16
Causal Inference

Roughly speaking, the statement "X causes Y" means that changing the value of X will change the distribution of Y. When X causes Y, X and Y will be associated but the reverse is not, in general, true. Association does not necessarily imply causation. We will consider two frameworks for discussing causation. The first uses **counterfactual** random variables. The second, presented in the next chapter, uses **directed acyclic graphs.**

16.1 The Counterfactual Model

Suppose that X is a binary treatment variable where $X = 1$ means "treated" and $X = 0$ means "not treated." We are using the word "treatment" in a very broad sense. Treatment might refer to a medication or something like smoking. An alternative to "treated/not treated" is "exposed/not exposed" but we shall use the former.

Let Y be some outcome variable such as presence or absence of disease. To distinguish the statement "X is associated Y" from the statement "X causes Y" we need to enrich our probabilistic vocabulary. Specifically, we will decompose the response Y into a more fine-grained object.

We introduce two new random variables (C_0, C_1), called **potential outcomes** with the following interpretation: C_0 is the outcome if the subject is

not treated ($X = 0$) and C_1 is the outcome if the subject is treated ($X = 1$). Hence,

$$Y = \begin{cases} C_0 & \text{if } X = 0 \\ C_1 & \text{if } X = 1. \end{cases}$$

We can express the relationship between Y and (C_0, C_1) more succinctly by

$$Y = C_X. \tag{16.1}$$

Equation (16.1) is called the **consistency relationship**.

Here is a toy dataset to make the idea clear:

X	Y	C_0	C_1
0	4	4	*
0	7	7	*
0	2	2	*
0	8	8	*
1	3	*	3
1	5	*	5
1	8	*	8
1	9	*	9

The asterisks denote unobserved values. When $X = 0$ we don't observe C_1, in which case we say that C_1 is a **counterfactual** since it is the outcome you would have had if, counter to the fact, you had been treated ($X = 1$). Similarly, when $X = 1$ we don't observe C_0, and we say that C_0 is **counterfactual**. There are four types of subjects:

Type	C_0	C_1
Survivors	1	1
Responders	0	1
Anti-responders	1	0
Doomed	0	0

Think of the potential outcomes (C_0, C_1) as hidden variables that contain all the relevant information about the subject.

Define the **average causal effect** or **average treatment effect** to be

$$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0). \tag{16.2}$$

The parameter θ has the following interpretation: θ is the mean if everyone were treated ($X = 1$) minus the mean if everyone were not treated ($X = 0$). There are other ways of measuring the causal effect. For example, if C_0 and C_1 are binary, we define the **causal odds ratio**

$$\frac{\mathbb{P}(C_1 = 1)}{\mathbb{P}(C_1 = 0)} \div \frac{\mathbb{P}(C_0 = 1)}{\mathbb{P}(C_0 = 0)}$$

and the **causal relative risk**

$$\frac{\mathbb{P}(C_1 = 1)}{\mathbb{P}(C_0 = 1)}.$$

The main ideas will be the same whatever causal effect we use. For simplicity, we shall work with the average causal effect θ.

Define the **association** to be

$$\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0). \tag{16.3}$$

Again, we could use odds ratios or other summaries if we wish.

16.1 Theorem (Association Is Not Causation). *In general,* $\theta \neq \alpha$.

16.2 Example. Suppose the whole population is as follows:

X	Y	C_0	C_1
0	0	0	0*
0	0	0	0*
0	0	0	0*
0	0	0	0*
1	1	1*	1
1	1	1*	1
1	1	1*	1
1	1	1*	1

Again, the asterisks denote unobserved values. Notice that $C_0 = C_1$ for every subject, thus, this treatment has no effect. Indeed,

$$\begin{aligned}
\theta &= \mathbb{E}(C_1) - \mathbb{E}(C_0) = \frac{1}{8}\sum_{i=1}^{8} C_{1i} - \frac{1}{8}\sum_{i=1}^{8} C_{0i} \\
&= \frac{0+0+0+0+1+1+1+1}{8} - \frac{0+0+0+0+1+1+1+1}{8} \\
&= 0.
\end{aligned}$$

Thus, the average causal effect is 0. The observed data are only the X's and Y's, from which we can estimate the association:

$$\begin{aligned}
\alpha &= \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0) \\
&= \frac{1+1+1+1}{4} - \frac{0+0+0+0}{4} = 1.
\end{aligned}$$

Hence, $\theta \neq \alpha$.

To add some intuition to this example, imagine that the outcome variable is 1 if "healthy" and 0 if "sick". Suppose that $X = 0$ means that the subject

does not take vitamin C and that $X = 1$ means that the subject does take vitamin C. Vitamin C has no causal effect since $C_0 = C_1$ for each subject. In this example there are two types of people: healthy people $(C_0, C_1) = (1, 1)$ and unhealthy people $(C_0, C_1) = (0, 0)$. Healthy people tend to take vitamin C while unhealthy people don't. It is this association between (C_0, C_1) and X that creates an association between X and Y. If we only had data on X and Y we would conclude that X and Y are associated. Suppose we wrongly interpret this causally and conclude that vitamin C prevents illness. Next we might encourage everyone to take vitamin C. If most people comply with our advice, the population will look something like this:

X	Y	C_0	C_1
0	0	0	0^*
1	0	0	0^*
1	0	0	0^*
1	0	0	0^*
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1

Now $\alpha = (4/7) - (0/1) = 4/7$. We see that α went down from 1 to 4/7. Of course, the causal effect never changed but the naive observer who does not distinguish association and causation will be confused because his advice seems to have made things worse instead of better. ■

In the last example, $\theta = 0$ and $\alpha = 1$. It is not hard to create examples in which $\alpha > 0$ and yet $\theta < 0$. The fact that the association and causal effects can have different signs is very confusing to many people.

The example makes it clear that, in general, we cannot use the association to estimate the causal effect θ. The reason that $\theta \neq \alpha$ is that (C_0, C_1) was not independent of X. That is, treatment assignment was not independent of person type.

Can we ever estimate the causal effect? The answer is: sometimes. In particular, random assignment to treatment makes it possible to estimate θ.

16.3 Theorem. *Suppose we randomly assign subjects to treatment and that* $\mathbb{P}(X = 0) > 0$ *and* $\mathbb{P}(X = 1) > 0$. *Then* $\alpha = \theta$. *Hence, any consistent estimator of* α *is a consistent estimator of* θ. *In particular, a consistent estimator is*

$$\widehat{\theta} \quad = \quad \widehat{\mathbb{E}}(Y|X = 1) - \widehat{\mathbb{E}}(Y|X = 0)$$

$$= \quad \overline{Y}_1 - \overline{Y}_0$$

is a consistent estimator of θ*, where*

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} Y_i X_i, \qquad \overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-X_i),$$

$n_1 = \sum_{i=1}^{n} X_i$*, and* $n_0 = \sum_{i=1}^{n}(1-X_i)$*.*

PROOF. Since X is randomly assigned, X is independent of (C_0, C_1). Hence,

$$\begin{aligned}
\theta &= \mathbb{E}(C_1) - \mathbb{E}(C_0) \\
&= \mathbb{E}(C_1|X=1) - \mathbb{E}(C_0|X=0) \qquad \text{since } X \amalg (C_0, C_1) \\
&= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \qquad \text{since } Y = C_X \\
&= \alpha.
\end{aligned}$$

The consistency follows from the law of large numbers. ■

If Z is a covariate, we define the **conditional causal effect** by

$$\theta_z = \mathbb{E}(C_1|Z=z) - \mathbb{E}(C_0|Z=z).$$

For example, if Z denotes gender with values $Z = 0$ (women) and $Z = 1$ (men), then θ_0 is the causal effect among women and θ_1 is the causal effect among men. In a randomized experiment, $\theta_z = \mathbb{E}(Y|X=1, Z=z) - \mathbb{E}(Y|X=0, Z=z)$ and we can estimate the conditional causal effect using appropriate sample averages.

Summary of the Counterfactual Model
Random variables: (C_0, C_1, X, Y).
Consistency relationship: $Y = C_X$.
Causal Effect: $\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0)$.
Association: $\alpha = \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0)$.
Random assignment $\implies$ $(C_0, C_1) \amalg X$ $\implies$ $\theta = \alpha$.

16.2 Beyond Binary Treatments

Let us now generalize beyond the binary case. Suppose that $X \in \mathcal{X}$. For example, X could be the dose of a drug in which case $X \in \mathbb{R}$. The counterfactual vector (C_0, C_1) now becomes the **counterfactual function** $C(x)$ where

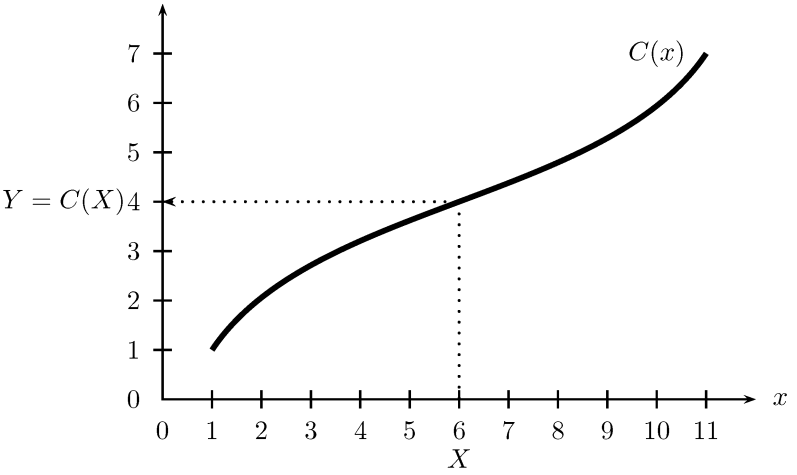

FIGURE 16.1. A counterfactual function $C(x)$. The outcome Y is the value of the curve $C(x)$ evaluated at the observed dose X.

$C(x)$ is the outcome a subject would have if he received dose x. The observed response is given by the consistency relation

$$Y \equiv C(X). \tag{16.4}$$

See Figure 16.1. The **causal regression function** is

$$\theta(x) = \mathbb{E}(C(x)). \tag{16.5}$$

The regression function, which measures association, is $r(x) = \mathbb{E}(Y|X = x)$.

16.4 Theorem. *In general, $\theta(x) \neq r(x)$. However, when X is randomly assigned, $\theta(x) = r(x)$.*

16.5 Example. An example in which $\theta(x)$ is constant but $r(x)$ is not constant is shown in Figure 16.2. The figure shows the counterfactual functions for four subjects. The dots represent their X values X_1, X_2, X_3, X_4. Since $C_i(x)$ is constant over x for all i, there is no causal effect and hence

$$\theta(x) = \frac{C_1(x) + C_2(x) + C_3(x) + C_4(x)}{4}$$

is constant. Changing the dose x will not change anyone's outcome. The four dots in the lower plot represent the observed data points $Y_1 = C_1(X_1), Y_2 = C_2(X_2), Y_3 = C_3(X_3), Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. Although there is no causal effect, there is an association since the regression curve $r(x)$ is not constant. ■

16.3 Observational Studies and Confounding

A study in which treatment (or exposure) is not randomly assigned is called an **observational study.** In these studies, subjects select their own value of the exposure X. Many of the health studies you read about in the newspaper are like this. As we saw, association and causation could in general be quite different. This discrepancy occurs in non-randomized studies because the potential outcome C is not independent of treatment X. However, suppose we could find groupings of subjects such that, within groups, X and $\{C(x) : x \in \mathcal{X}\}$ are independent. This would happen if the subjects are very similar within groups. For example, suppose we find people who are very similar in age, gender, educational background, and ethnic background. Among these people we might feel it is reasonable to assume that the choice of X is essentially random. These other variables are called **confounding variables.**[1] If we denote these other variables collectively as Z, then we can express this idea by saying that

$$\{C(x) : x \in \mathcal{X}\} \amalg X|Z. \tag{16.6}$$

Equation (16.6) means that, within groups of Z, the choice of treatment X does not depend on type, as represented by $\{C(x) : x \in \mathcal{X}\}$. If (16.6) holds and we observe Z then we say that there is **no unmeasured confounding**.

16.6 Theorem. *Suppose that (16.6) holds. Then,*

$$\theta(x) = \int \mathbb{E}(Y|X = x, Z = z) dF_Z(z) dz. \tag{16.7}$$

If $\widehat{r}(x, z)$ is a consistent estimate of the regression function $\mathbb{E}(Y|X = x, Z = z)$, then a consistent estimate of $\theta(x)$ is

$$\widehat{\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} \widehat{r}(x, Z_i).$$

[1] A more precise definition of confounding is given in the next chapter.

