# A chill intro to causal inference via propensity scores

*George Berry (geb97@cornell.edu, @george_berry)*

*3/10/2019, version 1.0.1*

## Abstract

Observational causal inference is hard. This introduction aims to present causal inference in an accessible and straightforward way, with terminology familiar to information scientists. Particular attention is devoted to propensity score methods, which are commonly used and also a frequent subject of confusion. Simulations demonstrate the key arguments, and pointers to reference material are provided.

## Introduction

The last couple of years, I've been assigned papers to review at information science conferences using propensity score methods (PSM). These papers usually make an argument that goes: we use PSM, we throw 5-10 features in our model and achieve balance on those covariates, therefore our estimate is causal.

This process isn't likely to estimate a causal effect. It also incorrectly suggests that causal inference can be accomplished in a black-box framework similar to standard machine learning modeling. In this text, I'll provide a friendly introduction to observational causal inference using terminology hopefully familiar to IS (and CS) folks (for example, we'll use the word "features" rather than "variables"). My hope is that better understanding of causal inference will cause it to be more widely used in IS applications.

The big point of this text is **if you would like to use any observational causal inference technique (including PSM), you must engage with assumptions about how causality operates in your data**. It is not enough to have a bunch of data and predict who is likely to be treated. At the same time, **you should consider using causal inference because lots of important questions require it**.

This text aims to provide a quick and intuitive overview using the minimum possible terminology. Unfortunately, there's still a fair amount. I've included pointers to comprehensive reference material, but I don't include proofs.

**Why this text focuses on propensity scores:** PSM appears similar to predictive modeling at first: we are predicting treatment based on features. Propensity scores, initially, appear to allow causal inference using the same tools we use for prediction all the time. But they suffer from all of the normal problems of observational causal inference, and we need to make explicit causal (rather than statistical) assumptions for propensity score-based treatment effects to be valid.

**What you'll find here**

1. An introduction to causal inference using the potential outcomes model
2. A discussion of the assumptions underlying propensity score methods (PSM)
3. Using a framework called **causal graphs**, a demonstration that you need to worry about causal assumptions when using PSM
4. Simulations showing how to estimate causal effects in various idealized scenarios
5. A baseline set of things you should consider and discuss when using PSM or other observational causal inference methods
6. Other observational causal inference methods beyond PSM you might want to consider

**Further reading**

This is a brief introduction. I'd recommend the following books.

- **Counterfactuals and causal inference by Morgan and Winship**: Treatment of causal inference for social scientists using both causal graphs and potential outcomes, with examples largely from sociology, which means the authors carefully think through causal inference in the messy domains sociologists study.
- **Causal inference: the mixtape by Cunninham**: A contemporary approach with lots of examples and problems to work through (Stata is the language used). Available for free as a PDF. Chance the Rapper quotes included.
- **Causality by Pearl**: Comprehensive treatment of causal graphs: this is a great theory book that serves as an important reference when tough questions about causality come up.
- **Mostly harmless econometrics: an empiricist's companion by Angrist and Pischke**: Causal inference in a regression-focused framework with examples from economics. Lots of emphasis on quasi-experimental techniques. You will really learn to apply the law of iterated expectations.

## Example

To provide a concrete case throughout, I'll refer to the following **example**: let's say we are studying sociologists on Twitter (#academictwitter!). We want to understand the causal effect of professors self-disclosing their academic position on Twitter ($D$) on having a lot of followers ($Y$). We have a hypothesis that says when a professor (indexed by $i$) discloses their position ($D_i = 1$), they are more likely to have a lot of followers ($Y_i = 1$). We also measure whether someone is an avid social media user before disclosure ($X_i$), and we believe that avid social media usage causes both disclosure ($X \rightarrow D$) and having lots of followers ($X \rightarrow Y$).

We can represent these assumptions in the **causal graph** Figure 1 (see (Pearl 2009) for a full treatment and (Morgan and Winship 2015) for an accessible introduction). An arrow from one node to the other encodes our assumptions about the causal process.

