LECTURE 2, PART III: CAUSAL INFERENCE

Text references: Chapter 16 from *All of Statistics* by Larry Wasserman (on Canvas)

Association versus Causation

There is much confusion about the difference between causation and association. 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.

For example, there is a strong linear relationship between death rate due to breast cancer and fat intake. So,

$$\text{DEATH BY CANCER} = \beta_0 + \beta_1 \text{FAT } + \epsilon \tag{1}$$

where $\beta_1 > 0$. Does that mean that FAT causes breast cancer? Consider two interpretations of (1).

ASSOCIATION (or correlation). Fat intake and breast cancer are associated. Therefore, if I observe someone's fat intake, I can use equation (1) to predict their chance of dying from breast cancer.

> CAUSATION. Fat intake causes Breast cancer. Therefore, if I observe someone's fat intake, I can use equation (1) to predict their chance of dying from breast cancer. Moreover, if I change someone's fat intake by one unit, their risk of death from breast cancer changes by β_1.

If the data are from a **randomized study** (X is randomly assigned) then the causal interpretation is correct. If the data are from an **observational study**, (X is not randomly assigned) then the association interpretation is correct. To intuitively see why the causal interpretation is wrong in the observational study, consider the following example:

Example. Suppose that people with high fat intake are the rich people. And suppose, for the sake of the example, that rich people smoke a lot. Further, suppose that smoking does cause cancer. Then it will be true that high fat intake predicts high cancer rate. But changing someone's fat intake may not change their cancer risk.

How can we make these ideas precise? The answer is to use either **directed acyclic graphs** (DAG's) as in the figure above [see AOS Chapter 17 or Shalizi Chapters 21-24 if interested] or to use **counterfactuals**. In this course, we will discuss causation using the idea of counterfactual random variables which, as we shall see, relate naturally to regression analysis.

Counterfactual Model for Binary Treatments

[Reference: Sec 16.1 in *All of Statistics*]

1. Suppose that X is a **binary treatment variable** where $X = 1$ means "treated" and $X = 0$ means "not treated". 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.

2. Let Y be some **outcome variable** such as presence or absence of disease.

3. To distinguish the statement "X is associated Y" from the statement "X causes Y" we need to enrich our probabilistic vocabulary. 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; that is if $X = 0$; and $C(1)$ is the outcome if the subject is treated; that is, if $X = 1$. Hence, the response Y is given by

More succinctly, we write

$$Y = C(X)$$

which is called the **consistency relationship**. You can think of the potential outcomes $\{C(0), C(1)\}$ as hidden variables that contain all the relevant information about the subject.

To be clear: Suppose that n subjects receive binary treatments $X_1, ..., X_n$ and that their respective potential outcomes are given by $\{C_j(0), C_j(1)\}$ for $j = 1, \ldots, n$. What are the respective *observed* outcomes? What are the unobserved variables?

Toy Example 1. (Denote the *counterfactuals* by *; these are outcomes you would have had if, counter to the fact, you had been treated or not treated.)

Now define the **average causal effect** or **average treatment effect**, θ:

Define the **association**, α:

We can estimate the association α from data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and by LLN, $\widehat{\alpha}$ is a consistent estimator of α:

Note: In general, **association is not causation**, that is, $\theta \neq \alpha$.

Toy Example 2. Suppose that we want to investigate the causal relationship between treatment X and health Y, but we have two subpopulations: population A (wealthy subjects) and population B (poor). Only the wealthy can afford treatment and they tend to be (or to simplify, is always) healthy for other reasons, and the poor cannot afford treatment and they tend to be (or, to simplify, is always) unhealthy. Then an **observational** prospective study would look something like this:

We see that the association $\widehat{\alpha}$ is a poor estimate of θ!

Compare the above to a **randomized** study that may look like this:

Bottomline: [See AOS Theorem 16.1 and 16.3]

Beyond Binary Treatments

[Reference: Sec 16.2 in *All of Statistics*]

Now generalize beyond binary treatments to a continuous random variable X. Let Y be the outcome variable. In the binary treatment case with $x \in \{0, 1\}$, we only had two potential outcomes $\{C(0), C(1)\}$ for each subject. For continuous treatment variables, we will introduce a **counterfactual function** $C(x)$, where $C(x)$ is the outcome a subject would have if he received dose $x \in \mathbb{R}$.