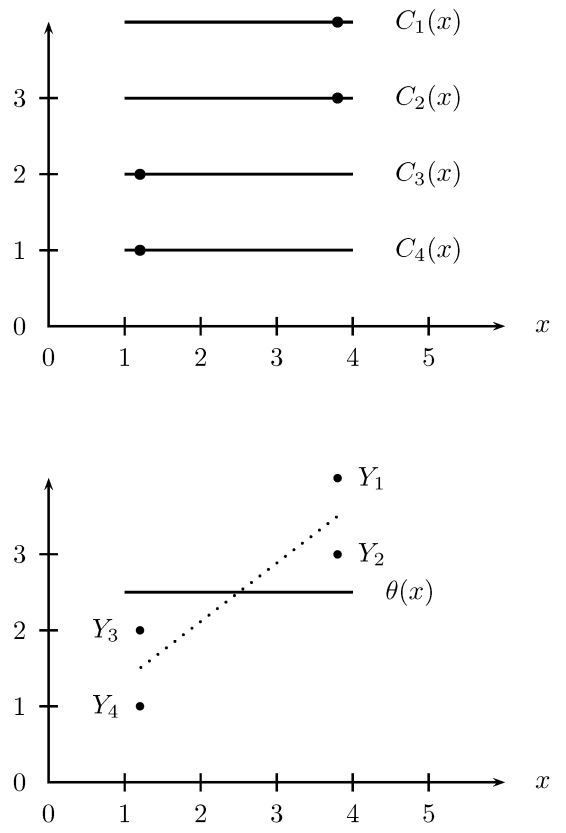

FIGURE 16.2. The top plot shows the counterfactual function $C(x)$ for four subjects. The dots represent their X values. Since $C_i(x)$ is constant over x for all i, there is no causal effect. Changing the dose will not change anyone's outcome. The lower plot shows the causal regression function $\theta(x) = (C_1(x) + C_2(x) + C_3(x) + C_4(x))/4$. The four dots represent the observed data points $Y_1 = C_1(X_1)$, $Y_2 = C_2(X_2)$, $Y_3 = C_3(X_3)$, $Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. There is no causal effect since $C_i(x)$ is constant for all i. But there is an association since the regression curve $r(x)$ is not constant.

In particular, if $r(x,z) = \beta_0 + \beta_1 x + \beta_2 z$ *is linear, then a consistent estimate of* $\theta(x)$ *is*

$$\widehat{\theta}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 \overline{Z}_n \tag{16.8}$$

where $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ *are the least squares estimators.*

16.7 Remark. It is useful to compare equation (16.7) to $\mathbb{E}(Y|X=x)$ which can be written as $\mathbb{E}(Y|X=x) = \int \mathbb{E}(Y|X=x, Z=z) dF_{Z|X}(z|x)$.

Epidemiologists call (16.7) the **adjusted treatment effect**. The process of computing adjusted treatment effects is called **adjusting (or controlling) for confounding.** The selection of what confounders Z to measure and control for requires scientific insight. Even after adjusting for confounders, we cannot be sure that there are not other confounding variables that we missed. This is why observational studies must be treated with healthy skepticism. Results from observational studies start to become believable when: (i) the results are replicated in many studies, (ii) each of the studies controlled for plausible confounding variables, (iii) there is a plausible scientific explanation for the existence of a causal relationship.

A good example is smoking and cancer. Numerous studies have shown a relationship between smoking and cancer even after adjusting for many confounding variables. Moreover, in laboratory studies, smoking has been shown to damage lung cells. Finally, a causal link between smoking and cancer has been found in randomized animal studies. It is this collection of evidence over many years that makes this a convincing case. One single observational study is not, by itself, strong evidence. Remember that when you read the newspaper.

16.4 Simpson's Paradox

Simpson's paradox is a puzzling phenomenon that is discussed in most statistics texts. Unfortunately, most explanations are confusing (and in some cases incorrect). The reason is that it is nearly impossible to explain the paradox without using counterfactuals (or directed acyclic graphs).

Let X be a binary treatment variable, Y a binary outcome, and Z a third binary variable such as gender. Suppose the joint distribution of X, Y, Z is

	$Y=1$	$Y=0$	$Y=1$	$Y=0$
$X=1$	.1500	.2250	.1000	.0250
$X=0$	.0375	.0875	.2625	.1125
	$Z=1$ (men)		$Z=0$ (women)	

The marginal distribution for (X, Y) is

	$Y=1$	$Y=0$	
$X=1$	.25	.25	.50
$X=0$	.30	.20	.50
	.55	.45	1

From these tables we find that,

$$\begin{aligned} \mathbb{P}(Y=1|X=1) - \mathbb{P}(Y=1|X=0) &= -0.1 \\ \mathbb{P}(Y=1|X=1, Z=1) - \mathbb{P}(Y=1|X=0, Z=1) &= 0.1 \\ \mathbb{P}(Y=1|X=1, Z=0) - \mathbb{P}(Y=1|X=0, Z=0) &= 0.1. \end{aligned}$$

