Causality

Lecture Notes

Niklas Pfister

August 21, 2024

1 Causal Models

A causal model, in its most basic form, is an extension of a statistical model that not only describes a system in its observed states but also under a set of well-defined hypothetical interventions. Conceptually, a statistical model specifies a set of distributions,

$$\mathcal{P} \subseteq \{P \mid P \text{ probability distribution on } \mathcal{X}\},$$

such that the observed data is assumed to be a random sample $X \sim P_0$ for a fixed $P_0 \in \mathcal{P}$. In contrast, for a fixed index set $\mathcal{I}$, a causal model specifies a set of distribution valued functions with domain $\mathcal{I}$,

$$\mathcal{P}_{\mathcal{I}} \subseteq \{\overline{P} \mid \overline{P} : \mathcal{I} \to \mathcal{P}\}.$$

We call $\mathcal{I}$ the set of *interventions* and assume that for all interventions $i \in \mathcal{I}$ data is modeled as a random sample $X \sim \overline{P}_0(i)$ for a fixed $\overline{P}_0 \in \mathcal{P}_{\mathcal{I}}$. We further assume that the set of interventions can be divided into a set of *observed interventions* $\mathcal{I}_{\text{obs}} \subseteq \mathcal{I}$ – often this only contains a single element corresponding to the observed setting – and a set of *hypothetical interventions* $\mathcal{I}_{\text{hyp}} \subseteq \mathcal{I}$. Formally, we assume $\mathcal{I}_{\text{obs}} \cap \mathcal{I}_{\text{hyp}} = \emptyset$ and $\mathcal{I}_{\text{obs}} \cup \mathcal{I}_{\text{hyp}} = \mathcal{I}$. Similar to how in a statistical model we define statistical estimands as real-valued functionals on $\mathcal{P}$, we define *causal estimands* as functionals $\Psi : \mathcal{P}_{\mathcal{I}} \to \mathbb{R}$.

Example 1 (Average causal effects)**.** *Assume a setting in which we want to determine whether taking an asprin helps removing a headache. Consider now a study population where for each participant we measure some covariates X capturing some characteristics of the participants (e.g., age and gender), a response $Y \in \{0, 1\}$ encoding whether they had a headache at the end of the experiment and a treatment indicator $T \in \{0, 1\}$ encoding whether they received an asprin or not. We can use a statistical model describing (T, X, Y) for each participant, for example, by assuming they are i.i.d. draws from a distribution P_0. The distribution P_0, however, is insufficient to formally express the original goal of determining the effect of asprin as it does not allow us to describe changes in the model.*

Instead, we need to first specify what an effect of asprin means. One way of doing this is to imagine two hypothetical experiments. In the first we give all participants the asprin (denoted by 'ALL'*) and in the second we give none of the participants an asprin (denoted by* 'NONE'*). If we would observe data from both of these experiments, we could define an* average causal effect *of asprin as the difference in the expectation of Y in the two experiments.*

Formally, we define the set of interventions $\mathcal{I} := \{\varnothing, \text{ALL}, \text{NONE}\}$, where $\varnothing$ corresponds to observed setting described above (i.e., $\mathcal{I}_{\text{obs}} := \{\varnothing\}$ and $\mathcal{I}_{\text{hyp}} := \{\text{ALL}, \text{NONE}\}$). As in the statistical model, we now assume that under each intervention $i \in \mathcal{I}$ the measurements for all participants are modeled as i.i.d. draws from

$$(T, X, Y) \sim \overline{P}_0(i),$$

where $\overline{P}_0(i)$ is a fixed distribution. The average causal effect is then defined by

$$\mathbb{E}_{\overline{P}_0(\text{ALL})}[Y] - \mathbb{E}_{\overline{P}_0(\text{NONE})}[Y],$$

which captures the difference in expectation of having a headache across the two hypothetical interventions. This target quantity can be expressed as the causal estimand $\Psi : \mathcal{P}_{\mathcal{I}} \to \mathbb{R}$ defined for all $\overline{P} \in \mathcal{P}_{\mathcal{I}}$ by

$$\Psi(\overline{P}) = \mathbb{E}_{\overline{P}(\text{ALL})}[Y] - \mathbb{E}_{\overline{P}(\text{NONE})}[Y].$$

Just as in an (observational) statistical model an estimand is not necessarily identifiable even if infinite data are available. However, while for (most) statistical estimands identifiability follows from assuming regularity conditions on the statistical model $\mathcal{P}$ (e.g., smoothness), the unidentifiability for causal estimands can be more substantial as it can depend on interventional distributions for which no data is available.

Definition 1 (Causal identifiability)**.** *Let $\mathcal{P}_{\mathcal{I}}$ be a causal model based on the interventions $\mathcal{I} = \mathcal{I}_{\text{obs}} \cup \mathcal{I}_{\text{hyp}}$. Let $\Psi : \mathcal{P}_{\mathcal{I}} \to \mathbb{R}$ be a causal estimand. We say Ψ is* causally identifiable *if there exists $\Phi : \{\overline{P}|_{\mathcal{I}_{\text{obs}}} \mid \overline{P} \in \mathcal{P}_{\mathcal{I}}\} \to \mathbb{R}$ satisfying for all $\overline{P} \in \mathcal{P}_{\mathcal{I}}$ that*

$$\Phi(\overline{P}|_{\mathcal{I}_{\text{obs}}}) = \Psi(\overline{P}).$$

A causal model as defined here is not capable of modeling counterfactual statements. While there are certainly valid reasons to consider such statements (e.g., applications in fairness), we avoid them here as they involve delicate philosophical reasoning that if not explicitly needed is best avoided. We use the abstract notion of a causal model to provide a unified perspective on the more specialized causal models that exist in the literature and have been developed (often independently of each other) for different types of applications. Each model has its own advantages and disadvantages and none is strictly superior to another. Most importantly, all of them can be shown to induce distributions for observed and hypothetical interventions and are hence causal models as defined above. Moreover, any valid causal analysis based solely on these distributions results in the same conclusions.

Potential outcome models Potential outcome models (sometimes also called Rubin causal models) were originally developed to formalize a notion of causality in (randomized) control trials and are still the most prevalent models used in clinical and biomedical settings. They start from a well-defined set of hypothetical interventions and introduce individual random variables (called potential outcomes) for each possible intervention.

Advantages: The potential outcomes are easy to communicate to practitioners as most humans are comfortable thinking in terms of counterfactuals. Moreover, the framework avoids explicitly specifying causal mechanisms (as long as they are not needed). Lastly, since these models start from individual units, it can be easier to specify certain delicate causal assumptions between these units (e.g., interference) than with other causal models.

Disadvantages: A common criticism is that these models obscure which quantities are counterfactual (and potentially non falsifiable) and which are interventional (and falsifiable by experimentation) [e.g., Dawid, 2021]. A further disadvantage is that as soon as multiple sequential causal relations are present (e.g., in mediation analysis or other settings with complex causal structure) the models become complex and clearly communicating the underlying causal assumptions gets challenging. Lastly, the underlying mathematical formalism is subtle given that it starts from finite populations.

References: Rubin [2005], Imbens and Rubin [2015]

Graphical causal models Graphical casual models (sometimes also called Pearlian graphical models) aim to provide a concise and intuitive description of causal relations between multiple causal variables. They extend probabilistic graphical models, which parameterize distributions via conditional independence constraints, to additionally model interventions that consist of changing specific variables while keeping the others untouched. These models originated in the computer science community, which relies heavily on probabilistic graphical models to efficiently parameterize complex multivariate distributions.

Advantages: Due to its graphical representation it is straight-forward to communicate a causal ordering and describe complex causal relations. Moreover, a causal graph provides a simple language to communicate causal assumptions, which makes it easy to discuss assumptions with domain experts that may not have a background in causality.

Disadvantages: A common concern with these models is that they model all causal mechanisms at once, even if these are not needed for the analysis. In particular, they (in general) assume that interventions on all variables are feasible, which in practice is often unreasonable. A further drawback is that functional constraints cannot be directly included (if this is required a structural causal model is a better choice).