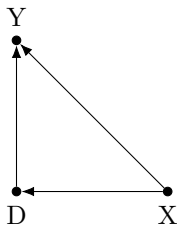



Figure 1: A causal graph where arrows represent causal relationships. $Y$ is the outcome we care about (being highly followed), $D$ is the treatment we are studying (disclosing status) $X$ is a feature that causes both treatment and outcome (avid social media usage before disclosure). Arrows represent our assumptions about the causal process.

## PSM notation and the fundamental problem of causal inference

PSM is pretty straightforward, with one subtle point. First, some notation about the counterfactual model of causal inference. This is often called the Neyman-Rubin model (Sekhon 2009).

- $Y$ is the **outcome** we care about (being highly followed).
- $D$ is the **treatment status** (disclosing position), with $D = 1$ meaning treated (disclosed) and $D = 0$ meaning control (didn't disclose).
- $X$ are **features**, also called covariates, conditioning variables, or observables. $X$ are assumed to cause the treatment. We'll consider a single feature, being highly followed before status disclosure.
- $i$ indexes **units**, for instance people observed in our dataset. When we refer to an individual's treatment or outcome, we write $D_i$ instead of $D$ and $Y_i$ instead of $Y$.
- $Y^1, Y^0$ are **potential outcomes**, representing the outcome under treatment and control, respectively. The subtle point is that we only observe one of these potential outcomes for each unit $i$. We can write $Y_i$ in terms of potential outcomes and treatment status like this: $Y_i = D_i * Y_i^1 + (1 - D_i) * Y_i^0$.
- $p(D = 1|X)$ is the **propensity score**, or probability of being treated at some point in feature space.
- $\delta$ represents the **average treatment effect (ATE)**. This can be thought of as the average difference between an individual's potential outcomes under treatment and control, written $E[Y_i^1 - Y_i^0]$. We will call the true treatment effect $\delta_{ATE}$, and denote various estimators of the ATE by $\hat{\delta}_{ATE}$.

The **fundamental problem of causal inference** is that we can't observe both of the potential outcomes $Y_i^1, Y_i^0$ for any given unit $i$. The problem from this perspective is a missing data problem. The counterfactual outcome ($Y^0$ for the treated and $Y^1$ for the control) is unobserved for all units, and we need to make a between-units comparison to estimate the ATE. You can see this clearly with this hypothetical data.

| $i$ | $D_i$ | $Y_i^1$ | $Y_i^0$ | $\delta_i = Y_i^1 - Y_i^0$ |
|---|---|---|---|---|
| Alice | 1 | 1 | - | - |
| Bob | 0 | - | 0 | - |
| Carol | 0 | - | 1 | - |
| David | 1 | 1 | - | - |

Table 1: A demonstration that we can only observe one of the two potential outcomes for each unit. For instance, Alice's control outcome, $Y_i^0$ is unobserved because she is in the treatment condition. This means that all individual level treatment effects $\delta_i$ are unobserved. The observed outcome is written $Y_i = D_i * Y_i^1 + (1 - D_i) * Y_i^0$.

The straightforward way (we'll call it "randomized", because it usually only works when you have randomized treatment assignment) to estimate average treatment effects (ATE) is

$$\hat{\delta}_{\text{ATE}}^{\text{randomized}} = E[Y_i^1 - Y_i^0] = E[Y_i^1|D = 1] - E[Y_i^0|D = 0].$$

This estimator takes the average outcomes for the treatment group and control group and differences them. It works when there is random assignment because the equality $E[Y_i^1 - Y_i^0] = E[Y_i^1|D = 1] - E[Y_i^0|D = 0]$ holds. To see this, note that under random assignment the average potential outcomes in the treatment and control groups are the same: $E[Y_i^1|D = 1] = E[Y_i^1|D = 0]$ and $E[Y_i^0|D = 1] = E[Y_i^0|D = 0]$.

If the treatment in Table 1 isn't randomly assigned, we say it suffers **selection bias**. When there is selection bias, factors including individual choices and the operation of social systems influence the treatment status $D$ in non-random ways. In this case, the ATE estimated with $\hat{\delta}_{\text{ATE}}^{\text{randomized}}$ is likely biased because the equality $E[Y_i^1 - Y_i^0] = E[Y_i^1|D = 1] - E[Y_i^0|D = 0]$ does not hold.

We need two types of additional information to estimate the ATE in a setting with selection bias.

1. Features $X$ which we believe cause the treatment assignment $D$. You might call this **technical information**, since it results from a process of data collection and data cleaning.
2. Knowledge about the causal process, expressed as assumptions, which restrict how we allow causation to operate in our model. You might call this **causal knowledge**.

For instance, maybe David has a lot of followers ($Y_i = 1$) because he was an avid social media user before going to graduate school ($X_i = 1$), and his decision to disclose his professional status ($D_i = 1$) came because of his avid social media usage ($X \to D$). We need to be able to both measure the feature $X$ and encode the assumption that $X$ causes $D$ in a model. Without both of these things, we can easily mistake David's pre-existing popularity for the causal effect of disclosing his job title.

To make this really clear, let's assume for a moment that we could avoid the fundamental problem of causal inference, and observe both potential outcomes for each unit. This is displayed in Table 2, with the data that's hidden from us in real-world applications in orange.

If we had access to the information in Table 2, we could correctly estimate the ATE. We could average within-individual differences with an "omniscient" ATE estimator:

| $i$ | $D_i$ | $Y_i^1$ | $Y_i^0$ | $\delta_i = Y_i^1 - Y_i^0$ |
|---|---|---|---|---|
| Alice | 1 | 1 | 0 | 1 |
| Bob | 0 | 0 | 0 | 0 |
| Carol | 0 | 1 | 1 | 0 |
| David | 1 | 1 | 1 | 0 |

Table 2: An example of the case where we could hypothetically observe both potential outcomes for each unit. The data hidden in real-world applications is shown in orange. Having this data allows us to observe individual-level treatment effects (last column, shown in blue).

$$\delta_{\text{ATE}} = \hat{\delta}_{\text{ATE}}^{\text{omniscient}} = E[Y_i^1 - Y_i^0] = \frac{1}{4}((1-0) + (0-0) + (1-1) + (1-1)) = 0.25$$

We see that the causal effect of disclosing status $D$ on being highly followed $Y$ is $\delta_{\text{ATE}} = 0.25$. The treatment only has an effect for Alice. Even with the treatment, Bob would not become highly followed. Carol was highly followed without the treatment, and David would have been highly followed without the treatment.

Now let's restrict ourselves to estimating the ATE with only the data from Table 1, representing the conditions we face in real life.

$$\delta_{\text{ATE}} \neq \hat{\delta}_{\text{ATE}}^{\text{randomized}} = E[Y_i^1 | D = 1] - E[Y_i^0 | D = 0] = \frac{1+1}{2} - \frac{0+1}{2} = 1 - 0.5 = 0.5.$$

The treatment effect is over-estimated. If we used this estimate we would conclude that $D$ had a larger effect on $Y$ than it really does. An intervention designed around getting professors to disclose status $D$ to get them lots of followers $Y$ would show disappointing results.

Given that we never have access to the data in Table 2, and we'd still like to estimate the effect of $D$ on $Y$, what can we do? Use the information contained in features $X$, plus causal assumptions, to estimate a credible ATE.

## Assumptions for using $X$ to model treatment assignment

Now that we understand the basic problem of causal inference, let's discuss the assumptions you'll read about when consulting reference material on PSM. We'll see that the most important assumption for PSM cannot be tested. In other words, **you cannot make a purely statistical or technical argument for PSM, you must appeal to a substantive process causing selection into treatment**. All non-experimental causal inference faces similar problems.

PSM says that if we have relevant information on the process $X$, we can estimate an unbiased ATE using only the propensity score, $p(D = 1|X)$ and the observed outcomes $Y$. Let's get specific about the assumptions needed for $X$ in the form of a propensity score to allow unbiased estimation of the ATE (these are adapted from the presentation in (Cunningham 2018)).