Similar to before, the observed response is given by the consistency rela-

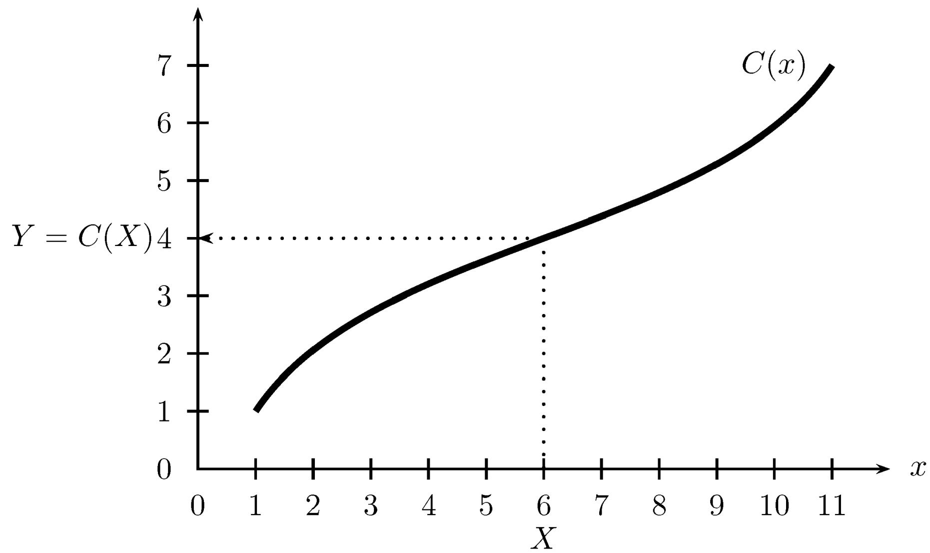

FIGURE 16.1. A counterfactual function $C(x)$. The outcome Y is the value of the curve $C(x)$ evaluated at the observed dose X.

Figure 1: From "All of Statistics" by Wasserman

tion

$$Y \equiv C(X).$$

To be clear: Suppose that there are n subjects whose observations will be denoted $(X_1, Y_1), ..., (X_n, Y_n)$. For each subject j, there is a counterfactual function $C_j(x)$ that gives, for each x, the potential observation (possible value of Y_j) that will be observed if $X_j = x$. Each subject has a $C_j(x)$ function of the sort plotted in Figure 1. They can be all different shapes. Different subjects functions can cross each other. Some might be decreasing while others are not monotone at all.

In the setting with continuous treatment variables, the tables that appeared in the Toy Examples are much harder to display because (i) each $\{C_j(0), C_j(1)\}$ pair (one for each row of the table) needs to be replaced by an entire function $C_j(x)$, and (ii) all of the values of $C_j(x)$ are counterfactual except for the one that corresponds to that subjects value of X.

Since there are more than two "treatments" with a continuous X, there is no natural analog to the causal effect $\theta = \mathbb{E}[C(1)] - \mathbb{E}[C(0)]$. Instead, we define the **causal regression function** $\theta(x)$ as

Similarly, we replace the association $\alpha = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 1]$ with the **regression function** $r(x)$, where

Take a close look at Fig 16.2 in AOS p.258.

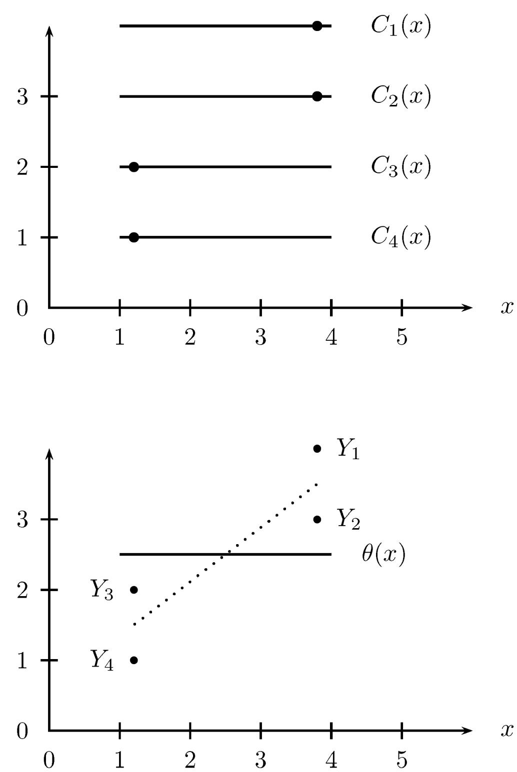

FIGURE 16.2. The top plot shows the counterfactual function $C(x)$ for four subjects. The dots represent their X values. Since $C_i(x)$ is constant over x for all i, there is no causal effect. Changing the dose will not change anyone's outcome. The lower plot shows the causal regression function $\theta(x) = (C_1(x)+C_2(x)+C_3(x)+C_4(x))/4$. 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.

Figure 2: From "All of Statistics" by Wasserman

Important result [AOS Theorem 16.4]: In general, $\theta(x) \neq r(x)$. However, when X is randomly assigned, $\theta(x) = r(x)$. Can you show this?

Follow-up question: So what do we do if we have an observational study? How do we attempt to study the causal effect of one variable X on another variable Y?

Observational Studies and Confounding

[Reference: Sec 16.3 in *All of Statistics*]

Key: *Adjust for confounding bias* by measuring so-called **confounding variables**, or variables that depend on both X and Y. Suppose Z is the entire collection of confounding variables such that

Roughly speaking, the general idea (in matching by Z as well as adjusting for confounding by regression) is to:

1. Partition the population into groups that are (approximately) homogeneous relative to Z

2. Assess the effect of X on Y in each homogeneous group

3. Average the results over Z

Toy Example 2 revisited

Note that **within groups of** Z, the choice of treatment X does not depend on type, as represented by the counterfactual functions $\{C(x) : x \in \mathbb{R}\}$. In other words, [AOS Eq.16.6]

The way we **adjust for confounders** in regression is to regress Y on X **and** Z (where Z is the collection of confounding variables). More formally, define the conditional regression function $r(x, z)$:

Important theorem [AOS Theorem 16.6] (**adjusted treatment effect**):

Revisit Toy Example 2 again: