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.*

For years, causal inference was studied by statisticians, epidemiologists and economists. The machine learning community was largely uninterested. This has changed. The ML community now has an active research program in causation. This is because it is now recognized that many problems that were once treated as prediction problems are actually causal questions. Questions like: "If I place this ad on a web page, will people click on it?" and "If I recommend a product will people buy it?" are causal questions, not predictive questions.

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 θ that characterizes the causal influence of X on Y and find a way to estimate θ. 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.* **As we shall see, 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 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 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

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

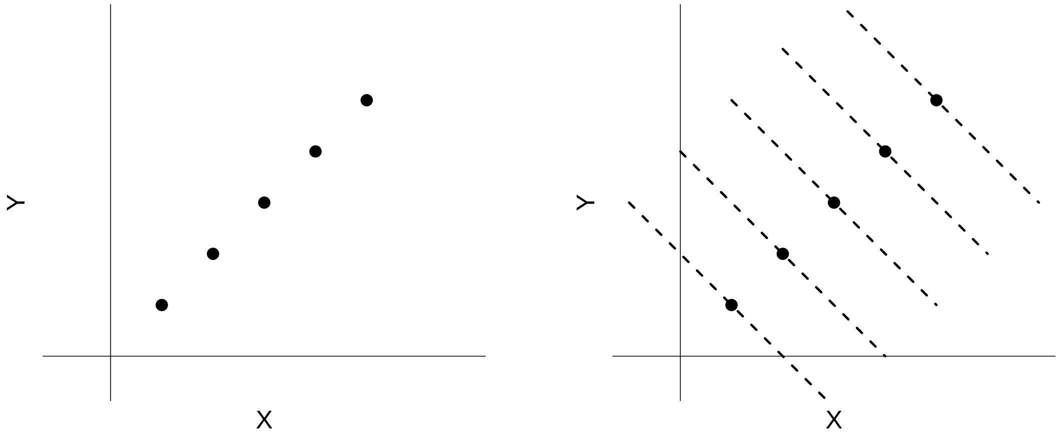

Figure 1: *Left: X and Y have positive association. Right: The lines are the counterfactuals, i.e. what would happen to each person if I changed their X value. Despite the positive association, the causal effect is negative. If we increase X everyone's Y values will decrease.*

More succinctly

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

We have replaced the random variables (X, Y) with the more detailed variables (X, Y_0, Y_1, Y) where $Y = XY_1 + (1 - X)Y_0$. When X is continuous, the counterfactual is a function $Y(\cdot)$. Then $Y(x)$ is value of the function $Y(\cdot)$ when $X = x$. The observed Y is $Y \equiv Y(X)$.

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 θ has the following interpretation: θ 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 θ? In general the answer is no. We can estimate

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

but α is not equal to θ. 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.

Let's formalize this. Let $\mathcal{P}$ be the set of distributions for (X, Y_0, Y_1, Y) such that $P(X = 0) > \delta$ and $P(X = 1) > \delta$ for some $\delta > 0$. (We have no hope if we do not have positive probability of observing exposed and unexposed subjects.) Recall that $Y = XY_1 + (1-X)Y_0$. The observed data are $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$. Let $\theta(P) = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$. An estimator is uniformly consistent if, for every $\epsilon > 0$,

$$\sup_{P \in \mathcal{P}} P(|\widehat{\theta}_n - \theta(P)| > \epsilon) \to 0$$

as $n \to \infty$.

Theorem 2 *In general, there does not exist a uniformly consistent estimator of θ.*

Proof. It is easy constrict $p(x, y_0, y_1)$ and and $q(x, y_0, y_1)$ such that $\theta(p) \neq \theta(q)$ and yet $p(x, y) = q(x, y)$. ■

In the case that X is continuous, the causal quantity (or rather, an example of a causal quantity) is

$$\theta(x) = \mathbb{E}[Y(x)]$$

which, in general, is NOT equal to $m(x) = \mathbb{E}[Y|X = x]$.

2.1 Two Ways to Make θ Estimable

Fortunately, there are two ways[1] to make θ 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:

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

Of course, we can't estimate θ if we always assign $X = 1$ or $X = 0$. We assume that $0 < \delta \leq P(X = 1) \leq 1 - \delta < 1$ for some δ. Let $\mathcal{P}$ be all such distributions.
Warning! Note that X is not independent of Y.

Theorem 3 *If X is randomly assigned, then $\theta = \alpha$ where*

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

A uniformly consistent estimator of α (and hence θ) is the plug-in estimator

$$\widehat{\alpha} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i} - \frac{\sum_{i=1}^n (1 - X_i) Y_i}{\sum_{i=1}^n (1 - X_i)}.$$

That is, for every $\epsilon > 0$,

$$\sup_{P \in \mathcal{P}} P(|\widehat{\alpha} - \theta| > \epsilon) \to 0$$

as $n \to \infty$.

Proof. Since X is independent of (Y_0, Y_1), we have

$$\begin{aligned} \alpha &= \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0) \\ &= \mathbb{E}(Y_1|X = 1) - \mathbb{E}(Y_0|X = 0) \quad \text{since } Y = XY_1 + (1 - X)Y_0 \\ &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \theta \quad \text{since } (Y_0, Y_1) \amalg X. \end{aligned}$$

Hence, random assignment makes θ equal to α. To prove the consistency of $\widehat{\alpha}$, note that we can write $\widehat{\alpha} = (A_n/B_n) - (C_n/D_n)$. Also note that

$$\alpha = \frac{\mathbb{E}[YX]}{\mathbb{E}[X]} - \frac{\mathbb{E}[Y(1 - X)]}{\mathbb{E}[1 - X]} \equiv \frac{A}{B} - \frac{C}{D}.$$

Let ϵ be a small positive constant. By Hoeffding's inequality and the union bound, with high probability, $A_n/B_n < (A + \epsilon)/(B - \epsilon) < (A/B) + \epsilon\Delta_1$ for some positive constant Δ_1.

[1] A third way is to use instrumental variables but we won't discuss that.

Similarly, $A_n/B_n > (A/B) - \epsilon\Delta_2$, say. A similar argument applies to the second term and the result follows. ■

Similarly, we can construct a test ϕ for testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ such that we have type I error control

$$\sup_{P \in \mathcal{P}_0} P(\phi = 1) \leq \alpha$$

and with non-trivial power: for any $\epsilon > 0$,

$$\inf_{P \in \mathcal{P}_\epsilon} P(\phi = 1) \to 1$$

where $\mathcal{P}_\epsilon$ is the set of distribution with $|\theta| \geq \epsilon$. We can also construct a confidence set (using Hoeffding's inequality or the CLT) such that

$$\inf_{P \in \mathcal{P}} P(\theta \in C) \geq 1 - \alpha.$$

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

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

In other words, under randomization, $\theta(x) = m(x)$ where $m(x) = \mathbb{E}(Y|X = x)$ is the uusal regression function. So you can use everything you know about regression estimation and then you are estimating the causal effect.

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, α 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 4 *Suppose that*

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

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 θ 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. It appears that the machine learning community has ignored this goal and has focused instead on the quixotic goal of causal discovery.

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 θ 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.

More generally,

$$\begin{aligned}\theta(x) = \mathbb{E}[Y(x)] = \mathbb{E}[Y(x)|Z=z]dP(z) &= \int \mathbb{E}[Y(x)|Z=z, X=x]dP(z)\\ = \int \mathbb{E}[Y|Z=z, X=x]dP(z) &= \int m(x,z)dP(z)\end{aligned}$$

where $m(x,z) = \mathbb{E}[Y|Z=z, X=x]$ is the usual regression function. We can insert an estimate $\widehat{m}$ and replace the integral over z eith an average:

$$\widehat{\theta}(x) = \frac{1}{n}\sum_i \widehat{m}(x, Z_i).$$

However, you should not use cross-validation to choose the smoothing parameter. You need to use methods known as *semi-parametric inference* to get an accurate estimate.

An alternative is to use *matching* which I will explain in class.