1. **Conditional independence assumption (CIA)**: This assumption says that $(Y^1, Y^0) \perp\!\!\!\perp D|X$. In words, the potential outcomes are independent of the treatment, once we know the information in $X$. The information in $X$ can be contained in a single propensity score $p(D = 1|X)$ if we like, in which case the CIA is written $(Y^1, Y^0) \perp\!\!\!\perp D|p(D = 1|X)$ (to understand why the two CIA statements are equivalent, see (Rosenbaum and Rubin 1983)). It's important to stop right here and recognize that we are stating something that is not testable because we do not observe both potential outcomes for any observation, as expressed in Table 1. We always have exactly one potential outcome for each unit, either $Y_i^1$ or $Y_i^0$, but never both. You may encounter the CIA under the names like **ignorability** or **unconfoundedness**, although I find these terms vague.
2. **Common support (CS)**: for each strata of features $X = x$, there must be positive probability of being both in treatment and control. For instance, we could not conduct a study on the effects of social media disclosure if all professors declared their status. There would be no observations in the control group. This assumption can be tested, and is usually addressed by limiting the analysis to $X$'s which appear in both treatment and control.

An additional factor often discussed with PSM is called **covariate balance** (Imai and Ratkovic 2014), which means that the treatment and control groups, after propensity score weighting, are about the same on the observables $X$. This method has been shown to improve estimates of $\delta_{\text{ATE}}^{\text{PSM}}$. It's important to note that **covariate balance does not mean your ATE estimates are valid**. We'll see this in a second with simulations, where our biased and unbiased estimates both have balanced covariates.

## CIA (subtly) implies it is not enough to predict the treatment

It's not clear from the terse statement of the CIA, $(Y^1, Y^0) \perp\!\!\!\perp D|X$, why we need to engage in reasoning about the causal process generating selection into treatment. But it's true: saying "we have a lot of features, surely the ATE must be relatively unbiased" is not a valid argument. In fact, including additional features in your model can increase bias.

The reasoning is easy to see using a visual framework for causal inference called **causal graphs** (Pearl 2009; Morgan and Winship 2015), which I previewed earlier.

## Causal graphs and backdoor paths

We want to understand the effect $D \to Y$. To do this, Pearl (2009) lays out the **backdoor criterion (BDC)** for identifying the effect of one feature on another. If we can satisfy the BDC (we'll discuss how in a bit), we can identify the effect of $D$ on $Y$. **Satisfying the conditional independence assumption for a propensity score model is equivalent to satisfying the backdoor criterion** (see discussion in (Pearl 2009; Cunningham 2018)).

The BDC has a big advantage: you can often check visually whether the assumed causal model represented by a causal graph satisfies the BDC. I like to do this in a quiet coffee shop, with pen and paper in my unlined notebook from Norway. If it's hard, there's even software to do it for you automatically (check out the `dagitty` R package). On the other hand, I have no idea how to directly check that assumptions satisfy the CIA.

Let's take another look at the causal graph we previewed earlier, where $X$ causes the treatment $D$ and outcome $Y$, and the treatment $D$ causes $Y$.

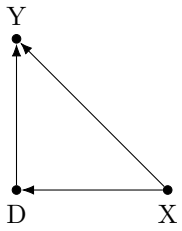



Figure 2: A causal graph representing our assumptions about how two features, the treatment $D$ and a feature $X$, cause an outcome $Y$.

For the BDC, we start with the feature whose causal effect we care about, $D$. Then we trace paths in the graph backwards from $D$ to the outcome we care about, $Y$. We do this by starting with each feature pointing to $D$ and tracing each path backwards to $Y$. It's okay if the paths overlap in some places.

