# Causal Inference

Prediction and causation are very different. Typical questions are:

| | |
|---|---|
| Prediction: | Predict $Y$ after **observing** $X = x$ |
| Causation: | Predict $Y$ after **setting** $X = x$. |

Causation involves predicting the effect of an intervention. For example:

| | |
|---|---|
| Prediction: | Predict health given that a person takes vitamin C |
| Causation: | Predict health if I give a person vitamin C |

The difference between passively observing $X = x$ and actively intervening and setting $X = x$ is significant and requires different techniques and, typically, much stronger assumptions. This is the area known as *causal inference*.

# 1 Preliminaries

Before we jump into the details, there are a few general concepts to discuss.

## 1.1 Two Types of Causal Questions

There are two types of causal questions. The first deals with questions like this: do cell phones cause brain cancer? In this case, there are variables $X$ and $Y$ and we want to know the causal effect of $X$ on $Y$. The challenges are: find a parameter $\theta$ that characterizes the causal influence of $X$ on $Y$ and find a way to estimate $\theta$. This is usually what we mean when we refer to *causal inference*.

The second question is: given a set of variables, determine the causal relationship between the variables. This is called *causal discovery*. **This problem is statistically impossible** despite the large number of papers on the topic.

## 1.2 Two Types of Data

Data can be from a controlled, randomized experiment or from an observational study. In the former, $X$ is randomly assigned to subjects. In the latter, it is not randomly assigned. In randomized experiments, causal inference is straightforward. In observational (non-randomized) studies, the problem is much harder and requires stronger assumptions and also requires subject matter knowledge. Statistics and Machine Learning cannot solve causal problems without background knowledge.

## 1.3 Two Languages for Causation

There are two different mathematical languages for studying causation. The first is based on *counterfactuals*. The second is based on *causal graphs*. It will not seem obvious at first, but the two are mathematically equivalent (apart from some small details). Actually, there is a third language called *structural equation models* but this is very closely related to causal graphs.

## 1.4 Example

Consider this story. A mother notices that tall kids have a higher reading level than short kids. The mother puts her small child on a device and stretches the child until he is tall. She is dismayed to find out that his reading level has not changed.

The mother is correct that height and reading skill are **associated**. Put another way, you can use height to predict reading skill. But that does not imply that height *causes* reading skill. This is what statisticians mean when they say:

**correlation is not causation.**

On the other hand, consider smoking and lung cancer. We know that smoking and lung cancer are associated. But we also believe that smoking causes lung cancer. In this case, we recognize that intervening and forcing someone to smoke does change his probability of getting lung cancer.

## 1.5 Prediction Versus Causation

The difference between prediction (association/correlation) and causation is this: in prediction we are interested in

$$\mathbb{P}(Y \in A | X = x)$$

which means: the probability that $Y \in A$ given that we **observe** that $X$ is equal to $x$. For causation we are interested in

$$\mathbb{P}(Y \in A | \mathsf{set}\ X = x)$$

which means: the probability that $Y \in A$ given that we **set** $X$ equal to $x$. Prediction is about passive observation. Causation is about active intervention. The phrase **correlation is not causation** can be written mathematically as

$$\mathbb{P}(Y \in A | X = x) \neq \mathbb{P}(Y \in A | \mathsf{set}\ X = x).$$

Despite the fact that causation and association are different, people confuse them up all the time, even people trained in statistics and machine learning. On TV recently there was a report that good health is associated with getting seven hours of sleep. So far so good. Then the reporter goes on to say that, therefore, everyone should strive to sleep exactly seven hours so they will be healthy. Wrong. That's confusing causation and association. Another TV report pointed out a correlation between people who brush their teeth regularly and low