3 Causal Graphs and Structural Equations

Another way to capture the difference between $P(Y \in A|X=x)$ and $P(Y \in A|\mathsf{set}\ X=x)$ is to represent the distribution using a directed graph. Then we capture the second statement by performing certain operations on the graph. Specifically, we break the arrows into the some variables to represent an intervention.

A Directed Acyclic Graph (DAG) is a graph for a set of variables with no cycles. The graph defines a set of distributions of the form

$$p(y_1, \ldots, y_k) = \prod p(y_j|\mathsf{parents}(y_j)$$

where $\mathsf{parents}(y_j)$ are the parents of y_j. A **causal graph** is a DAG with extra information. A DAG is a causal graph if it correctly encodes the effect of setting a variable to a fixed value.

Consider the graph G in Figure 2. Here, X denotes treatment, Y is response and Z is a confounding variable. To find the causal distribution $p(y|\mathsf{set}\ X = x)$ we do the following steps:

1. Form a new graph G_* by removing all arrow into X. Now set X equal to x. This corresponds to replacing the joint distribution $p(x, y, z) = p(z)p(x|z)p(y|x, z)$ with the new distribution $p_*(y, z) = p(z)p(y|x, z)$. The factor $p(x|z)$ is removed because we know regard x as a fixed number. (Actually, $p(x|z)$ is replaced with a point mass at x.)
2. Compute the distribution of y from the new distribution:

$$p(y|\mathsf{set}\ X = x) \equiv p_*(y) = \int p_*(y, z)dz = \int p(z)p(y|x, z)dz.$$

Now we have that

$$p(y|\mathsf{set}\ X = 1) - p(y|\mathsf{set}\ X = 0) = \int p(y|1, z)p(z)dz - \int p(y|0, z)p(z)dz.$$

Hence,

$$\begin{aligned}
\theta &= \mathbb{E}[Y|\mathsf{set}\ X = 1] - \mathbb{E}[Y|\mathsf{set}\ X = 0] \\
&= \int yp(y|1, z)p(z)dz - \int yp(y|0, z)p(z)dz = \mathbb{E}[Y|X = 1, Z = z]p(z)dz - \mathbb{E}[Y|X = 0, Z = z]p(z)dz \\
&= \int \mu(1, z)p(z)dz - \int \mu(0, z)p(z)dz
\end{aligned}$$

This is precisely the same equation as (2). Both approaches lead to the same formulas for the causal effect. Of course, if there were unobserved confounding variables, then the formula for θ would involve these variables and the causal effect would be non-estimable (as before).

In a randomized experiment, there would be no arrow from Z to X. (That's the point of randomization). In that case the above calculations shows that $\theta = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$ which again agrees with the counterfactual approach.

In general, the DAG approach and the counterfactual approach lead to the same formulas for causal effects. They are two different languages for the same thing.

The formulas derived from a causal graph will only be correct if the causal graph is correct. Right now, we are assuming that the the correct causal structure is known to us, and is based on subject matter knowledge. For example, we know that rain cases wet lawns but wet lawns don't cause rain.

Example 5 *You may have noticed a correlation between rain and having a wet lawn, that is, the variable "Rain" is not independent of the variable "Wet Lawn" and hence* $p_{R,W}(r, w) \neq$

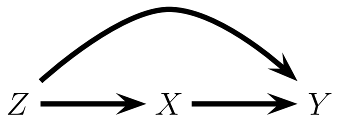

Figure 2: A basic causal graph. The arrows represent the effect of interventions. For example, the arrow from X to Y means that changing X effects the distribution of Y.

$p_R(r)p_W(w)$ *where* R *denotes Rain and* W *denotes Wet Lawn. Consider the following two DAGs:*

$$\text{Rain} \longrightarrow \text{Wet Lawn} \qquad \text{Rain} \longleftarrow \text{Wet Lawn.}$$

The first DAG implies that $p(w,r) = p(r)p(w|r)$ *while the second implies that* $p(w,r) = p(w)p(r|w)$ *No matter what the joint distribution* $p(w,r)$ *is, both graphs are correct. Both imply that* R *and* W *are not independent. But, intuitively, if we want a graph to indicate causation, the first graph is right and the second is wrong. Throwing water on your lawn doesn't cause rain. The reason we feel the first is correct while the second is wrong is because the interventions implied by the first graph are correct.*