In the graph in Figure 2, the only such backdoor path is $D \leftarrow X \rightarrow Y$. The BDC states that we identify the causal effect of $D$ on $Y$ if both

1. We "block" all such backdoor paths with appropriate control features
2. No feature we use as a control feature is a descendant of the treatment $D$

Blocking a backdoor path means:

1. We control for the middle feature in a path, $B$ in $A \rightarrow B \rightarrow C$
2. We control for the middle feature in a source, $B$ in $A \leftarrow B \rightarrow C$
3. We **don't** control for the middle feature in a sink, $B$ in $A \rightarrow B \leftarrow C$, or any of $B$'s descendants. The middle feature in a sink is referred to as a **collider**.

In Figure 2, the only feature that lies along a backdoor path that we need to control for is $X$. So if we estimate a model for the effect of $D \rightarrow Y$ and include $X$ as a control, we will identify the effect of $D$ on $Y$ (assuming we don't mess up the functional form too bad, but let's not worry about that here).

## Ideal world

Let's say that the causal assumptions represented in Figure 2 are correct. And let's say we can observe $D, X, Y$. Then we can estimate the effect of $D \rightarrow Y$ by controlling for $X$. **Controlling for** a feature just means including it in a multivariate regression, or in your propensity score model.

You can see two ways of "controlling for $X$" in the following simulation.

1. Estimating a regression $Y = \beta_0 + \beta_1 D + \beta_2 X + \epsilon$.
2. Estimating a propensity score model $P(D = 1|X)$ using logistic regression.

Note that you can also do **propensity score matching** using a package such as `MatchIt` (Ho et al. 2018). But you probably shouldn't use this method, and should instead rely on inverse probability weighting for estimating treatment effects (King and Nielsen 2015).

In these simulations, the true effect of $D$ on $Y$ is 1/2.

Here's some setup code.

```
library(tidyverse)
set.seed(1234)
ATE_propensity = function(Y, D, propensity) {
    return(
        mean(Y * D / propensity) - mean(Y * (1 - D) / (1 - propensity))
    )
}

covariate_balance = function(X, D, propensity) {
    return(
        mean(X * D / propensity) - mean(X * (1 - D) / (1 - propensity))
    )
}

N = 10000
Ux = rnorm(N)
Uy = rnorm(N)

X = Ux
D_prob = 1/(1 + exp(-1/2 * X))
D = rbinom(N, 1, D_prob)
Y = 1/2 * D + 1/2 * X + Uy

df = data.frame(X=X, Y=Y, D=D)
```

Now we estimate the ATE with propensity scores. The true effect is 1/2, so this looks good as expected.

```
propensity = predict(glm(D~X, data=df, family='binomial'), type='response')
print(ATE_propensity(Y, D, propensity))
```

```
## [1] 0.4979767
```

You can see that balance on $X$ is good when weighting by propensity. This measure of balance is taken from (Imai and Ratkovic 2014).

```
print(covariate_balance(X, D, propensity))
```

```
## [1] 0.002944151
```

We get basically the same result with a multivariate regression.

```
print(coef(lm(Y~D+X, data=df)))
```

```
##  (Intercept)            D            X
```

```
## -0.007015131  0.498395401  0.511144921
```

## Omitted variable bias (omitted feature bias)

Omitted variable bias (OVB) is the most widely discussed reason an estimated ATE could be biased. In this case, we forget to include a feature we should have included. Usually, this is because we can't measure the thing we need to measure.

This is easy to see in Figure 3, where the necessary control feature $X$ is unobserved (indicated by a hollow point in the causal graph).

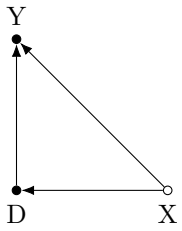



Figure 3: Here, the feature $X$ is unobservable, leading to omitted variables bias in the estimate of $D$ on $Y$ since there is no way to control for $X$.

This messes up the ATE estimate because a backdoor path is unblocked. In this case, there is nothing we can do to estimate a valid ATE. Note that there's no propensity score to estimate here because there are no features $X$.

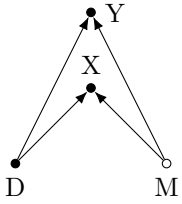
```
print(coef(lm(Y~D, data=df)))
```

```
## (Intercept)           D
##  -0.1223037   0.7358898
```

## Collider bias

Collider bias (CB) is the opposite of OVB: it's when we include something in the model we should not have included. The included feature then biases the ATE.