References: Pearl [2009]

Simultaneous and structural equation models Simultaneous and structural equation models extend conventional statistical regression models to allow the endogenous (dependent) variables not only to depend on exogenous (independent) variables but also on the other endogenous variables. By explicitly distinguishing exogenous from endogenous variables these models can be used to formalize causal effects on the endogenous variables given interventions on the exogenous variables. These models originated in the econometrics literature as a way to add causal meaning to specific parts of a regression model and have in particular been used in the context of instrumental variables.

Advantages: The closeness to conventional (non-causal) regression models means that they integrate particularly well with existing statistical regression techniques. Furthermore, similar to the potential outcome models, they only model a specific predefined set of interventions and hence avoid unnecessary assumptions.

Disadvantages: The causal assumptions underlying these models are not made explicit meaning that they can be easily misinterpreted or wrongly applied (if the required assumptions are not satisfied).

References: Newey et al. [1999], Duncan [2014]

Structural causal models Structural causal models (sometimes also called functional causal models) are a refinement of graphical causal models that additionally specify the functional form of all causal mechanisms. One can see them as a hybrid of a graphical causal model and a simultaneous equation model. While in most cases they are defined by specifying a causal structure among all variables, they can in fact also be used (similar to simultaneous equation models) to only model a single causal mechanism and leaving the remaining parts unspecified.

Advantages: They always induce a corresponding (partial) causal graphical model, hence making the implied causal assumptions and structure easy to communicate and discuss. Moreover, by explicitly modeling the causal mechanism, it is easy to specify functional causal assumptions that can be easily mapped to statistical assumptions and procedures.

Disadvantages: A common concern with structural causal models is that by specifying the causal mechanism explicitly they can be easily misinterpreted by using the model to reason about intervention or counterfactual statements for which the model was not intended. Such ambiguities about the causal implication of the model can be avoided by always explicitly stating which interventions the model is intended to model.

References: [Pearl, 2009, Bongers et al., 2021, Peters et al., 2017]

Further models In practice, one often uses a mix or slight variations of the above models. Moreover, additional causal models also exist. Notably, single-world intervention graphical models (SWIGs) [Richardson and Robins, 2013] that attempt to add a rigorous graphical language to potential outcome models and the decision theoretic framework for causality [Dawid, 2012, 2021], which advocates for being more explicit about which causal implications of the structural and graphical causal frameworks are actually used. In the end, which causal model is best suited for a given application, boils down to personal preference and the application at hand. However, no matter what model you use, you should always be able to understand and communicate the causal assumption that it implies.

2 Potential Outcome Models

Assume we are given a set of observed variables and are interested in understanding how a subset of these variables (called *target variables*) is causally affected by intervening on a second subset of these variables (called *intervention variables*). The set of target variables should consist of all variables for which we intend to model its distribution, this can in particular mean that a variable is both a target and an intervention variable (e.g., in the instrumental variable setting). A potential outcome model provides a mathematical language to describe these causal relations. To construct it, we introduce for each of the observed variables either (i) a single new random

variable if it is not a target variable or (ii) a (potentially infinite) set of random variables – called *potential outcomes* – indexed by the values the intervention variables (except itself) can attain if it is a target variable. All of the new variables are assumed to live on the same probability space.

To make this more concrete, let $(T_1, X_1, Y_1), \dots, (T_n, X_n, Y_n) \in \mathcal{T} \times \mathcal{X} \times \mathcal{Y}$ denote random variables corresponding to observations from n different units (e.g., participants in a study). We call $\mathbf{Y} = (Y_1, \dots, Y_n)$ responses, $\mathbf{X} = (X_1, \dots, X_n)$ covariates and $\mathbf{T} = (T_1, \dots, T_n)$ treatments. We now are interested in understanding how the responses are affected by the treatments, hence the responses are target variables and the treatments intervention variables. To construct the potential outcome model, we therefore introduce the following new random variables

$$(\bar{T}_i, \bar{X}_i, (\bar{Y}_i(\mathbf{t}))_{\mathbf{t} \in \mathcal{T}^n})_{i \in [n]}, \quad \text{with joint distribution } P_{\text{full}}. \tag{1}$$

The variables $(\bar{T}_i, \bar{X}_i)$ should be thought of as copies of the observed treatment and covariates, while the potential outcomes $(\bar{Y}_i(\mathbf{t}))_{\mathbf{t} \in \mathcal{T}^n}$ are the random values of the responses under the intervention $\mathbf{T} = \mathbf{t}$. We call the set of probability distributions induced by (1) a *potential outcome model*. This is indeed a causal model according to the abstract definition provided in Section 1, since the distribution P_{full} induces for all $\mathbf{t} \in \mathcal{T}^n$ an interventional distribution $P_{\text{full}}^{\mathbf{Y}(\mathbf{t})}$ over the responses. Similar to a statistical model, we can restrict a potential outcome model by adding additional assumptions on the possible distributions induced by (1). In most applications of potential outcome models one adds the following three assumptions which we discuss in Section 2.1: Assumption 1 (consistency), Assumption 2 (no interference) and Assumption 3 (single unit). When working with a potential outcome model it is important to be aware that the potential outcomes for a single unit can never be observed at the same time. This counterfactual nature is relevant both when constructing targets of inference as well as when making assumptions on the model (see Remark 1).

Remark 1 (Counterfactuals)**.** *Potential outcomes are sometimes referred to as counterfactuals. We avoid this terminology, because potential outcomes are not necessarily counterfactual quantities. In fact one of the potential outcomes, the one for which the treatment is observed, is always factual. More specifically, assume the treatment* $\mathbf{T} = \mathbf{t}$ *is observed, then the potential outcomes* $Y_i(\mathbf{t})$ *are factual while for all* $\mathbf{t}' \in \mathcal{T}$ *with* $\mathbf{t}' \neq \mathbf{t}$ *the potential outcomes* $Y_i(\mathbf{t}')$ *are counterfactual. When working with potential outcomes one needs to be particularly careful at two stages of the analysis: (1) When constructing a target of inference (i.e., a causal estimand) based on potential outcomes and (2) when making assumptions based on potential outcomes. In both cases, it is easy to (accidentally) end up with either a counterfactual quantity or a counterfactual assumption (sometimes called cross-world assumption). If not explicitly of interest, one should avoid counterfactuals as they are both philosophically delicate and potentially non-verifiable via experimentation.*

The potential outcome model in (1) only models how the responses are affected by interventions on the treatment. If we are interested in other causal questions, we need to specify different sets of intervention and target variables. For example, if we are interested in treatment effects mediated by the covariates X we would additionally need to consider X as a target and intervention variable, leading to the potential outcomes $T_i(\mathbf{x})$, $X_i(\mathbf{t})$ and $Y_i(\mathbf{t}, \mathbf{x})$.

2.1 Consistency, no interference and single unit assumptions

We start from the potential outcome model given in (1) and discuss the core assumptions made in most (but not all) applications of the potential outcome model. The first assumption known as *consistency* connects (1) to the observed data.

Assumption 1 (Consistency)**.** *For all* $i \in [n]$ *it holds that*

$$T_i = \bar{T}_i, \quad X_i = \bar{X}_i \quad \textit{and} \quad Y_i = \bar{Y}_i(\bar{\mathbf{T}}).$$

Without consistency the potential outcome model is meaningless as it has no connection to the observed data. Whether or not it holds needs to be argued using knowledge about the data

generating process. Since consistency directly connects the random variables in (1) to the data, we follow the standard convention and drop the bar from the notation. It is nevertheless useful to think of the observed random variables as being separate from the potential outcome model random variables, even if this is not visible in the notation.

Unlike in structural causal models potential outcome models do not explicitly specify the causal ordering. Nevertheless, depending on precise model, several assumptions on the causal order are implicitly encoded. For example, in the model (1) the treatment is assumed to precede the response implying that there is no feedback (or cycle) between the treatment and response. Moreover, if X was also a target variable but we did not add potential outcomes for X, we would be implicitly assuming that X is not causally affected by the treatment. Such mistakes can be avoided by clearly specifying the target and intervention variables and adding additional assumptions on the causal ordering only later.

Since we defined different potential outcomes for all values of the treatment of all units simultaneously, we can model settings in which the potential outcomes of a single unit can depend on the treatments of the remaining units. This can for example occur in clinical trials on the efficacy of vaccines where the potential outcome (getting a disease or not) of one individual might be affected by whether individuals close to that individual received the treatment. Considering these types of dependencies can be difficult, therefore if possible, one often assumes it does not occur.

Assumption 2 (No interference)**.** *For all $i \in [n]$ and all $\mathbf{t}, \mathbf{t}' \in \mathcal{T}^n$ with $t_i = t'_i$ it holds that*

$$Y_i(\mathbf{t}) = Y_i(\mathbf{t}').$$

Assumption 1 and Assumption 2 together are called the Stable Unit-Treatment Value Assumption (SUTVA). As long as no-interference holds, one can avoid defining the potential outcomes for all $\mathbf{t} \in \mathcal{T}$. Therefore, whenever we assume SUTVA we use the simplified potential outcome model

$$(T_i, X_i, (Y_i(t))_{t\in\mathcal{T}})_{i\in[n]} \quad \text{with joint distribution } P_{\text{sutva}}. \tag{2}$$

Finally, while in certain settings it is desirable to assume that different units are affected differently the unit-level effects are often not identifiable without strong assumptions. It is therefore common to assume that the unit-level data are i.i.d..

Assumption 3 (Single unit)**.** *Define for all $i \in [n]$ the variables*

$$W_i \coloneqq (T_i, X_i, (Y_i(\mathbf{t}))_{\mathbf{t}\in\mathcal{T}^n})$$

and assume that $W_1, \dots, W_n$ are independent and identically distributed.

The single unit assumption makes most sense in combination with the no-interference assumption, as it otherwise strongly restricts the type of allowed interference. It can be justified by arguing that if the sequence $W_1, \dots, W_n$ is exchangeable, then, in the limit as n tends to infinity,[1] de Finitti's representation theorem allows us to construct an identically distributed i.i.d. sequence. Under the single unit assumption, the potential outcome model can be fully specified by the distribution of a single unit. Therefore, whenever we assume SUTVA and single unit we use the simplified potential outcome model

$$(T, X, (Y(t))_{t\in\mathcal{T}}) \quad \text{with joint distribution } P_{\text{sutva-iid}}. \tag{3}$$

2.2 Average causal effects

We now assume we are given a potential outcome model

$$(T, X, (Y(t))_{t\in\mathcal{T}}) \quad \text{with joint distribution } P. \tag{4}$$

[1]The case of an infinite population is sometimes referred to as superpopulation in the potential outcome literature.

that satisfies Assumption 1, Assumption 2 and Assumption 3. The most common causal estimand in the causal literature is the average causal effect. It is defined differently depending on whether the treatment variable T is binary or continuous. For binary treatments ($\mathcal{T} = \{0, 1\}$) it is given by

$$\text{ACE} \coloneqq \mathbb{E}[Y(1)] - \mathbb{E}[Y(0))],$$

while for continuous treatments it is defined by

$$\text{ACE} \coloneqq \mathbb{E}[\tfrac{d}{dt}\mathbb{E}[Y(t)]|_{t=T}],$$

where we assume sufficient smoothness for the derivative to exist. If a single number summarizing the causal effect is insufficient, one can also consider the causal dose-response curve, given by

$$t \mapsto \mathbb{E}[Y(t)].$$

In the following section, we discuss the most common assumptions used in the potential outcome model to ensure that these types of causal effects are identifiable (see Definition 1).

2.3 Identifiability conditions

To achieve identifiability of average causal effects in the potential outcome model (4), we can consider hypothetical experiments in which the treatment is randomly assigned. In such a case, it makes sense to assume that the treatment assignment is independent of the potential outcomes.

Assumption 4 (Ignorability)**.** *It holds for all $t \in \mathcal{T}$ that*

$$Y(t) \perp\!\!\!\perp T.$$

Using ignorability we can directly express the average causal effect (in the binary case) as

$$\begin{aligned}\text{ACE} &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}[Y(T) \mid T = 1] - \mathbb{E}[Y(T) \mid T = 0] \\ &= \frac{\mathbb{E}[Y\mathbb{1}(T = 1)]}{\mathbb{P}(T = 1)} - \frac{\mathbb{E}[Y\mathbb{1}(T = 0)]}{\mathbb{P}(T = 0)},\end{aligned}$$

where in the second equality we used ignorability and in the third equality we used consistency and the definition of the conditional expectation. Here we additionally assumed that $\mathbb{P}(T = 0) > 0$ and $\mathbb{P}(T = 1) > 0$, which will be formalized below in Assumption 6. In settings in which we cannot assume that the treatment is randomized it is no longer realistic to assume ignorability. Instead, one therefore often considers a weaker notion, that is motivated by a randomized control trial in which the treatment is assigned randomly based on a set of covariates X.

Assumption 5 (Conditional Ignorability)**.** *It holds for all $t \in \mathcal{T}$ that*

$$Y(t) \perp\!\!\!\perp T \mid X.$$

Conditional ignorability alone is not sufficient to ensure identifiability of the causal effect, since it does not ensure that all possible treatments $t \in \mathcal{T}$ are observed across all subgroups $X = x$. We therefore make the following additional assumption.

Assumption 6 (Positivity)**.** *Assume* (4) *induces a density p. Then, it holds for all $t \in \mathcal{T}$ that*

$$\mathbb{P}(p(t \mid X) > 0) = 1.$$

Given that both conditional ignorability and positivity are true, we can express the average causal effect in terms of the distribution of the observed variables (T, X, Y). In the case of a binary

treatment, we get

$$\begin{aligned}
\text{ACE} &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&= \mathbb{E}[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]] \\
&= \mathbb{E}[\mathbb{E}[Y(T) \mid X, T=1] - \mathbb{E}[Y(T) \mid X, T=0]] \\
&= \mathbb{E}[\mathbb{E}[Y \mid X, T=1] - \mathbb{E}[Y \mid X, T=0]] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[Y\mathbb{1}(T=1) \mid X]}{\mathbb{P}(T=1 \mid X)} - \frac{\mathbb{E}[Y\mathbb{1}(T=0) \mid X]}{\mathbb{P}(T=0 \mid X)}\right], \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{Y\mathbb{1}(T=1)}{\mathbb{P}(T=1 \mid X)} \mid X\right] - \mathbb{E}\left[\frac{Y\mathbb{1}(T=0)}{\mathbb{P}(T=0 \mid X)} \mid X\right]\right], \\
&= \mathbb{E}\left[\frac{Y\mathbb{1}(T=1)}{\mathbb{P}(T=1 \mid X)} - \frac{Y\mathbb{1}(T=0)}{\mathbb{P}(T=0 \mid X)}\right],
\end{aligned}$$

where in the third equality we used conditional ignorability, in the fourth equality we used consistency and fifth equality we used the properties of the conditional expectation. More generally, for potentially non-binary T, we get for $t \in \mathcal{T}$ that

$$\begin{aligned}
\mathbb{E}[Y(t)] &= \mathbb{E}[\mathbb{E}[Y \mid X, T=t]] \\
&= \int\int y p(y \mid x, t) \mu(dy) p(x) \mu(dx) \\
&= \int\int y \frac{p(y, x \mid t) p(x)}{p(x \mid t)} \mu(dy) \mu(dx) \\
&= \int\int y \frac{p(t)}{p(t \mid x)} p(y, x \mid t) \mu(dy) \mu(dx) \\
&= \mathbb{E}\left[Y \frac{p(t)}{p(t \mid X)} \,\middle|\, T=t\right].
\end{aligned}$$

This short computation provides two useful representations for identifying the interventional distribution $\mathbb{E}[Y(t)]$. Firstly, the adjustment representation $\mathbb{E}[\mathbb{E}[Y \mid X, T=t]]$, which if used directly requires an estimate of the conditional expectation function $(x,t) \mapsto \mathbb{E}[Y \mid X=x, T=t]$. Secondly, the propensity weighted representation $\mathbb{E}[Y \frac{p(t)}{p(t|X)} \mid T=t]$, which if the propensity score $p(t \mid x)$ is known (or easy to estimate) only requires an estimate of the simpler conditional expectation function $t \mapsto \mathbb{E}[\tilde{Y} \mid T=t]$ for $\tilde{Y} = Y \frac{p(t)}{p(t|X)}$.

Remark 2 (Identifiability for continuous treatments)**.** *We skipped over a technical detail regarding identifiability in the case of a continuous treatment variable T (i.e., dominated by the Lebesgue measure). Even if there exists a joint density p for the observational distribution over (T, X, Y), the positivity assumption is insufficient to uniquely identify the density as it can be modified on sets of measure zero. This further implies that the conditional expectation function*

$$(t, x) \mapsto \mathbb{E}[Y \mid T=t, X=x]$$

is also not uniquely identified. To avoid this unidentifiability additional regularity conditions on the density (e.g., assuming it is continuous) are required. We do not consider these issues further and simply assume that sufficient regularity for identifiability of the density is given.

2.3.1 Adjusting with propensity scores

The following result is due to Rosenbaum and Rubin [1983].

Theorem 1 (Adjusting with propensity scores)**.** *Assume $T \in \{0,1\}$ and the potential outcome model satisfies Assumptions 1, 2, 3 and 5. Then, for all $t \in \{0,1\}$ it holds that*

$$Y(t) \perp\!\!\!\perp T \mid \pi(X),$$

where $\pi(X) \coloneqq \mathbb{E}[T \mid X]$ is the propensity score.

Proof. We begin by showing that

$$X \perp\!\!\!\perp T \mid \pi(X). \tag{5}$$

This independence is also called the *balance equation* and only requires that T is binary. To prove (5), it is enough to show that

$$\mathbb{E}[T \mid \pi(X)] = \mathbb{E}[T \mid X, \pi(X)]$$

since T is binary. First, expanding the left hand side we get

$$\mathbb{E}[T \mid \pi(X)] = \mathbb{E}[\mathbb{E}[T \mid \pi(X), X] \mid \pi(X)] = \mathbb{E}[\mathbb{E}[T \mid X] \mid \pi(X)] = \pi(X)$$

and similarly expanding the right hand side we get

$$\mathbb{E}[T \mid \pi(X), X] = \mathbb{E}[T \mid X] = \pi(X).$$

This completes the proof of (5).

Next, fix $t \in \{0, 1\}$ and an arbitrary bounded measurable function $g : \mathbb{R} \to \mathbb{R}$. Then, it holds that

$$\begin{aligned}
\mathbb{E}[g(Y(t)) \mid \pi(X), T] &= \mathbb{E}[\mathbb{E}[g(Y(t)) \mid X, \pi(X), T] \mid \pi(X), T] \\
&= \mathbb{E}[\mathbb{E}[g(Y(t)) \mid X, T] \mid \pi(X), T] \\
&= \mathbb{E}[\mathbb{E}[g(Y(t)) \mid X] \mid \pi(X), T] \\
&= \mathbb{E}[\mathbb{E}[g(Y(t)) \mid X] \mid \pi(X)] \\
&= \mathbb{E}[g(Y(t)) \mid \pi(X)]
\end{aligned}$$

Here, for the first and second equality we used the tower property and that we can drop $\pi(X)$ from the conditioning, respectively. For the third equality we used conditional ignorability (i.e., $Y(t) \perp\!\!\!\perp T \mid X$). For the fourth equality we used (5). Finally, for the last equation we used that the tower property again.

Since this holds for all bounded measurable g, it follows by the definition of the conditional expectation [see e.g. Dawid, 1979] that

$$Y(t) \perp\!\!\!\perp T \mid X.$$

This completes the proof of Theorem 1. □

References

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

A. P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979. doi: 10.1111/j.2517-6161.1979.tb01052.x.

P. Dawid. The decision-theoretic approach to causal inference. *Causality: Statistical perspectives and applications*, pages 25–42, 2012.

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

O. D. Duncan. *Introduction to structural equation models*. Elsevier, 2014.

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

W. K. Newey, J. L. Powell, and F. Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999. doi: 10.1111/1468-0262.00037.

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

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

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.

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. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.