatment effect is said to be *confounded* with other aspects of the population.

The field of causal inference has proposed and developed a variety of techniques for coping with this difficulty, the most common of which is some form of regression adjustment (meant here to include propensity score estimators and matching estimators, etc), which entails estimating average causal effects as (weighted) averages of (estimated) conditional expectations. The key assumption that justifies this process is referred to as *conditional unconfoundedness*, which asserts that the measured covariates adequately account for all of the systematic differences between the treated and untreated individuals in our observational sample;

formalizing this assumption can be approached in a number of ways, which we turn to now. Only after the notation of these formalisms has been introduced can our causal estimands and their estimators be defined.

2.1 Potential outcomes

The potential outcomes framework casts causal inference as a missing data problem: causal estimands are contrasts between pairs of outcomes that are mutually unobservable — when we see one, we cannot see the other. At present, the standard reference for the potential outcomes framework is Imbens and Rubin [2015], which contains extensive citations to the primary literature.

Let Y^1 and Y^0 refer to the "potential outcomes" when $Z = 1$ and $Z = 0$. For individual i, the *individual treatment effect* will be defined as the difference between the potential outcomes:

$$\tau_i = Y_i^1 - Y_i^0.$$

Other treatment effects, such as a ratio rather than a difference, are sometimes considered, but in this paper we focus on the difference. Because the potential outcomes (Y^1, Y^0) are never observed simultaneously, individual treatment effects can never be estimated directly[1].

However, *average* treatment effects can be identified (learned from data) provided certain assumptions are satisfied. The causal estimand this paper will focus on is the average treatment effect, or ATE:

$$\bar{\tau} \equiv \mathbb{E}[Y^1 - Y^0]. \tag{1}$$

The precise population over which this expectation is taken will be discussed in more detail in section 2.5. The standard assumptions that allow this average effect to be estimated are:

1. Stable unit treatment value assumption (SUTVA), which consists of two conditions:
   (a) *Consistency*: The observed data is related to the potential outcomes via the identity
   $$Y = Y^1 Z + Y^0(1 - Z), \tag{2}$$
   which describes the "gating" role of the observed treatment assignment, Z.
   (b) *No Interference*: for any sample of size n with $Y \in \mathcal{Y}$ and $Z \in \mathcal{Z}$, $(Y_i^1, Y_i^0) \perp\!\!\!\perp Z_j$ for all $i, j \in \{1, ..., n\}$ with $j \neq i$, which rules out interference between observational units.
2. Positivity: $0 < \mathbb{P}(Z = 1 \mid X = x) < 1$ for all $x \in \mathcal{X}$
3. Conditional unconfoundedness: $(Y^1, Y^0) \perp\!\!\!\perp Z \mid X$

Imagining concrete violations of these conditions is intuition-building. Consistency can be violated under non-compliance, so that treatment assignment doesn't match treatment actually received. No interference can be violated, for example, if we were studying the effect of individual tutoring on student grades in a certain classroom and students study together; Jimmy's treatment assignment may impact Sally's grade. Positivity is violated if certain individuals can never receive treatment, rendering their contribution to the average treatment effect unlearnable. And finally, conditional unconfoundedness can be violated, for example, if both treatment assignment and the outcome variable share a common cause (that is not account for in X). However, this is not the only way conditional unconfoundedness can be violated, and exploring other possibilities in full generality is the topic of the remainder of the paper.

Taken together, the above assumptions enable identification of average treatment effects because they imply the following equality, the left-hand side of which is estimable:

$$\mathbb{E}_X[\mathbb{E}[Y \mid X, Z = 1] - \mathbb{E}[Y \mid X, Z = 0]] = \mathbb{E}[Y^1 - Y^0].$$

[1] Even with repeated measurements from the same individual across time, the treatment effect that is estimable is an average across time; the observational unit there is individual-per-time and at that level the ITE cannot be estimated.

In more detail, the equivalence is established as follows:

$$\begin{aligned}\mathbb{E}_X[\mathbb{E}[Y \mid X, Z=1] &= \mathbb{E}_X[\mathbb{E}[Y^1 Z + Y^0(1-Z) \mid X, Z=1]] \\ &= \mathbb{E}_X[\mathbb{E}[Y^1 \mid X, Z=1] = \mathbb{E}[Y^1]. \\ \mathbb{E}_X[\mathbb{E}[Y \mid X, Z=0] &= \mathbb{E}_X[\mathbb{E}[Y^1 Z + Y^0(1-Z) \mid X, Z=0]] \\ &= \mathbb{E}_X[\mathbb{E}[Y^0 \mid X, Z=0] = \mathbb{E}[Y^0].\end{aligned}$$

An alternative parametrization is:

$$Y_i = Y_i^0 + \tau_i Z_i$$

where

$$\tau_i = Y_i^1 - Y_i^0,$$

which emphasizes that τ_i itself can differ across units and, as a random variable, can be *dependent* on the treatment assignment so that $\tau \not\perp\!\!\!\perp Z$. This treatment effect parametrization will be used extensively in this manuscript.

The following question motivates our investigation: If X satisfies conditional unconfoundedness, could there be a function of X with a reduced range that also satisfies conditional unconfoundedness? That is, can X be reduced in dimension while still providing valid causal effect estimation? Answering this question requires a more detailed examination of *how* conditional unconfoundedness is achieved in any particular data generating process, which is facilitated by the introduction of causal diagrams.

2.2 Causal diagrams

2.2.1 Graph theory for causal identification

Causal diagrams provide a more fine-grained look at confounding, as they consider the full joint distribution of the response, treatment, and control variables regressors. The graphical approach to causality has its earliest roots in the work of Sewell Wright [Wright, 1918, 1920, 1921], but attained its mature modern form in the prodigious work of Judea Pearl [Pearl, 1987, Pearl and Verma, 1987, 1995, Pearl, 1995]. See Pearl [2009a] for a textbook treatment and comprehensive references. The presentation here loosely follows the expository treatment in Shalizi [2021].

Recall that any joint density over p random variables may be expressed in *compositional form*, as a product of conditional densities:

$$f(x_1, x_2, \ldots, x_p) = f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_1, x_2) \ldots f(x_p \mid x_1, x_2, \ldots, x_{p-1}),$$

where the density functions $f(\cdot)$ and $f(\cdot \mid \cdot)$ refer to different densities depending on their arguments. The labeling of the variables is arbitrary, and so we can chain together these marginal and conditional distributions in any order (though of course that will lead to different forms). Some of these variables might exhibit *conditional independence*, meaning that, for example

$$f(x_1 \mid x_2, x_3) = f(x_1 \mid x_2)$$

which is equivalently expressed as

$$X_1 \perp\!\!\!\perp X_3 \mid X_2.$$

The relationship to *directed (acyclic) graphs* (DAG) is straightforward: draw a node for each variable and draw a line from X_j going into X_i if X_j appears in the conditional distribution of X_i. This graph is *directed*, with the arrow pointing from X_j *to* X_i. We say that X_j is a "parent" of X_i and that X_i is the "child" of X_j.

From the graph, the joint distribution may be expressed as

$$f(x_1, \ldots, x_p) = \prod_{j=1}^{p} f(x_j \mid \text{parents}(x_j)).$$

This leads us to the *Markov property*, which is

$$X_j \perp\!\!\!\perp \text{non-descendants}(X_j) \mid \text{parents}(X_j),$$

where "descendant" refers to children, grandchildren, great-grandchildren, etc. We can see this by dividing through by the marginal distribution of parents(X_j) and observing that the resulting distribution is a product of terms involving either X_j or non-descendants(X_j), but not both. The Markov property allows one to efficiently deduce conditional independence relationships and underpins Pearl's algorithm (which will be described shortly).

Finally, a complete treatment of confounding in the causal diagram framework requires the following definition:

Definition 1. *A* collider *is a node/variable V in a DAG that sits on an undirected path between two other nodes/variables, X_j and X_i, and the paths both have arrows pointing* into *V.*

Conditioning on a collider induces dependence between its parents. For a classic example of this phenomenon, suppose that a certain college grants admission only to applicants with high test scores and/or athletic talent. Even if we grant that in the general population these talents may be independent, but among admitted students, these two attributes become highly dependent. If we know that a student is not athletic, then we know for sure that they must be academically gifted and vice-versa. While this is a basic result in probability theory, Pearl's work emphasized its significance to the problem of regression adjustment for causal effect estimation. Sometimes this phenomenon is known by "Berkson's paradox" [Berkson, 1946].

With a DAG in hand, it is possible to deduce – rather than assume – conditional unconfoundedness: Pearl developed an algorithm for determining subsets of variables in X (i.e., its coordinate dimensions) that define valid regression estimators. The inputs to this algorithm are a directed acyclic graph (DAG) that characterizes the causal relationships between variables; such a graph describes a particular compositional representation of the joint distribution, reflecting conditional independences that are implied by the *stipulated* causal relationships. The prohibition on cycles rules out positive feedback self-causation.

Here we present Pearl's algorithm in a somewhat simplified form, assuming that the graph contains no descendants of Z other than Y (and thus no descendants of Y either). That is, Z has no out arrows other than Y, and Y has no out arrows at all.

Given an input DAG, $\mathcal{G}$ and a subset of nodes S, the "backdoor" algorithm proceeds as follows:

1. Identify all (undirected) paths between Z and Y.
2. Consider each variable along each of these paths and make sure that at least one of them is "blocked".
   (a) A variable W is blocked if
      i. W is not a collider and is in the set S or
      ii. W is a collider and neither W nor any of its descendants is in the set S.
3. Return TRUE if every "backdoor" path between Z and Y (all paths except the direct causal arrow from Z to Y), is blocked. Otherwise return FALSE.

Sets of variables satisfying the backdoor criterion — those sets where the algorithm returns TRUE — are valid adjustment sets in the sense that Y and Z *would be* conditionally independent, given those variables, *if there were no causal relationship* between Y and Z. By ruling out all other possible sources of association, any observed association may be interpreted as arising from a causal relationship. Causal identification is thus a sort of process of elimination.