CB is named after **colliders**, which are features along a backdoor path which are caused by two or more other features, such as $B$ in $A \rightarrow B \leftarrow C$. The following causal graph illustrates this, where $X$ is now a collider caused by $D$ and $M$.



It's easy to remember collider bias because, in this example, it kinda looks like a Star Trek communicator.

```
Ux = rnorm(N)
Uy = rnorm(N)
Ud = rnorm(N)
Um = rnorm(N)

D_prob = 1/(1 + exp(-Ud))
D = rbinom(N, 1, D_prob)
M = Um
X = 1/2 * D + 1/2 * M + Ux
Y = 1/2 * D + 1/2 * M + Uy

df = data.frame(X=X, Y=Y, D=D, M=M)
```

Here, the right strategy to estimate the effect of $D$ on $Y$ is to simply use $D$ with no controls. There's no unblocked backdoor path from $D$ to $Y$ because a collider $X$ lies along the only backdoor path, $D \rightarrow X \leftarrow M \rightarrow Y$. Colliders block the path for you, and nothing else needs to be done. Including the collider in the model unblocks the path. We can see this in the simulation below.

```
print(coef(lm(Y~D, data=df)))
```

```
## (Intercept)           D
## -0.01033564  0.52997529
```

What happens when we start controlling for $X$? Things go wrong. First, the propensity score ATE.

```
propensity = predict(glm(D~X, data=df, family='binomial'), type='response')
print(ATE_propensity(Y, D, propensity))
```

```
## [1] 0.4354768
```

Note that balance in $X$ is still good after weighting by propensity score.

```
print(covariate_balance(X, D, propensity))
```

```
## [1] -0.001066885
```

Then the regression version. Same results: biased treatment effect.

```
print(coef(lm(Y~D+X, data=df)))
```

```
## (Intercept)           D           X
## -0.01513107  0.43586583  0.19976483
```

From the standard prediction point of view, collider bias is particularly problematic. In a prediction context we usually throw a lot of stuff into the model. However, causal inference is a different beast, and using a "kitchen sink" model can prevent us from estimating the effect of interest.

Taken together, collider bias and omitted variable bias mean that discussing specific assumptions about the causal process is necessary for estimating an ATE using PSM.

### Why does collider bias happen?

Collider bias appears mysterious at first. Here's an intuitive explanation: once we know the value of a collider $X$, it gives us information about the relationship between the causes of $X$, $D$ and $M$. For instance, if $X$ is large, it's likely that either $D$ or $M$ is large, but not both.

Here's a presentation with a simple example. We'll condition on the collider $X$, and you'll be able to see that this conditioning induces correlation between the causes of the collider, $D$ and $M$. This happens even though the overall correlation between $D$ and $M$ is zero.

```
Ux = rnorm(N)
Uy = rnorm(N)
Ud = rnorm(N)
Um = rnorm(N)

D = Ud
M = Um
X = 1/2 * D + 1/2 * M + Ux
Y = 1/2 * D + 1/2 * M + Uy

df = data.frame(X=X, Y=Y, D=D, M=M)
```

```
print(cor.test(df$D, df$M))
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$D and df$M
## t = -1.4008, df = 9998, p-value = 0.1613
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.033598908  0.005593543
## sample estimates:
##         cor
## -0.01400806
```

Now, let's condition on values of the collider $M$. We'll condition on values of $X$ greater than 1, and then greater than 2. In both cases, we see that our intuition proves correct: conditioning on values of a collider induces correlation among its causes.

```
df_collider = df %>% filter(X > 1)
cor.test(df_collider$D, df_collider$M)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_collider$D and df_collider$M
## t = -7.1117, df = 2057, p-value = 1.575e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1967928 -0.1124658
## sample estimates:
##        cor
## -0.1549114
```

```
df_collider = df %>% filter(X > 2)
cor.test(df_collider$D, df_collider$M)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_collider$D and df_collider$M
## t = -2.8525, df = 545, p-value = 0.004502
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.20305620 -0.03783454
## sample estimates:
##        cor
## -0.1212854
```

## Proxy control scheme

Sometimes we can't measure a feature $X$ we need to model selection into treatment $D$. But causal graphs suggest a variety of ways to reduce bias in this case. I've taken this example from (Elwert and Winship 2014). If we can measure some downstream outcome $M$ of the unmeasurable feature $X$, we can reduce bias by conditioning on $X$. There are many such strategies, and I'd encourage you to consult (Elwert and Winship 2014; Pearl 2009) for more.

```
Ux = rnorm(N)
Uy = rnorm(N)
Um = rnorm(N)

X = Ux
D_prob = 1/(1 + exp(-1/2 * X))
D = rbinom(N, 1, D_prob)
Y = 1/2 * D + 1/2 * X + Uy
M = 1/2 * X + Um
```

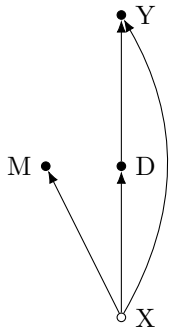



Figure 4: Even when $X$ is unmeasurable, we can reduce bias by conditioning on $M$. Note that the absence of a link $M \to Y$ encodes our assumption that $M$ does not cause $Y$.

```
df = data.frame(X=X, Y=Y, D=D, M=M)
```

```
print(coef(lm(Y~D, data=df)))
```

```
## (Intercept)           D
##  -0.1252937   0.7324187
```

We reduce the bias in $D \to Y$ by including $M$.

```
propensity = predict(glm(D~M, data=df, family='binomial'), type='response')
print(ATE_propensity(Y, D, propensity))
```

```
## [1] 0.6930174
```

Balance looks good.

```
print(covariate_balance(M, D, propensity))
```

```
## [1] 4.838147e-05
```

Then the regression version. Same results.

```
print(coef(lm(Y~D+M, data=df)))
```

```
## (Intercept)           D           M
##  -0.1057538   0.6932698   0.1859442
```

## Prediction, causation, and PSM's appeal

It's quite common in information and computer science applications to address problems as **prediction problems**. Prediction is useful for many tasks, but there are sets of problems where understanding causation is necessary (Kleinberg et al. 2015). When we are concerned with making things better in some real social system, both prediction and causation have important uses. To oversimplify, prediction is important when we have access to an intervention that can be targeted at people having a certain type of experience, and causation is

important when we wish to understand the mechanics of a system to evaluate its operation and potential in-system changes.

The example of rain is given by (Kleinberg et al. 2015). Prediction is useful for knowing when to carry an umbrella, and causation is useful for knowing what factors generate rain. Causal knowledge in this case is not necessary for knowing how to avoid getting wet in a storm (look at the prediction, bring an umbrella), but causal knowledge is necessary to understand something like the effects of climate change on rainfall patterns.

One of the confusing things about the prediction/causation divide is that similar notation and terms are used for both. However, as (Pearl 2009) convincingly argues, when we're doing observational causal inference we need to make non-statistical assumptions about causation. And Pearl provides clear language, which is distinct from statistical language, to describe causal reasoning.

My hunch is that one reason for use of PSM in IS applications is that it feels a lot like our standard prediction framework. We just move the prediction one level back, from predicting $Y$ to predicting $D$. But as shown above, without engaging in reasoning about the process generating $D$, we can't be sure we're estimating the causal effect of interest.

## A list of questions to ask yourself

It's also helpful to think about causal inference with observational data as a spectrum. Estimates can be more or less credible, and pretty much everyone agrees that estimates are open to revision based on new information about relationships between variables, new inference techniques, and new data. We should still do our best to estimate causal effects, but we don't have to project absolute certainty.

To have a shot at estimating a credible causal effect, here's an (incomplete) list of things you should consider and discuss.