To summarize, we *seem* to have the following information:

Mathematical Statement	English Statement?
$\mathbb{P}(Y=1\|X=1) < \mathbb{P}(Y=1\|X=0)$	treatment is harmful
$\mathbb{P}(Y=1\|X=1, Z=1) > \mathbb{P}(Y=1\|X=0, Z=1)$	treatment is beneficial to men
$\mathbb{P}(Y=1\|X=1, Z=0) > \mathbb{P}(Y=1\|X=0, Z=0)$	treatment is beneficial to women

Clearly, something is amiss. There can't be a treatment which is good for men, good for women, but bad overall. This is nonsense. The problem is with the set of English statements in the table. Our translation from math into English is specious.

> **The inequality $\mathbb{P}(Y=1|X=1) < \mathbb{P}(Y=1|X=0)$ does not mean that treatment is harmful.**

The phrase "treatment is harmful" should be written mathematically as $\mathbb{P}(C_1 = 1) < \mathbb{P}(C_0 = 1)$. The phrase "treatment is harmful for men" should be written $\mathbb{P}(C_1 = 1|Z = 1) < \mathbb{P}(C_0 = 1|Z = 1)$. The three mathematical statements in the table are not at all contradictory. It is only the translation into English that is wrong.

Let us now show that a real Simpson's paradox cannot happen, that is, there cannot be a treatment that is beneficial for men and women but harmful overall. Suppose that treatment is beneficial for both sexes. Then

$$\mathbb{P}(C_1 = 1|Z = z) > \mathbb{P}(C_0 = 1|Z = z)$$

for all z. It then follows that

$$\begin{aligned}\mathbb{P}(C_1 = 1) &= \sum_z \mathbb{P}(C_1 = 1|Z = z)\mathbb{P}(Z = z) \\ &> \sum_z \mathbb{P}(C_0 = 1|Z = z)\mathbb{P}(Z = z) \\ &= \mathbb{P}(C_0 = 1).\end{aligned}$$

Hence, $\mathbb{P}(C_1 = 1) > \mathbb{P}(C_0 = 1)$, so treatment is beneficial overall. No paradox.

16.5 Bibliographic Remarks

The use of potential outcomes to clarify causation is due mainly to Jerzy Neyman and Donald Rubin. Later developments are due to Jamie Robins, Paul Rosenbaum, and others. A parallel development took place in econometrics by various people including James Heckman and Charles Manski. Texts on causation include Pearl (2000), Rosenbaum (2002), Spirtes et al. (2000), and van der Laan and Robins (2003).

16.6 Exercises

1. Create an example like Example 16.2 in which $\alpha > 0$ and $\theta < 0$.

2. Prove Theorem 16.4.

3. Suppose you are given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an observational study, where $X_i \in \{0, 1\}$ and $Y_i \in \{0, 1\}$. Although it is not possible to estimate the causal effect θ, it is possible to put bounds on θ. Find upper and lower bounds on θ that can be consistently estimated from the data. Show that the bounds have width 1.

   Hint: Note that $\mathbb{E}(C_1) = \mathbb{E}(C_1|X = 1)\mathbb{P}(X = 1) + \mathbb{E}(C_1|X = 0)\mathbb{P}(X = 0)$.

4. Suppose that $X \in \mathbb{R}$ and that, for each subject i, $C_i(x) = \beta_{1i}x$. Each subject has their own slope β_{1i}. Construct a joint distribution on (β_1, X) such that $\mathbb{P}(\beta_1 > 0) = 1$ but $\mathbb{E}(Y|X = x)$ is a decreasing function of x, where $Y = C(X)$. Interpret.

5. Let $X \in \{0, 1\}$ be a binary treatment variable and let (C_0, C_1) denote the corresponding potential outcomes. Let $Y = C_X$ denote the observed

response. Let F_0 and F_1 be the cumulative distribution functions for C_0 and C_1. Assume that F_0 and F_1 are both continuous and strictly increasing. Let $\theta = m_1 - m_0$ where $m_0 = F_0^{-1}(1/2)$ is the median of C_0 and $m_1 = F_1^{-1}(1/2)$ is the median of C_1. Suppose that the treatment X is assigned randomly. Find an expression for θ involving only the joint distribution of X and Y.