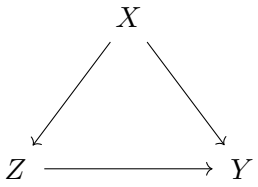

Figure 1: A simple triangle confounding diagram, where a control variable X causally influences both the treatment Z and the response Y. This graph does not clarify what information contained in (the potentially multidimensional) X is relevant for Z or Y or both or neither, only that knowing the value of X in its entirety permits causal estimation.

2.2.2 Functional causal models.

Causal DAGs may be associated with a functional causal model, a set of deterministic functions that take as inputs elements of X as well as independent ("exogenous") error terms. The basic triangle confounding graph corresponding to an (X, Y, Z) triple satisfying conditional unconfoundedness is shown in Figure 1. The corresponding functional causal model can be expressed as

$$\begin{aligned} Z &\leftarrow G(X, \epsilon_z) \\ Y &\leftarrow F(X, Z, \epsilon_y) \end{aligned} \tag{3}$$

where X, ϵ_z and ϵ_y are mutually independent (though all three may be vector-valued with non-independent elements). The exogenous errors (ϵ_z and ϵ_y) that appear in a single equation are suppressed in the graph. All of the stochasticity is inherited from the exogenous variables, while all of the deterministic relationships are reflected in the functions $G(\cdot)$ and $F(\cdot)$, which are explicitly endowed with a causal interpretation. Specifically, the potential outcomes are given by:

$$\begin{aligned} Y^1 &\leftarrow F(X, 1, \epsilon_y) \\ Y^0 &\leftarrow F(X, 0, \epsilon_y) \end{aligned} \tag{4}$$

where (X, ϵ_y) are drawn from their marginal distributions, irrespective of the value of the treatment argument. As was mentioned previously, throughout this paper we assume that X does not contain any causal descendants of Z.

Consider two ways to conceptualize the data generating process for both the potential outcome pairs, (Y^0, Y^1), and the observed response Y. On the one hand, the potential outcomes can be generated from the functional causal model, by fixing the Z argument to 0 or 1, irrespective of the implied distribution of $Z \mid X$. Procedurally, this would look like drawing X from its marginal distribution, drawing ϵ_y, and evaluating $F(X, 0, \epsilon_y)$ and $F(X, 1, \epsilon_y)$. The observed data can then be constructed via the consistency assumption $Y = F(X, 1, \epsilon_y)Z + F(X, 0, \epsilon_y)(1 - Z)$. Equivalently, Y may be drawn directly via $F(X, Z, \epsilon_y)$, where Z (the observed treatment assignment) was drawn according to $Z \mid X$ (as specified by the CDAG). This equivalence is especially instructive as to why $Y \mid Z = z$ and Y^z do not generally have the same distribution and, furthermore, why $Y \mid Z = z, X = x$ and $Y^z \mid X = x$ do have the same distribution (assuming, as we have above, that X is causally exhaustive).

The role of ϵ_y in defining the distribution of the potential outcomes is worth considering in more detail. Note that for a binary Z, any functional causal model F may be rewritten as

$$F(X, Z, \epsilon_y) = F(X, 0, \epsilon_y) + Z\left[F(X, 1, \epsilon_y) - F(X, 0, \epsilon_y)\right] = \mu(X, \epsilon_y) + Z\tau(X, \epsilon_y).$$

This formulation invites us to consider that ϵ_y may be multivariate, distinct elements of which may affect $\mu(X, \epsilon_y)$ and $\tau(X, \epsilon_y)$. Three particular cases are especially notable:

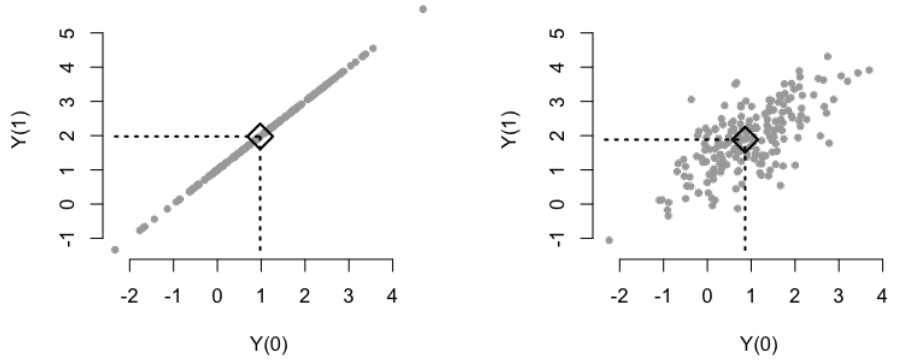

Figure 2: Left panel: Potential outcome distributions with a common additive univariate error and a homogeneous treatment effect (which shifts the line up one unit from the diagonal), articulated in Case 1 below. Right panel: Potential outcome distributions with a homogeneous treatment effect and distinct additive bivariate errors, $\epsilon_{y,0}$ and $\epsilon_{y,1}$, shown here with a positive correlation less than one, as described in cases 2 of the text.

1. $\mu(X, \epsilon_y) = \mu(X) + \epsilon_y$ and $\tau(X, \epsilon_y) = \tau(X)$: here, ϵ_y has the same effect on the two potential outcomes $F(X, 1, \epsilon_y)$ and $F(X, 0, \epsilon_y)$, so that their joint distribution is singular.

2. $\mu(X, \epsilon_y) = \mu(X) + \epsilon_{y,0}$ and $\tau(X, \epsilon_y) = \tau(X) + (\epsilon_{y,1} - \epsilon_{y,0})$ where the exogenous error is partitioned as $\epsilon_y = (\epsilon_{y,0}, \epsilon_{y,1})$. Here, $\epsilon_{y,0}$ and $\epsilon_{y,1}$ are distinct random variables that separately define the potential outcome distributions so that one effect of the treatment is in changing *which* exogenous influences affect the response.

3. $\mu(X, \epsilon_y) = \mu(X) + \epsilon_{y,\mu}$, $\tau(X, \epsilon_y) = \tau(X) + \epsilon_{y,\tau}$, where the exogenous errors is partitioned as $\epsilon_y = (\epsilon_{y,\mu}, \epsilon_{y,\tau})$. In this case, a distinct set of causal factors dictate exogenous variation in the prognostic (baseline) response and exogenous variation in the treatment effect itself. For example, variation in the baseline response may be due to environmental factors that are independent from genetic factors dictating one's response to a new drug.

These cases are visualized in Figure 2 with $\tau(X) = 1$. Empirically, these cases are indistinguishable in that they are "observationally equivalent" — because the potential outcomes are never jointly observed, most aspects of their joint distribution are fundamentally unidentified.

With a more detailed causal graph, a more detailed assessment of conditional unconfoundedness can be made. For instance, consider Figure 3, which is equivalent to the standard triangle digram in the sense that controlling for all of the elements of $X = (X_1, X_2, X_3, X_4)$ indeed satisfies conditional unconfoundedness. However, Pearl's algorithm reveals that (X_1, X_2) would suffice. By positing more information about the joint distribution of X, it is possible to absorb X_3 into ϵ_z and X_4 into ϵ_y, while redefining $X = (X_1, X_2)$, bringing us back to the triangle graph, but with a reduced set of control variables.

2.3 Structural equations: Mean regression models with exogenous additive errors

Finally, the classic econometric literature approaches causality in terms of mean regression models with additive (but not necessarily homoskedastic) error terms, which are referred to as "structural" models (although the term is often used informally and imprecisely in the applied literature). Heckman and Vytlacil [2005] reviews the structural model approach in econometrics in depth, noting that such methods

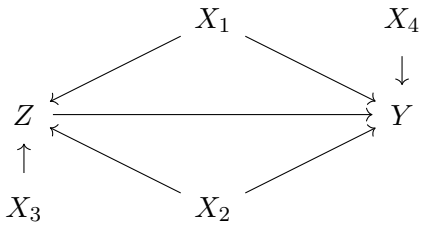

Figure 3: An elaboration of the triangle graph, depicting X_1 and X_2 as confounders, X_4 as a pure prognostic variable, and X_3 is an instrument.

have their origin in the study of dynamic macroeconomic systems. A seminal reference is Haavelmo [1943]. The mean regression perspective arises naturally if one takes a linear regression model as a starting point, but is straightforward to motivate starting from a generic functional causal model.

Define

$$\begin{aligned} \mu(x) &\equiv \mathbb{E}(F(x,0,\epsilon_y)), \\ \tau(x) &\equiv \mathbb{E}(F(x,1,\epsilon_y)) - \mu(x), \\ \upsilon(x,\epsilon_y) &\equiv F(x,0,\epsilon_y) - \mu(x), \\ \delta(x,\epsilon_y) &\equiv F(x,1,\epsilon_y) - F(x,0,\epsilon_y) - \tau(x) \end{aligned} \tag{5}$$

giving a "structural model"

$$Y = \mu(x) + \upsilon(x,\epsilon_y) + (\tau(x) + \delta(x,\epsilon_y))Z \tag{6}$$

where $\upsilon(x,\epsilon_y)$ and $\delta(x,\epsilon_y)$ are deterministic functions, both of which are mean zero integrating over ϵ_y (for any x): $\mathbb{E}(\upsilon(x,\epsilon_y)) = 0$ and $\mathbb{E}(\delta(x,\epsilon_y)) = 0$. In this formulation, conditional unconfoundedness may be expressed in terms of independence of the treatment, Z, and the error terms $\upsilon(x,\epsilon_y)$ and $\delta(x,\epsilon_y)$. Such models are commonly used in a simplified form, where $\delta(x,\epsilon_y)$ is assumed to be identically zero and $\tau(x)$ is assumed to be constant in x, but such assumptions are not intrinsic to the formalism.