1. Is it reasonable to think that $D$ can cause $Y$? If we have a switch to turn $D$ on, is it plausible to think $Y$ would move in response?
2. Do features $X$ which cause the treatment $D$ plausibly happen earlier in time?
3. Forget the data you have and use your intuition. If you had to dream up the things that caused $D$, what would they be?
4. Forget your intuition. What does prior literature say causes $D$?
5. Do you have a variable $Z$ which should not be caused by $D$? Does your model say $D$ causes $Z$ after controlling for $X$?

## Alternate observational causal inference methods to consider

PSM is a type of observational causal inference that's appropriate when you have a convincing model of how the treatment is determined, probably derived from past literature and a mixture of quantitative and qualitative research. It's similar to multiple regression, but offers some nice advantages because you can work with a single score.

Economists in particular have devised a bunch of additional observational causal inference and **quasi-experimental** techniques which seek to estimate treatment effects. I'll give a short rundown here to give readers an idea of what other options are out there. You can consult (Cunningham 2018) for an accessible introduction.

- **Instrumental variables (IV)**: IV relies on the presence of an **instrument**. This is a feature which affects the outcome $Y$ only through the treatment variable of interest $X$. It helps us deal with omitted variable $M$.

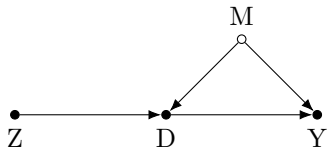



Figure 5: A basic example of an instrument $Z$ which lets us estimate the effect of $D$ on $Y$ even though there's an omitted variable $M$.

- **Difference-in-differences (DID)**: DID estimates the effect of an intervention by using a comparison case with the same time trend along the variable of interest. The canonical example is estimating the effect of minimum wage $D$ on employment $Y$ (Card and Krueger 1994). In 1992, New Jersey raised the minimum wage while neighboring Pennsylvania did not. This allowed studying fast food restaurants on either side of the state boundary to see whether employment changes happened in New Jersey. The critical assumption here is that NJ and PA had, and would continue to have had, similar time trends before and after the change. If this is true, observed differences can be attributed to the policy change.
- **Synthetic control (SC)**: Synthetic control is spiritually similar to DID (it's actually a generalization). It uses a set of control cases and a modeling procedure to combine the control cases into an optimal counterfactual for a given treatment case. The common case is cities: a city such as Seattle raises its minimum wage and we want to know the effect of the higher wage $D$ on employment $Y$. The SC strategy is to gather the big cities in the US (the "donor pool") and, looking at the pre-intervention period, take a weighted combination of the donor pool to produce a "synthetic Seattle" which matches Seattle's employment trends as closely as possible before the intervention. Then, we compare Seattle and synthetic Seattle after the intervention.
- **Regression discontinuity design (RDD)**: RDD notes that many interventions depend on cutoffs. For instance, you may only get into a good college if your SAT score is above a certain number. However, students just below and above the cutoff are likely to be quite similar, so we can study the effect of the intervention on the population around the cutoff by exploiting this discontinuity.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *The American Economic Review* 84 (4). http://web.sonoma.edu/users/c/cuellar/econ421/cardkrueger.pdf.

Cunningham, Scott. 2018. "Causal Inference: The Mixtape." http://scunning.com/cunningham_mixtape.pdf.

Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem

of Conditioning on a Collider Variable." *Annual Review of Sociology* 40 (1): 31–53. https://doi.org/10.1146/annurev-soc-071913-043455.

Ho, Daniel, Kosuke Imai, Gary King, Elizabeth Stuart, and Alex Whitworth. 2018. "Package 'MatchIt'." https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 243–63. https://doi.org/10.1111/rssb.12027.

King, Gary, and Richard Nielsen. 2015. "Why Propensity Scores Should Not Be Used for Matching," 32.

Kleinberg, By Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems" 105 (5): 491–95.

Morgan, SL, and C Winship. 2015. "Counterfactuals and Causal Inference: Methods and Principles for Social Research."

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

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55.

Sekhon, Jasjeet. 2009. *The Neyman— Rubin Model of Causal Inference and Estimation via Matching Methods*. Edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286546.003.0011.