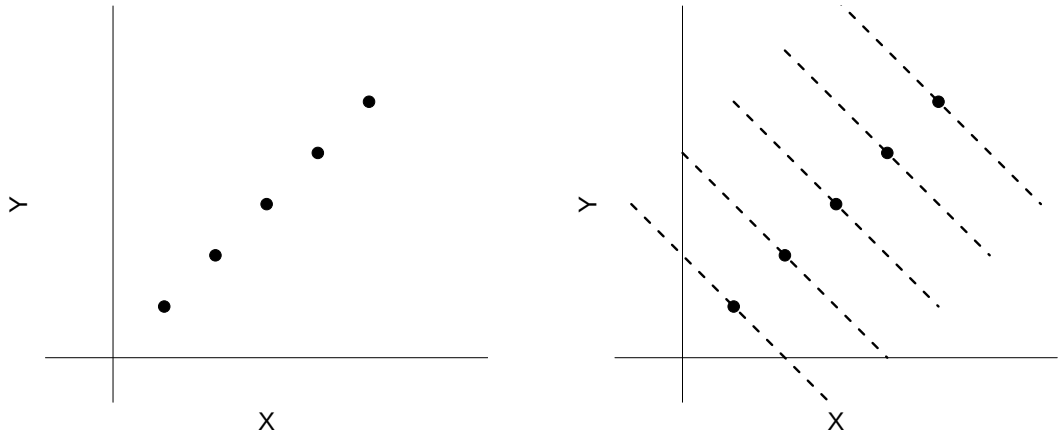


Figure 1: *Left: $X$ and $Y$ have positive association. Right: The lines are the counterfactuals, i.e. what would happen to each person if $I$ changed their $X$ value. Despite the positive association, the causal effect is negative. If we increase $X$ everyone's $Y$ values will decrease.*

rates of heart disease. An interesting correlation. Then the reporter (a doctor in this case) went on to urge people to brush their teeth to save their hearts. Wrong!

To avoid this confusion we need a way to discuss causation mathematically. That is, we need someway to make $\mathbb{P}(Y \in A|\mathsf{set}\ X = x)$ formal. As I mentioned earlier, there are two common ways to do this. One is to use **counterfactuals**. The other is to use **causal graphs**. There are two different languages for saying the same thing.

Causal inference is tricky and should be used with great caution. The main messages are:

1. Causal effects can be estimated consistently from randomized experiments.
2. It is difficult to estimate causal effects from observational (non-randomized) experiments.
3. All causal conclusions from observational studies should be regarded as very tentative.

Causal inference is a vast topic. We will only touch on the main ideas here.

# 2 Counterfactuals

Consider two variables $X$ and $Y$. We will call $X$ the "exposure" or the "treatment." We call $Y$ the "response" or the "outcome." For a given subject we see $(X_i, Y_i)$. What we don't see is what their value of $Y_i$ would have been if we changed their value of $X_i$. This is called

the counterfactual. The whole causal story is made clear in Figure 1 which shows data (left) and the counterfactuals (right).

Suppose now that $X$ is a binary variable that represents some exposure. So $X = 1$ means the subject was exposed and $X = 0$ means the subject was not exposed. We can address the problem of predicting $Y$ from $X$ by estimating $\mathbb{E}(Y|X = x)$. To address causal questions, we introduce *counterfactuals*. Let $Y_1$ denote the response if the subject is exposed. Let $Y_0$ denote the response if the subject is not exposed. Then

$$Y = \begin{cases} Y_1 & \text{if } X = 1 \\ Y_0 & \text{if } X = 0. \end{cases}$$

More succinctly

$$Y = XY_1 + (1 - X)Y_0. \tag{1}$$

If we expose a subject, we observe $Y_1$ but we do not observe $Y_0$. Indeed, $Y_0$ is the value we would have observed if the subject had been exposed. The unobserved variable is called a *counterfactual*. The variables $(Y_0, Y_1)$ are also called *potential outcomes*. We have enlarged our set of variables from $(X, Y)$ to $(X, Y, Y_0, Y_1)$. A small dataset might look like this:

| $X$ | $Y$ | $Y_0$ | $Y_1$ |
|---|---|---|---|
| 1 | 1 | * | 1 |
| 1 | 1 | * | 1 |
| 1 | 0 | * | 0 |
| 1 | 1 | * | 1 |
| 0 | 1 | 1 | * |
| 0 | 0 | 0 | * |
| 0 | 1 | 1 | * |
| 0 | 1 | 1 | * |

The asterisks indicate unobserved variables. Causal questions involve the the distribution $p(y_0, y_1)$ of the potential outcomes. We can interpret $p(y_1)$ as $p(y|\mathsf{set}\ X = 1)$ and we can interpret $p(y_0)$ as $p(y|\mathsf{set}\ X = 0)$. The *mean treatment effect* or *mean causal effect* is defined by

$$\theta = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \mathbb{E}(Y|\mathsf{set}\ X = 1) - \mathbb{E}(Y|\mathsf{set}\ X = 0).$$

The parameter $\theta$ has the following interpretation: $\theta$ is the mean response if we exposed everyone minus the mean response if we exposed no-one.

**Lemma 1** *In general,*

$$\mathbb{E}[Y_1] \neq \mathbb{E}[Y|X = 1] \quad \text{and} \quad \mathbb{E}[Y_0] \neq \mathbb{E}[Y|X = 0].$$

**Exercise:** Prove this.

Suppose now that we observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Can we estimate $\theta$? In general the answer is no. We can estimate

$$\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$$

but $\alpha$ is not equal to $\theta$. Quantities like $\mathbb{E}(Y|X = 1)$ and $\mathbb{E}(Y|X = 0)$ are predictive parameters. These are things that are commonly estimated in statistics and machine learning. In general, we cannot consistently estimate $\theta$.

## 2.1 Two Ways to Make $\theta$ Estimable

Fortunately, there are two ways to make $\theta$ estimable. The first is randomization and the second is adjusting for confounding.

**Randomization.** Suppose that we randomly assign $X$. Then $X$ will be independent of $(Y_0, Y_1)$. In symbols:

$$\text{random treatment assignment implies : } \ (Y_0, Y_1) \amalg X.$$

**Warning! Note that $X$ is not independent of $Y$.**

If $X$ is randomly assigned, then $\theta = \alpha$ where

$$\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0).$$

A consistent estimator of $\alpha$ (and hence $\theta$) is the difference of means $\overline{Y}_1 - \overline{Y}_0$.

To summarize: **If $X$ is randomly assigned then correlation = causation.** This is why people spend millions of dollars doing randomized experiments.

The same results hold when $X$ is continuous. In this case there is a counterfactual $Y(x)$ for each value $x$ of $X$. We again have that, in general,

$$\mathbb{E}[Y(x)] \ \neq \mathbb{E}[Y|X = x].$$

See Figure 1. But if $X$ is randomly assigned, then we do have $\mathbb{E}[Y(x)] = \mathbb{E}[Y|X = x]$ and so $\mathbb{E}[Y(x)]$ can be consistently estimated using standard regression methods. Indeed, if we had randomly chosen the $X$ values in Figure 1 then the plot on the left would have been downward sloping. To see this, note that $\theta(x) = \mathbb{E}[Y(x)]$ is defined to be the average of the lines in the right plot. Under randomization, $X$ is independent of $Y(x)$. So

$$\text{right plot} = \theta(x) = \mathbb{E}[Y(x)] = \mathbb{E}[Y(x)|X = x] = \mathbb{E}[Y|X = x] = \text{left plot}.$$

**Adjusting For Confounding.** In some cases it is not feasible to do a randomized experiment and we must use data from from observational (non-randomized) studies. Smoking and lung cancer is an example. Can we estimate causal parameters from observational (non-randomized) studies? The answer is: sort of.

In an observational study, the treated and untreated groups will not be comparable. Maybe the healthy people chose to take the treatment and the unhealthy people didn't. In other words, $X$ is not independent of $(Y_0, Y_1)$. The treatment may have no effect but we would still see a strong association between $Y$ and $X$. In other words, $\alpha$ might be large even though $\theta = 0$.

Here is a simplified example. Suppose $X$ denotes whether someone takes vitamins and $Y$ is some binary health outcome (with $Y = 1$ meaning "healthy.")

| $X$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|
| $Y_0$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| $Y_1$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |

In this example, there are only two types of people: healthy and unhealthy. The healthy people have $(Y_0, Y_1) = (1, 1)$. These people are healthy whether or not that take vitamins. The unhealthy people have $(Y_0, Y_1) = (0, 0)$. These people are unhealthy whether or not that take vitamins. The observed data are:

| $X$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|
| $Y$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0. |

In this example, $\theta = 0$ but $\alpha = 1$. The problem is that people who choose to take vitamins are different than people who choose not to take vitamins. That's just another way of saying that $X$ is not independent of $(Y_0, Y_1)$.

To account for the differences in the groups, we can measure **confounding variables**. These are the variables that affect both $X$ and $Y$. These variables explain why the two groups of people are different. In other words, these variables account for the dependence between $X$ and $(Y_0, Y_1)$. By definition, there are no such variables in a randomized experiment. The hope is that if we measure enough confounding variables $Z = (Z_1, \ldots, Z_k)$, then, perhaps the treated and untreated groups will be comparable, conditional on $Z$. This means that $X$ is independent of $(Y_0, Y_1)$ conditional on $Z$. We say that there is *no unmeasured* confounding, or that *ignorability holds*, if

$$X \;\amalg\; (Y_0, Y_1) \;\Big|\; Z.$$

The only way to measure the important confounding variables is to use subject matter knowledge. In other words, **causal inference in observational studies is not possible without subject matter knowledge.**

**Theorem 2** *Suppose that*

$$X \ \text{II} \ (Y_0, Y_1) \ \Big| \ Z.$$

*Then*

$$\theta \equiv \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \int \mu(1,z)p(z)dz - \int \mu(0,z)p(z)dz \tag{2}$$

*where*

$$\mu(x,z) = \mathbb{E}(Y|X=x, Z=z).$$

*A consistent estimator of* $\theta$ *is*

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(1,Z_i) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(0,Z_i)$$

*where* $\widehat{\mu}(x,z)$ *is an appropriate, consistent estimator of the regression function* $\mu(x,z) = \mathbb{E}[Y|X=x,Z=z]$.

**Remark:** Estimating the quantity in (2) well is difficult and involves an area of statistics called *semiparametric inference*. In statistics, biostatistics, econometrics and epidemiology, this is the focus of much research.

**Proof.** We have

$$\begin{aligned}
\theta &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \\
&= \int \mathbb{E}(Y_1|Z=z)p(z)dz - \int \mathbb{E}(Y_0|Z=z)p(z)dz \\
&= \int \mathbb{E}(Y_1|X=1,Z=z)p(z)dz - \int \mathbb{E}(Y_0|X=0,Z=z)p(z)dz \\
&= \int \mathbb{E}(Y|X=1,Z=z)p(z)dz - \int \mathbb{E}(Y|X=0,Z=z)p(z)dz \qquad (3)
\end{aligned}$$

where we used the fact that $X$ is independent of $(Y_0, Y_1)$ conditional on $Z$ in the third line and the fact that $Y = (1-X)Y_1 + XY_0$ in the fourth line. ■

The process of including confounding variables and using equation (2) is known as *adjusting for confounders* and $\widehat{\theta}$ is called the *adjusted treatment effect*. The choice of the estimator $\widehat{\mu}(x,z)$ is delicate. If we use a nonparametric method then we have to choose the smoothing parameter carefully. Unlike prediction, bias and variance are not equally important. **The**

**usual bias-variance tradeoff does not apply.** In fact bias is worse than variance and we need to choose the smoothing parameter smaller than usual. As mentioned above, there is a branch of statistics called *semiparametric inference* that deals with this problem in detail.

It is instructive to compare the casual effect

$$\theta = \int \mu(1,z)p(z)dz - \int \mu(0,z)p(z)dz$$

with the predictive quantity

$$\begin{aligned} \alpha &= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \\ &= \int \mu(1,z)p(z|X=1)dz - \int \mu(0,z)p(z|X=0)dz \end{aligned}$$

which are mathematically (and conceptually) quite different.

We need to treat $\widehat{\theta}$ cautiously. It is very unlikely that we have successfully measured all the relevant confounding variables so $\widehat{\theta}$ should be regarded as a crude approximation to $\theta$ at best.

In the case where $\mathbb{E}[Y|X=x, Z=z]$ is linear, the adjusted treatment effect takes a simple form. Suppose that $\mathbb{E}[Y|X=x, Z=z] = \beta_0 + \beta_1 x + \beta_2^T z$. Then

$$\theta = \int [\beta_0 + \beta_1 + \beta_2^T z]dP(z) - \int [\beta_0 + \beta_2^T z]dP(z) = \beta_1.$$

In a linear regression, the coefficient in front of $x$ is the causal effect of $x$ if (i) the model is correct and (ii) all confounding variables are included in the regression.

To summarize: the coefficients in linear regression have a causal intepretation if (i) the model is correct and (ii) every possible confounding factor is included in the model.