2.4 Relating the three frameworks

If every node in a causal diagram is observable, all remaining factors determining Y are attributable to the exogenous errors, which are, by definition, independent of the treatment assignment. In that case, it is easy to forge a connection between the three formalisms, as they all assert that

$$Y^z \mid X = x \quad \overset{d}{\sim} \quad Y \mid X = x, Z = z, \tag{7}$$

where (recall) $Y^z = F(x,z,\epsilon_y)$, with distribution induced by the distribution over ϵ_y. The above assertion essentially declares that the estimable conditional distributions which appear on the right hand side warrant a causal interpretation.

For sets of control variables that are *not* exhaustive, more care is needed in translating the formalisms, but a precise relationship can be obtained, as spelled out in the following lemma.

Lemma 1. *The assertions below (with their corresponding causal framework labeled in brackets) stand in the following logical relationship:* $1 \Rightarrow 2 \Leftrightarrow 3$.

1. $S = s(X)$ *satisfies the back-door criterion. [Causal DAGs]*
2. $S = s(X)$ *satisfies conditional unconfoundedness:* $(Y^0, Y^1) \perp\!\!\!\perp Z \mid S$. *[Potential Outcomes]*
3. *The response* Y *can be represented in terms of a mean regression model with error terms* $(\upsilon(s,X,\epsilon_y), \delta(s,X,\epsilon_y))$ $Z \mid s(X) = s$. *[Structural Equations]*

Figure 4: A typical causal DAG (CDAG) and its potential outcome counterpart, where $Y^* = (Y^0, Y^1)$.

Proof. Let X denote all of the variables in a complete causal diagram with the exception of the treatment variable Z and response variable Y, and consider the following causal model, written in terms of functional equations, potential outcomes, and a structural mean regression with additive exogenous errors:

$$\begin{aligned} Z &\leftarrow G(X, \epsilon_z), \\ Y^z &\leftarrow F(X, z, \epsilon_y) = \mu(X) + \upsilon(X, \epsilon_y) + (\tau(X) + \delta(X, \epsilon_y))z, \\ \begin{pmatrix} Y^0 \\ Y^1 \end{pmatrix} &\leftarrow \begin{pmatrix} \mu(X) + \upsilon(X, \epsilon_y) \\ \mu(X) + \tau(X) + \upsilon(X, \epsilon_y) + \delta(X, \epsilon_y) \end{pmatrix}. \end{aligned} \tag{8}$$

To see that 1 implies 2, recall that 1 means that S renders the treatment and response conditionally independent in the modified DAG with no causal arrow between Z and Y. But it is precisely such a graph that defines the relationship between Z and the potential outcomes $Y^0 = F(X, 0, \epsilon_y)$ and $Y^1 = F(X, 1, \epsilon_y)$, as shown in Figure 4.

To see that 2 and 3 are equivalent, re-parametrize the additive error model in terms of S, as follows:

$$\begin{aligned} Y^z &\leftarrow \mu(s) + \upsilon(s, X, \epsilon_y) + (\tau(s) + \delta(s, X, \epsilon_y))z \\ \mu(s) &\equiv \mathbb{E}(\mu(X) \mid S(X) = s) \\ \tau(s) &\equiv \mathbb{E}(\tau(X) \mid S(X) = s) \\ \upsilon(s, X, \epsilon_y) &\equiv \mu(X) - \mu(s) + \upsilon(X, \epsilon_y) \\ \delta(s, X, \epsilon_y) &\equiv \tau(X) - \tau(s) + \delta(X, \epsilon_y). \end{aligned} \tag{9}$$

For a fixed value of s, the mean terms $\mu(s)$ and $\tau(s)$ are constant, so that (Y^0, Y^1) stands in a one-to-one relationship with $\upsilon(s, X, \epsilon_y)$ and $\delta(s, X, \epsilon_y)$; therefore if the former are independent of Z, then so must be the latter, and vice-versa.

□

One might wonder why assumption 2 (equivalently, 3) does not imply 1. This can be done by exhibiting a counterexample where 2 and 3 hold but 1 does not. A concrete example is given in section 3.2.

2.5 Estimands, estimators, and sampling distributions

As described previously, by *treatment effect*, we mean the difference between the treated and untreated potential outcomes. By *average* treatment effect, we mean the average of this difference over some population of individuals. The functional causal model and a distribution over the exogeneous errors define an infinite hypothetical *population* from which the observed data is assumed to be a random sample. From this perspective, the population average treatment effect (PATE) may be expressed as

$$\mathbb{E}(\tau(X) + \delta(X, \epsilon)) = \mathbb{E}(\tau(X)), \tag{10}$$

where τ is a fixed-but-unknown function and the expectation is taken with respect to the data generating process defined by the CDAG and the associated functional causal model, so that X and ϵ are both being averaged over.

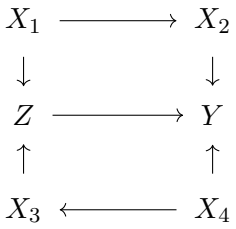

Figure 5: The "box diagram", which implies several valid control sets: any set containing at least one of $\{X_1, X_2\}$ and at least one of $\{X_3, X_4\}$.

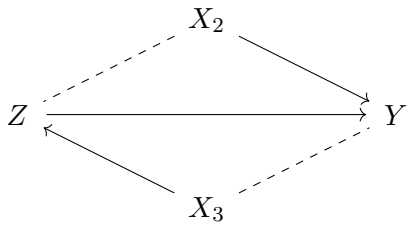

Figure 6: The "box diagram" with X_1 and X_4 omitted; a CDAG representation is no longer possible.

Other average causal effects, differing in terms of the (sub)population over which the average is taken, are likewise readily defined in terms of the functional causal model (FCM). For instance, if we wish to restrict our attention to the average treatment effect among individuals in our observed sample, we may define our estimand as the *sample average treatment effect*, or SATE:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\tau(x_i) + \delta(x_i, \epsilon_i)\right). \tag{11}$$

Note that the SATE and the PATE differ from one another in that, in general,

$$\mathbb{E}(\tau(X)) \neq \frac{1}{N}\sum_{i=1}^{N}\tau(x_i)$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\delta(x_i, \epsilon_i) \neq \mathbb{E}(\delta(X, \epsilon)) = 0.$$

Another average treatment effect of broad interest is the *conditional average treatment effect* (CATE), which defines an average treatment effect conditional on a set of covariate values. The population CATE,

$$\mathbb{E}(\tau(X) + \delta(X, \epsilon) \mid X = x) = \mathbb{E}(\tau(X) \mid X = x) = \tau(x), \tag{12}$$

is defined directly by the $\tau(x)$ function, by design. However, one may condition on $X \in \mathcal{S}$ for $\mathcal{S} \subset \mathcal{X}$ in which case

$$\mathbb{E}(\tau(X) \mid X \in \mathcal{S})$$

may be referred to as a *subgroup average treatment effect*, which is also a well-defined CATE.

The CATE is sometimes mistakenly reported in the literature as an *individual treatment effect* (ITE), which is a separate estimand that is only identified with more restrictive assumptions[2]. The ITE is defined at the unit level as the difference in potential outcomes; for unit i, the ITE is given by

$$F(X_i, Z_i = 1, \epsilon_i) - F(X_i, Z_i = 0, \epsilon_i) = \tau(x_i) + \delta(x_i, \epsilon_i). \tag{13}$$

This is unidentified without further assumptions on the nature of the error term when we do not have repeated measurements on individuals, as in general $\delta(x_i, \epsilon_i) \neq 0$ although it is mean-zero by construction; see Figure 2.

Finally, the average treatment effect among the treated (ATT) is defined as

$$\mathbb{E}(\tau(X) \mid Z = 1).$$

This estimand can be estimated under weaker conditions than the ATE; see section 3.11.

[2]Some authors use ITE to refer to *individualized* treatment effect, which they then define as the CATE. This is misleading and not recommended.

3 Examples

3.1 In what sense is randomization the "gold standard"?

It is often asserted, but in what sense is randomized treatment assignment the gold standard for treatment effect estimation? Surely solid-state physicists do not randomize their lab conditions and hope their sample size is large enough to reveal interesting results. Famously, esteemed physicist Ernst Rutherford quipped "If your experiment needs statistics, you ought to have done a better experiment" (Hammersley [1962]). The intuition behind this remark is that it is *control* that is central, not randomization. See section 3.6 for a definition of a control feature that evokes the experimental notion of "control".

Indeed, randomization is simply a way to guarantee control *on average* in the event that exact control is impossible, such as when crucial confounding factors are unobserved. This perspective in turn suggests that controlling for factors that we *can* observe and randomizing only for factors that we cannot observe would be the ideal approach. The following thought experiment amplifies this intuition.

Consider studying the effect of treatment Z on outcome Y in a sample of n pairs of identical twins and deciding how to allocate treatment across the $2n$ study participants. Completely randomized treatment assignment satisfies the assumptions outlined above and thus identifies the treatment effect. However, a naive randomization would sometimes accidentally treat both twins and leave other twin pairs untreated. This violates most people's intuition about why twin studies are interesting and useful, which is that giving one twin the treatment and the other a placebo implicitly "controls for" all of the shared biological and environmental factors that may impact the treatment effect. Randomization within each twin pair can protect against unmeasured factors that may confound the result, such as (perhaps) which twin was born first.

In this case, both Z and the twin pair index, X, are informative about the expected value of Y. Now consider four possible approaches to study the effect of Z on Y:

	Design	Estimator
1	Complete randomization	Unadjusted mean difference
2	Twin pair randomization	Unadjusted mean difference
3	Complete randomization	Adjusted mean difference
4	Twin pair randomization	Adjusted mean difference

where the unadjusted mean difference estimator is defined as

$$\bar{\tau}_U = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}$$

and the adjusted mean difference estimator is defined as

$$\bar{\tau}_A = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \left(\bar{Y}_{X=x,Z=1} - \bar{Y}_{X=x,Z=0} \right)$$

where $\mathcal{X}$ is the set of twin pairs and X is a variable that indexes twin pairs.

Each of the four approaches above identifies the ATE. However, adjusting for twin pairs (approaches 3 and 4) will tend to reduce variance over the unadjusted alternatives (1 and 2) and, similarly, designs that incorporate twin pairs in randomization (2 and 4) will also see a reduction in variance over the completely randomized alternatives (1 and 3).

3.2 Faithfulness

So why is it that satisfying the back-door criterion implies conditional unconfoundedness, but not vice-versa? This can happen, for instance, when Z and Y share a common cause, but it happens to have miraculously "counterbalanced" effects on Z and Y, such that the potential outcomes are marginally independent of Z. The condition of *faithfulness* is invoked to rule out such scenarios, asserting that the only conditional independencies exhibited by the data generating process are those implied by the causal

directed acyclic graph (CDAG). Here is a small numerical example of an "unfaithful" data generating process, corresponding to the causal diagram in Figure 1:

$$\begin{aligned}
\mathbb{P}(X = x) = p_x, \quad x \in \{1,2,3\}, \quad \mathbf{p} = \{p_1, p_2, p_3\} = \{16/60, 28/60, 16/60\}, \\
\mathbb{P}(Z = 1 \mid X) = \pi(X), \quad \pi(X) = \frac{5}{8}\mathbf{1}\{X == 1\} + \frac{5}{14}\mathbf{1}\{X == 2\} + \frac{5}{8}\mathbf{1}\{X == 3\}, \\
\mu(X) = \frac{1}{4}\mathbf{1}\{X == 1\} + \frac{1}{2}\mathbf{1}\{X == 2\} + \frac{3}{4}\mathbf{1}\{X == 3\}, \\
Y \leftarrow F(X, Z, \epsilon_y) = \mathbf{1}\{\epsilon_y < \mu(X) + \tau Z\}, \quad \tau = 0.2, \quad \epsilon_y \sim \text{Uniform}(0,1).
\end{aligned} \tag{14}$$

Note that $Y \mid X = x, Z = z \sim \text{Bernoulli}(\mu(x) + \tau z)$; above, this distribution is written in terms of an exogenous error ϵ_y to emphasize the functional causal model representation. The causal diagram has a backdoor path from Z to Y through X, and yet we may confirm that $Z \perp\!\!\!\perp (Y^1, Y^0)$ by direct calculation:

$$\mathbb{P}(Y^0 = 1 \mid Z = 1) = \frac{\sum_{x=1}^3 \mu(x)\pi(x)p_x}{\mathbb{P}(Z=1)} = \frac{\sum_{x=1}^3 \mu(x)(1-\pi(x))p_x}{\mathbb{P}(Z=0)} = \mathbb{P}(Y^0 = 1 \mid Z = 0)$$

and similarly for $\mathbb{P}(Y^1 \mid Z)$ which differs only by the constant $\tau = 0.2$. By contrast, in this example with binary Y and constant τ, no backdoor path would require that either $\pi(x)$ or $\mu(x)$ was constant in x, allowing it to be factored out of the above summations, yielding the desired independence. (Similar examples to this appear in several other papers, but this specific one was created from scratch for this paper.)

Faithfulness is an important condition in the context of inferring causal diagrams empirically, a problem known as *causal discovery*, but need not be invoked if a causal diagram is assumed to be in hand.

3.3 Mean conditional unconfoundedness

One of the major insights from a structural equations perspective is that conditional unconfoundedness is actually stronger than necessary for identifying treatment effects; the following weaker condition, *mean conditional unconfoundedness* is sufficient.

Definition 2. *A function s on covariate space $\mathcal{X}$ is said to satisfy* mean conditional unconfoundedness *if*

$$Z \perp\!\!\!\perp (\mu(X), \tau(X)) \mid s(X). \tag{15}$$

Lemma 2. *Mean conditional unconfoundedness is a sufficient condition for estimating average treatment effects.*

Proof. Denote the causal model as

$$Y^z \leftarrow \mu(X) + \upsilon(X, \epsilon_y) + (\tau(X) + \delta(X, \epsilon_y))z$$

where $\epsilon_y \perp\!\!\!\perp (Z, X)$, $\mathbb{E}(\upsilon(x, \epsilon_y)) = 0$, and $\mathbb{E}(\delta(x, \epsilon_y)) = 0$ for all x. We aim to show that

$$\mathbb{E}(Y^z \mid s(X) = s) = \mathbb{E}(Y \mid s(X) = s, Z = z),$$

from which the result follows by the estimability of the right hand side for both $z = 0$ and $z = 1$. Recalling the relationship between Y^z and $Y \mid Z = z$ described in Section 2.2.2, this is equivalent to showing that

$$\begin{aligned}
\mathbb{E}(\mu(X) + \upsilon(X, \epsilon_y) + (\tau(X) + \delta(X, \epsilon_y))z \mid s(X) = s) = \\
\mathbb{E}(\mu(X) + \upsilon(X, \epsilon_y) + (\tau(X) + \delta(X, \epsilon_y))z \mid s(X) = s, Z = z),
\end{aligned}$$

where the expectation over (X, ϵ_y) is with respect to its marginal distribution on the left hand side and with respect to its conditional distribution, given $Z = z$, on the right hand side. By the independence of ϵ_y, the mean zero errors for each x, and the linearity of expectation, this reduces to showing that

$$\mathbb{E}(\mu(X) + \tau(X)z \mid s(X) = s) = \mathbb{E}(\mu(X) + \tau(X)z \mid s(X) = s, Z = z).$$

By the assumption of mean conditional unconfoundedness, $Z \perp\!\!\!\perp (\mu(X), \tau(X)) \mid s(X)$, and the result follows. □

3.4 Propensity scores as stratification functions

Following Rosenbaum and Rubin [1983], the *propensity score* has become a central element in many applied analyses of causal effects.

Definition 3. *The* propensity score*, based on a vector of control variables x, is the conditional probability of receiving treatment:*

$$\pi(x) \equiv \mathbb{P}(Z = 1 \mid X = x). \tag{16}$$

It is common to interchangeably refer to the propensity *score*, which emphasizes a specific numerical value, $\pi(x)$, and the propensity *function*, which emphasizes the mapping, $\pi : \mathcal{X} \to (0, 1)$.

Rosenbaum and Rubin [1983] showed that $\pi(x)$ satisfies conditional unconfoundedness, from which it follows that

$$\text{ATE} = \mathbb{E}[Y^1 - Y^0] = \mathbb{E}_{\pi(X)}[\mathbb{E}[Y \mid \pi(X), Z = 1] - \mathbb{E}[Y \mid \pi(X), Z = 0]]. \tag{17}$$

This differs from the more general form of conditional unconfoundedness in that $\pi(X)$ is one-dimensional, while X itself typically involves many controls.

An especially common use of the propensity score in practice is via the inverse-propensity weighted (IPW) estimator

$$\bar{\tau}_{\text{ipw}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{Y_i Z_i}{\pi(X_i)} - \frac{Y_i(1 - Z_i)}{1 - \pi(X_i)} \right), \tag{18}$$

which is known to be consistent and has been widely studied theoretically. However, the propensity function itself is typically not known and so must be estimated.

How does the IPW estimator with estimated propensity scores differ from a regression adjustment. Well, in the case of a simple stratification estimator, it doesn't differ at all. It is straightforward to show, but not widely appreciated, that the empirical inverse propensity weighting (IPW) estimator is equivalent to $\bar{\tau}^x_{strat}$ under the following conditions:

1. $\mathcal{X}$ is discrete,
2. For all $x \in \mathcal{X}$, $N_{x,1} > 0$ and $N_{x,0} > 0$,
3. The propensity weighting function is estimated nonparametrically as $\hat{\pi}(x) = N_{x,1}/N_x$ for each $x \in \mathcal{X}$.

By direct calculation,

$$\begin{aligned}
\bar{\tau}^x_{ipw} &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i Z_i}{\hat{\pi}(X_i)} - \frac{Y_i(1 - Z_i)}{1 - \hat{\pi}(X_i)} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i Z_i}{N_{x_i,1}/N_{x_i}} - \frac{Y_i(1 - Z_i)}{N_{x_i,0}/N_{x_i}} \right) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i Z_i N_{x_i}}{N_{x_i,1}} - \frac{Y_i(1 - Z_i) N_{x_i}}{N_{x_i,0}} \right) \\
&= \frac{1}{n} \sum_{x \in \mathcal{X}} \left(\frac{N_x}{N_{x,1}} \left(\sum_{i:X_i = x} Y_i Z_i \right) - \frac{N_x}{N_{x,0}} \left(\sum_{i:X_i = x} Y_i(1 - Z_i) \right) \right) \\
&= \frac{1}{n} \sum_{x \in \mathcal{X}} \left(\frac{N_x}{N_{x,1}} \left(N_{x,1} \bar{Y}_{x,1} \right) - \frac{N_x}{N_{x,0}} \left(N_{x,0} \bar{Y}_{x,0} \right) \right) \\
&= \sum_{x \in \mathcal{X}} \frac{N_x}{n} \left(\frac{N_{x,1} \bar{Y}_{x,1}}{N_{x,1}} - \frac{N_{x,0} \bar{Y}_{x,0}}{N_{x,0}} \right) = \sum_{x \in \mathcal{X}} \frac{N_x}{n} \left(\bar{Y}_{x,1} - \bar{Y}_{x,0} \right) = \bar{\tau}^x_{strat}.
\end{aligned}$$

Notably, the propensity score might be a not-so-good stratification function, particularly in the case where π takes on many distinct values, leading to lots of stratification, but most of those strata are not confounded. The term of art for such strata is "instrumental"; this amounts to conditioning on variables that affect

treatment assignment but are not necessary for causal identification. It is typically unknown which elements of x are instrumental in this sense, but nonetheless it points out why thinking about regression adjustments fundamentally as propensity adjustment is misguided.

3.5 Prognostic scores

In data generating processes where variation in τ is independent of Z, the *prognostic score*, $\mathbb{E}(Y^0 \mid X = x) = \mu(x)$, is a sufficient control function. This follows because mean conditional unconfoundedness is satisfied trivially by $s(X) = \mu(X)$ when $\tau(X) \perp\!\!\!\perp Z$; see Lemma 2.

Like the propensity score, the prognostic score can be estimated from partially observed data — the propensity score can be estimated from (X, Z) pairs and the prognostic score can be estimated from control units only, $(X, Z = 0, Y)$, which in many contexts are more readily available than treated observations. See Hansen [2008] for a rigorous exposition of prognostic scores.

The vector-valued function (μ, τ) is a "generalized" prognostic score, containing both the usual prognostic score, as well as the treatment effect itself. This version of the prognostic score has received little attention, presumably because it "begs the question", in that one of its elements is the very estimand of interest. However, note that conditioning on a random variable is not about the values of that variable per se, but is rather about the level sets of the function defining that random variable. In particular, any one-to-one function of (μ, τ) also satisfies mean conditional unconfoundedness; knowledge of the treatment effect itself is not required, merely knowledge of which strata have distinct treatment effects.

3.6 Constant control function

The previous two examples showed that propensity scores and prognostic scores are sufficient control functions; this example demonstrates a function that may be coarser than either one. Consider a function on $\mathcal{X}$ defined as follows:

Definition 4. *A function s on $\mathcal{X}$ is a* constant control *function if for all $x, x' \in \mathcal{X}$ such that $s(x) = s(x')$ at least one of the following holds*

- $\pi(x) = \pi(x')$,
- $\mu(x) = \mu(x')$ *and* $\tau(x) = \tau(x')$.

In other words, a constant control function is a coarsening of $\mathcal{X}$ such that on each level set defined by s, either $\pi(x)$ or $(\mu(x), \tau(x))$ are constant. The following lemma shows that a constant control function defines a random variable $S = s(X)$ such that $\mathbb{E}(Y \mid Z = z, S) = \mathbb{E}(Y^z \mid S)$.

Lemma 3. *Assume X satisfies conditional unconfoundedness and consider the random variable $S = s(X)$, where s is a constant control function; then S satisfies conditional unconfoundedness.*

Proof. The proof is almost immediate. Consider a stratum of $S = s(X)$ consisting of all $x \in \mathcal{X}$ such that $s(x) = s$. By the definition of a constant control function, either $\pi(x)$ is constant for such x or else $(\mu(x), \tau(x))$ is constant. In either case, $Z \perp\!\!\!\perp (\mu(X), \tau(X)) \mid S = s$. In the former case Z is randomized according to the constant value of $\pi(x)$; in the latter case $(\mu(X), \tau(X))$ is the constant random variable and so is trivially independent of Z. □

The intuition behind a constant control function is that one way to control for "systematic co-variation" is simply to remove all variation. Clearly, both $\pi(X)$ and $(\mu(X), \tau(X))$ are themselves constant control functions, as is X itself.

However, a constant control function may be coarser than either, as illustrated in Figure 7, which shows an example of a simple data generating process that has a constant control function. In this example, $\mathcal{S}$ comprises just two strata, although μ and π take 10 and 11 unique values, respectively, and $|\mathcal{X}| = 20$. The treatment effect is heterogeneous but unconfounded: $\tau_i \overset{\text{iid}}{\sim} \text{U}(5, 10)$. The second panel of Figure 7 shows the

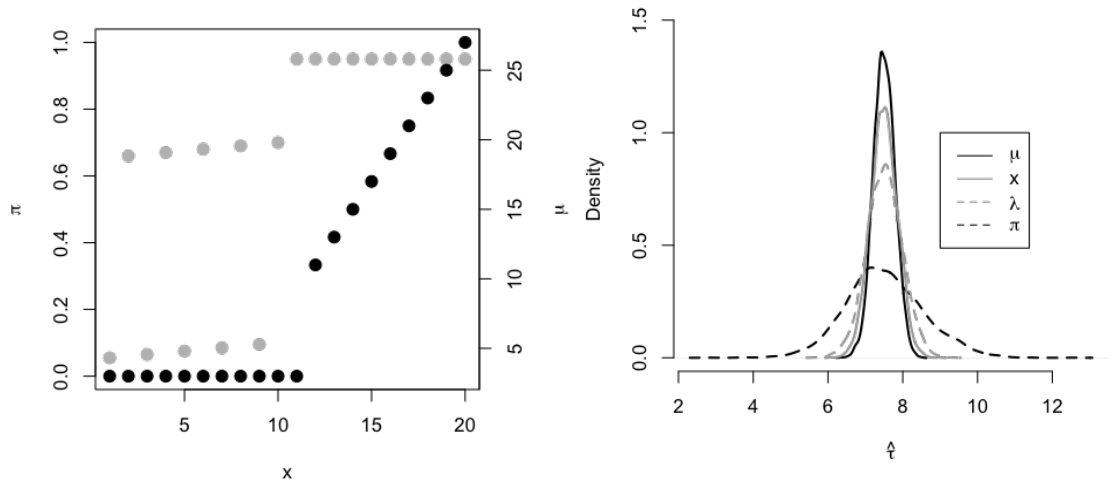

Figure 7: An example of a DGP admitting a simple constant control function, $\lambda = \mathbb{1}(x \leq 11)$. Here $\tau \sim \text{U}(5, 10)$ is heterogeneous and $x \in \{1, \ldots 20\}$. The left panel shows the μ values in black and the π values in gray. The right panel shows the sampling distributions of stratification estimators based on the level sets of different function: μ (solid black), x (solid gray), λ (dashed gray) and π (dashed black). All four estimators are unbiased, thought their variances differ markedly.

sampling distributions of four different stratification estimators: one using level sets of μ, one using level sets of π, one using all 20 values of x, and one using the two values of the minimal constant control function, indicating if $x \leq 11$. All four stratification estimators are unbiased, but exhibit differing variances: μ gives the lowest variance, followed by x, followed by the constant control function, followed by π.

3.7 Common cause confounding

Confounding is often explained to new students as a matter of adjusting for common causes. The graph corresponding to this simple explanation might look like the one in Figure 8, with three groupings of variables, only one of which connects the treatment Z and outcome variable Y. Consider variables X_j for $j = 0, \ldots, 7$ and a structural model for Z and Y with terms $\mu(x_1, x_2, x_3, x_4)$, $\tau(x_2, x_3, x_5, x_6)$, and $\pi(x_3, x_4, x_6, x_7)$. Note that X_0 plays no role in the structural model for Z or Y. In this case, $\{X_2, X_3, X_4\}$ is a necessary and sufficient set of controls. This would even be true if there were arrows connecting X_1 and $\{X_2, X_3, X_4\}$ and/or $\{X_2, X_3, X_4\}$ and $\{X_5, X_6, X_7\}$. However, it would not be true if there were arrows directly connecting X_1 and $\{X_5, X_6, X_7\}$. The next example illustrates the more realistic setting where confounding relationships are more complicated.

Note that for estimating the average treatment effect among the treated (ATT), one only needs to control for "prognostic confounding" and in this example that means that $\{X_2, X_3\}$ would suffice; see section 3.11 for more details.

3.8 The stacked-boxes diagram

Consider the causal diagram in Figure 5. Either the propensity controls (X_1, X_3) or the prognostic-moderation controls (X_2, X_4) are adequate for statistical control. However, Pearl's algorithm tells us that "mixed" variables also suffice, such as (X_1, X_4) or (X_2, X_3). Interestingly, such examples show that the notion of "instrumental" variables and "prognostic" variables are context dependent. Specifically, relative to a conditioning set of (X_2, X_3), additional stratification using X_4 is prognostic, while additional

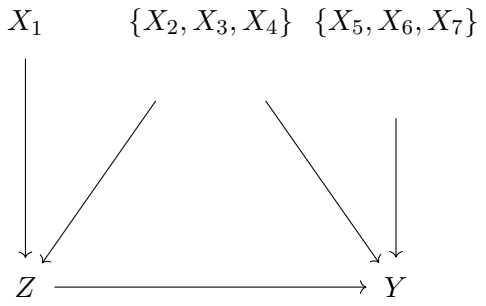

Figure 8: An elaboration of the basic triangle confounding graph. Direct inspection reveals that $\{X_2, X_3, X_4\}$ are necessary control variables for deconfounding. The structural model for Z and Y has terms $\mu(x_1, x_2, x_3, x_4)$, $\tau(x_2, x_3, x_5, x_6)$, and $\pi(x_3, x_4, x_6, x_7)$, as illustrated in the adjacent figure.

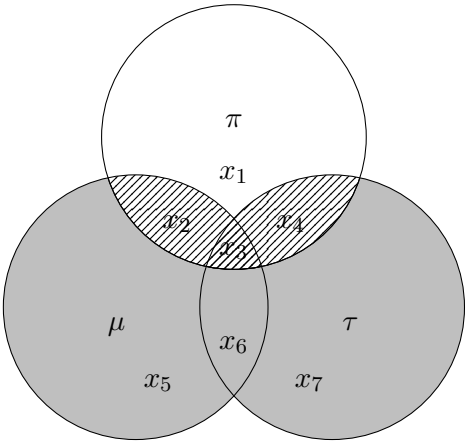

Figure 9: Necessary controls are variables appearing in both π and also in μ and/or τ. In this example, the necessary controls are $\{X_2, X_3, X_4\}$ as they fall in the cross-hatched region of the Venn diagram.

stratification on X_1 would be instrumental. Adding prognostic controls is often desirable, while adding instruments should be avoided, but such designations will fluctuate depending on what has already been included!

Similarly, invoking (conditionally) exogenous errors does not imply that the resulting mean components of the structural model are causal. In more detail, if the potential outcomes are defined in terms of the CDAG on the full set (X_1, X_2, X_3, X_4), a structural model can be derived that only involves (X_2, X_3), as follows:

$$\begin{aligned}
Y^0 &= F(x_1, x_2, x_3, x_4, z = 0, \epsilon_y) = F(x_2, x_4, z = 0, \epsilon_y) \\
Y^1 &= F(x_1, x_2, x_3, x_4, z = 1, \epsilon_y) = F(x_2, x_4, z = 1, \epsilon_y) \\
\mu(x_2, x_3) &\equiv \mathbb{E}(Y^0 \mid X_2, = x_2, X_3 = x_3) \\
\tau(x_2, x_3) &\equiv \mathbb{E}(Y^1 \mid X_2 = x_2, X_3 = x_3) - \mathbb{E}(Y^0 \mid X_2, = x_2, X_3 = x_3) \\
\upsilon(X_1, x_2, x_3, X_4, \epsilon_y) &\equiv F(X_1, x_2, x_3, X_4, z = 0, \epsilon_y) - \mu(x_2, x_3) \\
&= F(x_2, X_4, z = 0, \epsilon_y) - \mu(x_2, x_3) \\
\delta(X_1, x_2, x_3, X_4, \epsilon_y) &\equiv F(X_1, x_2, x_3, X_4, z = 1, \epsilon_y) - F(X_1, x_2, x_3, X_4, z = 0, \epsilon_y) - \tau(x_2, x_3) \\
&= F(x_2, X_4, z = 1, \epsilon_y) - F(x_2, X_4, z = 0, \epsilon_y) - \tau(x_2, x_3).
\end{aligned} \tag{19}$$

Noting that the resulting error terms now depend not only ϵ_y, but also on X_4, it is necessary to show that

$$(X_4, \epsilon_y) \perp\!\!\!\perp Z \mid (X_2, X_3).$$

But this follows from the fact that $\mathbb{E}(Z \mid X_2 = x_2, X_3 = x_3) = \mathbb{E}(\pi(X_1, X_3) \mid X_2 = x_2, X_3 = x_3) \equiv \pi(x_2, x_3)$ is free of X_4. In this model, $\mu(x_2, x_3)$, $\tau(x_2, x_3)$ and $\pi(x_2, x_3)$ must not be interpreted as causal functions, despite yielding the required exogenous errors; specifically, from the graph we know that X_2 has no causal impact on Z and X_3 has no causal impact on Y, as depicted in Figures 5.

3.9 CDAGs are non-unique

This example considers a data generating process that admits distinct CDAGs, depending on how the control variables are parametrized. This scenario is not commonly discussed, presumably because observed

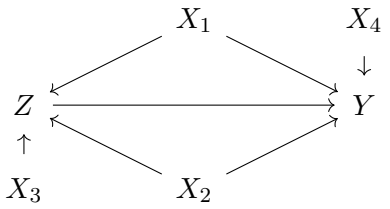

Figure 10: Causal graph in terms of original covariates

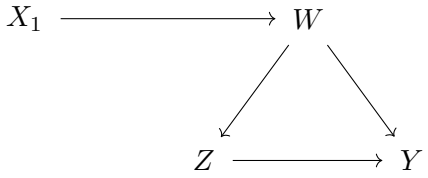

Figure 11: Causal graph under transformed covariates

measurements are taken to be designated by "nature", so to speak. However, reflecting on invertible transformations such as $(x_1, x_2) \to (x_1, x_1/x_2)$ highlights that functional causal models are, in fact, subject to changes of variables.

More concretely, consider the following DGP:

$$\begin{aligned} X_j &\overset{iid}{\sim} \text{Bernoulli}\,(p_j) \\ \pi(X) &= \beta_0 + \beta_1(2X_1X_2 - X_1 - X_2 + 1) + \beta_2 X_3 \\ Z &\sim \text{Bernoulli}\,(\pi(X)) \\ \mu(X) &= \alpha_0 + \alpha_1(2X_1X_2 - X_1 - X_2 + 1) + \alpha_2 X_4 \\ \tau(X) &= \tau \quad \text{(constant treatment effect)} \\ Y &= \mu(X) + \tau(X)Z + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right) \end{aligned}$$

Next, define random variable $W = (2X_1X_2 - X_1 - X_2 + 1)$, regarding X_2 as the exogenous variable in the functional model for $W \mid X_1$. Additionally, suppressing X_3 and X_4, as they represent exogenous variation, yields the causal graph in Figure 11.

From this graph, it is clear that conditioning on W satisfies conditional unconfoundedness. Most interestingly, $|\mu(\mathcal{X})| = |\pi(\mathcal{X})| = 4$. while $|\mathcal{W}| = 2$; thus W provides the smallest possible random variable for deconfounding.

3.10 Quantile treatment effects

The structural model perspective permits us to produce, starting from a given CDAG, a modified causal diagram that reflects only the mean dependencies. For estimation of average causal differences, such a graph suffices to identify valid control variable sets that are potentially smaller than any control set satisfying the back-door criterion on the original CDAG.

For example, consider the following data generating process:

$$\begin{aligned} X &\sim \text{Bernoulli}(1/2), \\ Z &\sim \text{Bernoulli}\,(1/4 + X/2) \\ Y &\sim \mathcal{N}\left(\tau Z, (\sigma + X)^2\right) \end{aligned}$$

For this DGP, $\mu(X) = 0$ and $\tau(X) = \tau$ are both constant in X, which implies that the null set satisfies mean conditional unconfoundedness; even though X is a common cause of Z and Y, it only affects the variance of Y, but not the mean. Therefore, the full joint distribution of X, Z, Y is the triangle diagram of Figure 6, while Figure 12 depicts the joint distribution of $(X, Z, \mathbb{E}(Y \mid X, Z))$, in which X is unconnected to $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z)$.

Note that while mean conditional unconfoundedness identifies the ATE, it does not identify other causal estimands. For instance, consider the quantile treatment effect (QTE), for $q \in (0, 1)$:

$$F_{Y^1}^{-1}(q) - F_{Y^0}^{-1}(q)$$

where F^{-1} denotes an inverse cumulative distribution function.

Integrating out X, $Y \mid Z = z$ is a mixture of two normal random variables, with PDF and CDF defined as

$$f(y \mid Z = z) = w_z \phi(y, \tau z, (\sigma + 1)^2) + (1 - w_z)\phi(y, \tau z, \sigma^2),$$
$$F(y \mid Z = z) = w_z \Phi(y, \tau z, (\sigma + 1)^2) + (1 - w_z)\Phi(y, \tau z, \sigma^2)$$

where $w_z = \mathbb{P}(X = 1 \mid Z = z)$. By contrast, the PDF and CDF of Y^z are given by

$$f(y^z \mid Z = z) = \frac{1}{2}\phi(y, \tau z, (\sigma + 1)^2) + \frac{1}{2}\phi(y, \tau z, \sigma^2),$$
$$F(Y^z \mid Z = z) = \frac{1}{2}\Phi(y, \tau z, (\sigma + 1)^2) + \frac{1}{2}\Phi(y, \tau z, \sigma^2).$$

Because $X \not\perp\!\!\!\perp Z$, $w_z \neq 1/2$ and therefore

$$F^{-1}_{Y^1}(q) - F^{-1}_{Y^0}(q) \neq F^{-1}_{Y|Z=1}(q) - F^{-1}_{Y|Z=0}(q),$$

as illustrated in Figure 13.

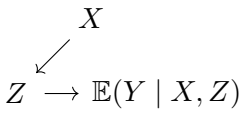

Figure 12: Mean causal graph

3.11 Partial randomization

Some estimands require weaker assumptions than estimating the average treatment effect over the whole population does. For example, the *average treatment effect among the treated*, or ATT, is defined as $\mathbb{E}(Y^1 - Y^0 \mid Z = 1) = \mathbb{E}(Y^1 \mid Z = 1) - \mathbb{E}(Y^0 \mid Z = 1)$[3]. This estimand is important in the program evaluation literature, see for example Heckman [1996] and Heckman et al. [1997].

Here we use structural model notation to compare the ATT to the ATE, as relates to the "naive" contrast that compares the average response among treated individuals to the average response among the untreated individuals. In terms of the population, the naive contrast estimates $\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)$. In terms of the structural model, this is equivalent to

$$\mathbb{E}(\mu(X) + \tau(X) \mid Z = 1) - \mathbb{E}(\mu(X) \mid Z = 0).$$

By definition, the exogenous errors are mean zero and vanish from the above expression. Now, randomization of Z implies that $(\mu(X), \tau(X)) \perp\!\!\!\perp Z$, which in turn implies that $\mathbb{E}(\mu(X) \mid Z = 1) = \mathbb{E}(\mu(X) \mid Z = 0)$ and therefore that

$$\mathbb{E}(\mu(X) + \tau(X) \mid Z = 1) - \mathbb{E}(\mu(X) \mid Z = 0) = \mathbb{E}(\tau(X) \mid Z = 1),$$

the ATT. Randomization further implies that $\mathbb{E}(\tau(X) \mid Z = 1) = \mathbb{E}(\tau(X))$, so that the ATE and the ATT are the same.

However, the above derivation also reveals that to estimate the ATT one only needs $\mathbb{E}(\mu(X) \mid Z = 1) = \mathbb{E}(\mu(X) \mid Z = 0)$, or what we might call *mean prognostic unconfoundedness*, which itself follows from

[3] In our experience, this potential outcomes notation for the ATT can give students fits, particularly the $\mathbb{E}(Y^0 \mid Z = 1)$ term. Such students may find the structural equation notation to be somewhat more transparent: $\mathbb{E}(\tau(X) \mid Z = 1)$ makes it clear that the probabilistic impact of conditioning on $Z = 1$ is to modify the distribution over X defining the expectation; there is no opportunity for cognitive interference from the fact that the "z" in Y^z is different from that in the condition $Z = z$.

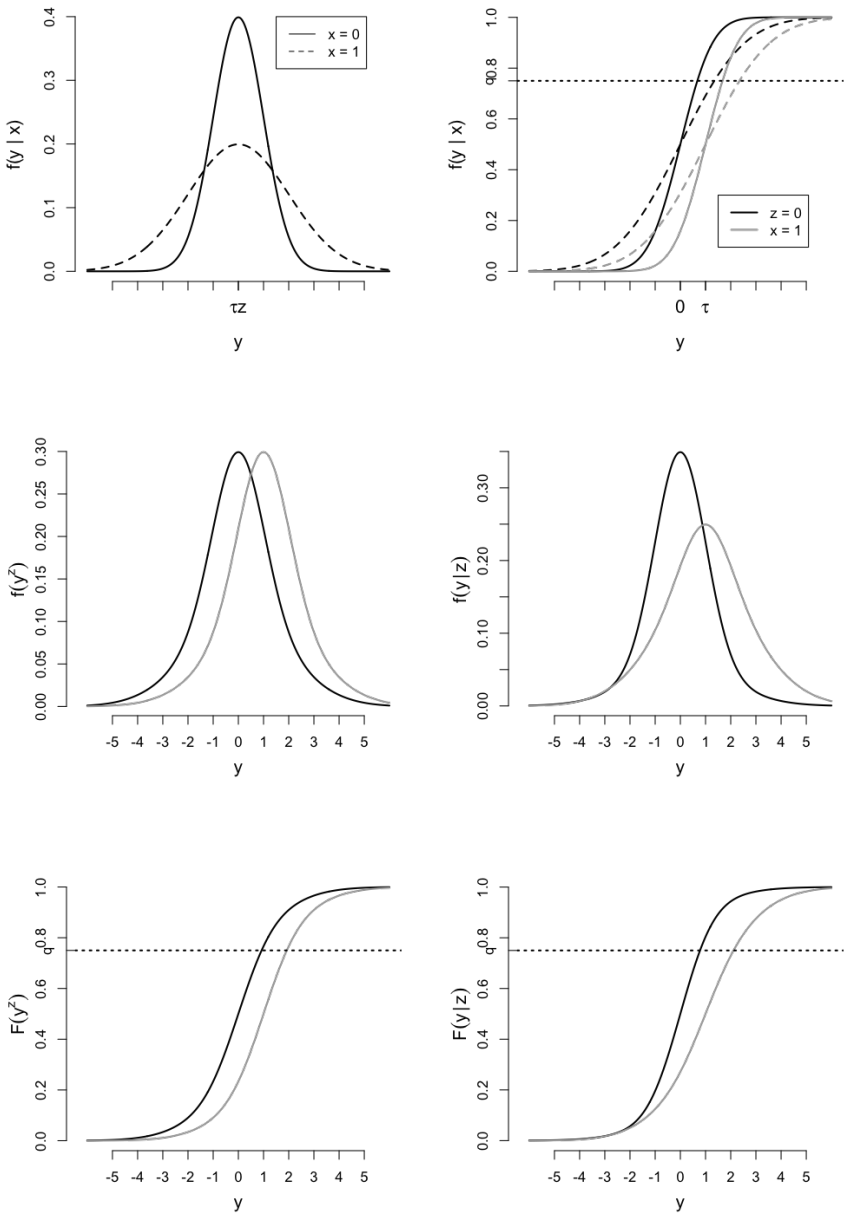

Figure 13: An illustration of a confounded quantile treatment effect with unconfounded ATE. The top two panels depict the density and CDF functions of the DGP from section 3.10 for the four combinations of $X \in \{0, 1\}$ and $Z \in \{0, 1\}$. For each value of X the change in the quantile is a constant shift to the right. The second row shows the densities of the potential outcome distributions and the conditional distribution of $Y \mid Z$, respectively, with X integrated out. In both cases, the resulting density is a mixture of two normals with different variances and a common mean. However, the potential outcomes densities are just translations of the same mixture density, whereas the conditional distribution of $Y \mid Z$ also differs in terms of the mixture weights. The bottom row depicts the same relationship, but in terms of the CDFs. Attempts to estimate the quantile treatment effect — shown here as the distance between the black and grey curves at the horizontal dashed line in left panel — using the analogous distance from the right panel would misestimate the effect.

$\mu(X) \perp\!\!\!\perp Z$, or *prognostic unconfoundedness*. Thus, when the ATT is the sole interest, one only needs to rule out prognostic confounding[4]. Meanwhile, treatment effect confounding, $\tau(X) \not\perp\!\!\!\perp Z$, entails that the ATT and ATE are different, so that the ATE remains unknown even with the ATT in hand.

As it relates to feature selection, it is notable that a smaller feature set may allow estimating the ATT than would be required for estimating the ATE. The following DGP is a concrete example:

$$\begin{aligned} X_1 \sim \text{Bernoulli}(1/2), \quad X_2 \sim \text{Bernoulli}(1/2), \\ Z \mid X_1, X_2 \sim \text{Bernoulli}(0.25 + 0.5X_2), \\ Y \mid X_1, X_2, Z \sim \mathcal{N}(X_1 + (1 + 2X_2)Z, \sigma^2). \end{aligned}$$

In this example, $\tau(X) = \tau(X_2) = 1+2X_2$, $\mu(X) = \mu(X_1) = X_1$, and the ATE is $\mathbb{E}(\tau(X)) = 1+2\mathbb{E}(X_2) = 2$. The ATT, on the other hand, is $\mathbb{E}(\tau(X) \mid Z = 1) = 1 + 2\mathbb{E}(X_2 \mid Z = 1) = 3/2$. It is a nice simulation exercise to verify that the naive contrast is consistent for the ATT, but not the ATE.

3.12 A two-stage estimator using two distinct control features.

This example builds upon the ideas presented in the previous one, but returns to the goal of regression adjustments for the ATE.

Suppose we know that $\mu(X) \perp\!\!\!\perp Z \mid s_1(X)$ and $\tau(X) \perp\!\!\!\perp Z \mid s_2(X)$, for distinct functions (features) s_1 and s_2. One approach to estimating the ATE under this assumption would be to stratify on the common refinement of $s_1(X)$ and $s_2(X)$, thus guaranteeing that $(\mu(X), \tau(X)) \perp\!\!\!\perp Z \mid s(X) = s_1(X) \vee s_2(X)$. But an alternative two-stage approach is possible, which requires estimating fewer individual strata means. The procedure is:

1. Estimate $\mu(s_1(X)) = \mathbb{E}(Y \mid Z = 0, s_1(X))$ from the control data.
2. Define $R = Y - \mu(s_1(X))$.
3. Estimate $\mathbb{E}(R \mid Z = 1, s_2(X))$ from the treated data.
4. Compute the ATE as $\mathbb{E}_X(\mathbb{E}(R \mid Z = 1, s_2(X)))$, where the outer expectation is over X, with respect to its marginal distribution.

We may verify the validity of this estimator by first expressing the procedure as the following iterated expectation:

$$\begin{aligned} &\mathbb{E}_X\left(\mathbb{E}(Y - \mathbb{E}(Y \mid Z = 0, s_1(X)) \mid Z = 1, s_2(X))\right) \\ &\quad = \mathbb{E}_X\left(\mathbb{E}(Y \mid Z = 1, s_2(X))\right) - \mathbb{E}_X\left(\mathbb{E}(Y \mid Z = 0, s_1(X))\right) \\ &\quad = \mathbb{E}_X\left(\mathbb{E}(\mu(X) + \tau(X) \mid Z = 1, s_2(X))\right) - \mathbb{E}_X\left(\mathbb{E}(\mu(X) \mid Z = 0, s_1(X))\right) \\ &\quad = \mathbb{E}_X\left(\mathbb{E}(\tau(X) \mid Z = 1, s_2(X))\right) + \mathbb{E}_X\left(\mathbb{E}(\mu(X) \mid Z = 1, s_2(X))\right) - \mathbb{E}_X\left(\mathbb{E}(\mu(X) \mid Z = 0, s_1(X))\right). \end{aligned}$$

By the assumption that $\mu(X) \perp\!\!\!\perp Z \mid s_1(X)$, we find that $\mathbb{E}(\mu(X) \mid Z = 0, s_1(X)) = \mathbb{E}(\mu(X) \mid Z = 1, s_1(X))$, which in turn implies that the second and third terms above are both equal to $\mathbb{E}_X(\mu(X) \mid Z = 1)$ (just expressed as distinct iterated expectations) and thus cancel. By the assumption that $\tau(X) \perp\!\!\!\perp Z \mid s_2(X)$, the remaining term is equal to $\mathbb{E}(\tau(X) \mid s_2(X))$ and the desired result follows after taking the outer expectation: $\mathbb{E}(\tau(X)) = \mathbb{E}_X(\mathbb{E}(\tau(X) \mid s_2(X))$.

[4] An analogous argument works for $\mathbb{E}(\tau(X) \mid Z = 0)$, the average effect of the treatment on the control (untreated) population, or ATC. This is easiest to see by reparametrizing the structural model in terms of: $Z^* = 1 - Z$, $\mu^*(X) = \mu(X) + \tau(X)$, and $\tau^*(X) = -\tau(X)$. It then follows that the ATC may be estimated from the naive contrast so long as $\mu^*(X) \perp\!\!\!\perp Z$.

4 Additional Remarks and Further Reading

4.1 Overstated virtues of the propensity score.

Rosenbaum and Rubin [1983] is often cited in support of propensity score methods for causal inference, but its results are often over-stated. First, there is not one propensity score, but many, one corresponding to each valid set of control features. Second, a propensity score need not be minimal; it is the minimal balancing score for the complete set of features used to create it, but balancing on those features is not necessary to estimate causal effects. Third, a propensity score method that disregards important prognostic features can be much less efficient than a method that does incorporate such features.

4.2 Estimated versus True propensity scores.

In practice, the propensity score (corresponding to a given set of control features) is rarely known and so must be estimated. Hirano et al. [2003] is sometimes cited to put a positive spin on this state of affairs: estimating a propensity function is better than knowing it exactly! But the actual situation is more nuanced. The asymptotic analysis of Hirano et al. [2003] comparing the IPW estimator using true versus estimated propensity scores conceals the variety of specific ways the two estimators differ. Viewing the IPW as a stratification estimator in the discrete covariate setting puts these distinctions into immediate relief [Hahn and Herren, 2025]. One, the IPW using the true propensity scores uses different strata weights than the one using the estimated propensity scores, resulting in a higher variance estimator. Two, the IPW based on a true propensity score is able to collapse unnecessary strata, which can reduce the variance of the estimator. Three, collapsing unnecessary strata does not *always* reduce the variance, because the "extraneous" strata may be informative about *unconfounded* variation in the response. That is, an IPW estimator based on estimated propensity scores can have lower variance than one based on a true propensity score because it performs an implicit regression adjustment that is essentially unrelated to the propensity score.

4.3 Regression adjustments for randomized experiments.

Freedman [2008] is sometimes cited as a reason to avoid regression adjustment for causal effect estimation altogether. However, Freedman's result was more about model specification — or *mis*specification — than it was about regression adjustment per se. Provided that one undertakes a nonparametric adjustment, as advocated by Lin [2013], Freedman's main concerns are addressed. However, nonparametric adjustment poses its own challenges, in the form of high-variance estimators. Whether or not the inclusion of strong prognostic features is enough to offset the increased variability that comes with estimating a nonparametric model with limited data is impossible to say in any generality. However, since the publication of a method called CUPED in Deng et al. [2013] regression adjustment (specifically based on pre-treatment outcomes) has become much more widespread, particularly in the world of online experiments (so-called "A/B" testing).

4.4 The peril of colliders.

Greenland et al. [1999] introduce the "M-Graph" and the problem of conditioning on unblocked colliders. The issue was vigorously debated in a series of articles and replies in *Statistics in Medicine* between 2007 and 2009. Rubin [2007] suggested that all available pre-treatment covariates should be included in the conditioning set of any observational causal analysis, while others (Shrier [2008]; Sjölander [2009]; Pearl [2009b]) contended that such a strategy could incur collider bias. Rubin [2009] responded that unblocked colliders are a stylized problem that has few practical ramifications. This exchange in turn motivated further research, including Ding and Miratrix [2015], Rohde [2019], and Cinelli et al. [2020]. Here, we observed that should colliders appear in a set of control variables — along with the associated blocking variables — regularization can unintentionally induce collider bias, revealing that colliders are not only a

problem when their parents are unobserved. In particular, regularized regression approaches will struggle with colliders that are blocked by only a propensity-side ancestor.

4.5 Conditional unconfoundedness versus mean conditional unconfoundedness.

In a discussion of Angrist et al. [1996], Heckman [Heckman, 1996] makes a point similar to the one we make in section 3.10, that conditional unconfoundedness is stronger than necessary for estimating certain treatment effects. Angrist rejoins that identification based on "functional form" is undesirable. Here, we have taken the perspective of Heckman, as mean conditional unconfoundedness is the key notion for defining the principal deconfounding function, so it is perhaps worthwhile to unpack why. Our interest was in understanding the conditions according to which a particular set of control variables would yield a valid stratification estimator. From this perspective, a more *specific* assumption is *weaker* than a more general one: Conditional unconfoundedness implies mean conditional unconfoundedness, but not the other way around. It is the specificity of the *estimand* that permits the weaker (more general) assumption on the DGP. As we explored in section 3.10, mean conditional unconfoundedness does not permit estimation of quantile treatment effects. In order for mean conditional unconfoundedness to license estimation of quantile treatment effects, one would need to impose additional restrictions on the DGP, such as a fixed distributional shape around the unconfounded mean. But that is not our suggestion (nor do we believe it was Heckman's).

Interestingly, this distinction between conditional unconfoundedness and mean conditional unconfoundedness is at the heart of the the difference between general causal diagrams and more traditional path analysis. By focusing on correlations, the path diagram must only respect the mean causal relationships. Sometimes this is described by saying that path analysis "has a structural model, but no measurement model" (Wikipedia).

4.6 Methodological ecumenicalism

In section 2.4, it was shown that the potential outcomes, CDAG, and exogenous errors definitions of conditional unconfoundedness are substantively equivalent. This result allows us to conveniently move between the conventions of these alternative frameworks, which implicitly emphasize distinct aspects of the problem they all address — estimating treatment effects from data.

For example, the causal graph approach reminds us that sets of valid control variables are not unique and, consequently, we must not speak of *the* propensity score, but rather *a* propensity score and, perhaps, many candidate propensity scores (cf. section 3.4). This observation is fundamental to understanding how regularization will impact bias due to feature selection on graphs including colliders and instruments.

The potential outcomes approach reminds us that the exogenous errors need not be common among the treatment arms (cf. figure 1). More generally, because the potential outcome notation is intrinsically individualized, it emphasizes the idea that some individuals in a population may have distinct causal diagrams; in particular, some arrows may not appear in every individual's graph. This is not at odds with the graphical formalism; rather it emerges simply because the graph alone does not fully determine the data generating process. In this paper, this distinction is not particularly important, but in estimation techniques relying on instrumental variables, it becomes critical [Angrist et al., 1996].

From the exogeneous errors approach, we are reminded that full conditional unconfoundedness is not actually necessary to estimate particular causal effects (cf. section 3.10).

Synthesizing the three methods also clarifies common misunderstandings that can occur when operating solely within a single framework; for example, a mean regression model with exogeneous additive errors need not be structural (e.g., causal) in all of its arguments — rather, the exogeneity of the errors narrowly licenses a causal interpretation with respect to the treatment variable (cf. section 3.8).

4.7 Probabilistic graphical models.

Probabilistic graphical models [Koller and Friedman, 2009], also referred to as "Bayes nets", overlap with the material on causal graphs presented above. This is no mere coincidence, as Judea Pearl was a key figure in the development of both topics. However, causal diagrams are a strict subset, rather than a rebranding, of the many applications of Bayes nets; one that endows joint distributions with causal meaning. In particular, while a joint distribution on p random variables has $p!$ distinct compositional representations, only one of those corresponds to a causal representation; a generic Bayes net has no intrinsic relationship to matters of causality. Much of the interest in Bayes nets was motivated by the utility of the graphical representation in developing efficient algorithms for computing quantities such as conditional expectations or establishing conditional or marginal independence between particular variables, tasks that have nothing to do with causality per se. To put a finer point on it, causal diagrams are not "just" Bayes nets by another name.

4.8 Path analysis.

The graphical and exogenous errors approaches have a common antecedent in the area of *path analysis* and Structural Equation Modeling, or SEM. Path analysis (see Shipley [2016] for a thorough overview) grew out of the previously mentioned work of Sewell Wright, who was the first to combine mean regression and graphical representations with the intent of furnishing causal inferences. However, the modern areas of causal diagrams and structural regression with exogenous errors developed these ideas in distinct directions.

Modern causal diagrams generalized the path diagrams of Wright to represent arbitrary joint distributions on random variables rather than merely partial correlations. This generalization naturally shifted the focus away from an emphasis on covariance matrices and methods for estimating them.

Economists continued to pursue the mean regression perspective, but relaxed the exhaustiveness assumption — at least in the context of treatment effect estimation with a single treatment (as is common in applied economics work on policy evaluation). See remark **??** and section 3.8 for more details on this distinction. Consequently, the full causal structure among all variables was de-emphasized and the exogeneity assumption became central, making the graphical component of path analysis superfluous.

Meanwhile, path analysis was adopted by social scientists outside of economics starting in the 1950's and continued to be used and developed in its original mold, focused on tracing the causal impact of variables through intermediate effects on other variables (i.e. mediation analysis) and doing so primarily in the context of linear models. Consequently, research in path analysis strengthened relationships with linear factor models and methods for their estimation, topics that are far less prominent in contemporary work on causal diagrams or regression adjustment for causal effect estimation.

4.9 Other graphical model approaches

There are several other approaches to causal modeling using graphs. Richardson and Robins [2013] introduce Single-World Intervention Graphs (SWIGs) as a generalization of DAGs and potential outcomes. Dawid [2015] and Dawid [2021] discuss an approach to causal inference that relies on decision theory and offers alternative interpretations of both the DAG and potential outcome frameworks.

References

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

C. Cinelli, A. Forney, and J. Pearl. A crash course in good and bad controls. *Available at SSRN*, 3689437, 2020.

A. P. Dawid. Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Applications*, 2:273–303, 2015.

P. Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021.

A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132, 2013.

P. Ding and L. W. Miratrix. To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57, 2015.

D. A. Freedman. On regression adjustments in experiments with several treatments. *The annals of applied statistics*, 2(1):176–196, 2008.

S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12, 1943.

P. R. Hahn and A. Herren. True versus estimated propensity scores with discrete controls: a finite sample analysis. Technical report, 2025.

J. Hammersley. Monte carlo methods. In *Proceedings of the Seventh Conference on Design of Experiments for Army Research Development and Testing*, volume 7, pages 17–26. US Army Research Office, 1962.

B. B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

J. J. Heckman. Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, 91(434):459–462, 1996.

J. J. Heckman and E. Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.

J. J. Heckman and E. J. Vytlacil. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874, 2007.

J. J. Heckman, H. Ichimura, and P. E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

J. Pearl. Embracing causality in formal reasoning. In *AAAI*, pages 369–373, 1987.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. *Causality*. Cambridge University Press, 2009a.

J. Pearl. Remarks on the Method of Propensity Scores. *Statistics in Medicine*, 28(9):1416–1420, 2009b.

J. Pearl and T. Verma. *The logic of representing dependencies by directed graphs*. University of California (Los Angeles). Computer Science Department, 1987.

J. Pearl and T. S. Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

T. S. Richardson and J. M. Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

D. Rohde. A bayesian solution to the m-bias problem. *arXiv preprint arXiv:1906.07136*, 2019.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007.

D. B. Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423, 2009.

C. Shalizi. *Advanced data analysis from an elementary point of view*. Cambridge University Press, 2021. URL https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

B. Shipley. *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press, 2016.

I. Shrier. Letter to the Editor. *Statistics in Medicine*, 27(14):2740–1, 2008.

A. Sjölander. Propensity Scores and M-Structures. *Statistics in Medicine*, 28(9):1416–1420, 2009.

S. Wright. On the nature of size factors. *Genetics*, 3(4):367, 1918.

S. Wright. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences of the United States of America*, 6(6):320, 1920.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 22:557–585, 1921.