Look at the first graph and form the intervention $W = 1$ *where 1 denotes "wet lawn." Following the rules of intervention, we break the arrows into* W *to get the modified graph:*

$$\text{Rain} \qquad \boxed{\textbf{set} \ \ \text{Wet Lawn} = \mathit{1}}$$

with distribution $p^*(r) = p(r)$. *Thus* $\mathbb{P}(R = r \mid W := w) = \mathbb{P}(R = r)$ *tells us that "wet lawn" does not cause rain.*

Suppose we (wrongly) assume that the second graph is the correct causal graph and form the intervention $W = 1$ *on the second graph. There are no arrows into* W *that need to be broken so the intervention graph is the same as the original graph. Thus* $p^*(r) = p(r|w)$ *which would imply that changing "wet" changes "rain." Clearly, this is nonsense.*

Both are correct probability graphs but only the first is correct causally. We know the correct causal graph by using background knowledge.

Causal graphs can also be represented by *structural equation models.* The graph in Figure 2 can be written as:

$$\begin{aligned} Z &= g_1(U) \\ X &= g_2(Z, V) \\ Y &= g_3(Z, X, W) \end{aligned}$$

for some functions g_1, g_2, g_3 and some random variables (U, V, W). Intervening on X corresponds to replacing the second equation with

$$X = x.$$

4 Causal Discovery Is Impossible

We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible. There are claims that it is possible but these claims are based on some unusual and not very convincing asymptotics. Specifically, there are claims that the graph can be discovered with some procedure and that the procedure is correct with probability tending to 1 as $n \to \infty$. But the asymptotic statement is non-standard: there is no finite sample size, however large, that can ever approximate the infinite limit.

What's worse, if we try to form a confidence interval for the size of the causal effect, then the confidence is is infinite no matter how large the sample is. This is Panglossian asymptotics. To understand what is going on, let's consider two examples.

Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ where X_i is the income of the subject's parents when the subject was a child, and Y_i is income of the subject at age 50. In this case, the variables are time ordered. So we can have X causing Y but we cannot have Y causing X. We must **always** allow for the fact that there may be many unobserved confounding variables. We will denote these by $U = (U_1, \ldots, U_k)$ where k is potentially very large. There are eight possible graphs as shown in Figure 3.[2] Our main interest is in whether there is an arrow from X to Y.

Let's see how the graph discovery community reasons in this case. Suppose we observe a large sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let α be some measure of dependence between X and Y. It is possible to define a consistent estimator $\widehat{\alpha}$. The causal discovery algorithms work as follows in this example. Suppose we find that there is a strong association between X and Y. (We can formally test for dependence between X and Y.) This is consistent with graphs 4,5,6,7 and 8. Some of these graphs include an arrow from X to Y and some don't. The conclusion is that we cannot tell if X causes Y. In this case, the causal discovery algorithms are correct.

Now suppose instead that we find that there is no significant association between X and Y. This is consistent with the first three graphs. None of these graphs include an arrow from X to Y. However, the last graph is also consistent with X being independent of Y. This might seem counterintuitive when you look at this graph. But the correlation created by the path $U \longrightarrow X \longrightarrow Y$ can cancel out the correlation created by the path $U \longrightarrow Y$. Such a cancellation is called *unfaithfulness*. Such a cancellation is considered to be unlikely. And the set $\mathcal{B}$ of such unfaithful distributions is "small." (For example, if the joint distribution is Normal, then the parameters that correspond to unfaithful distributions have measure 0.) So it seems reasonable to restrict ourselves to faithful distributions. If we restrict to faithful distributions, then the only explanation for the independence of X and Y is the first three graphs. We conclude that X does not cause Y.

Let me summarize the logic. There is a measure of dependence α and a consistence estimator $\widehat{\alpha}$. We are interested in the causal effect θ. We showed earlier that θ is a function $p(x, y, u)$. In particular, $\theta = 0$ means there is no arrow from X to Y and $\theta \neq 0$ means there

[2] Actually, we should have a separate node for each U_j. And then there are many more possible graphs.

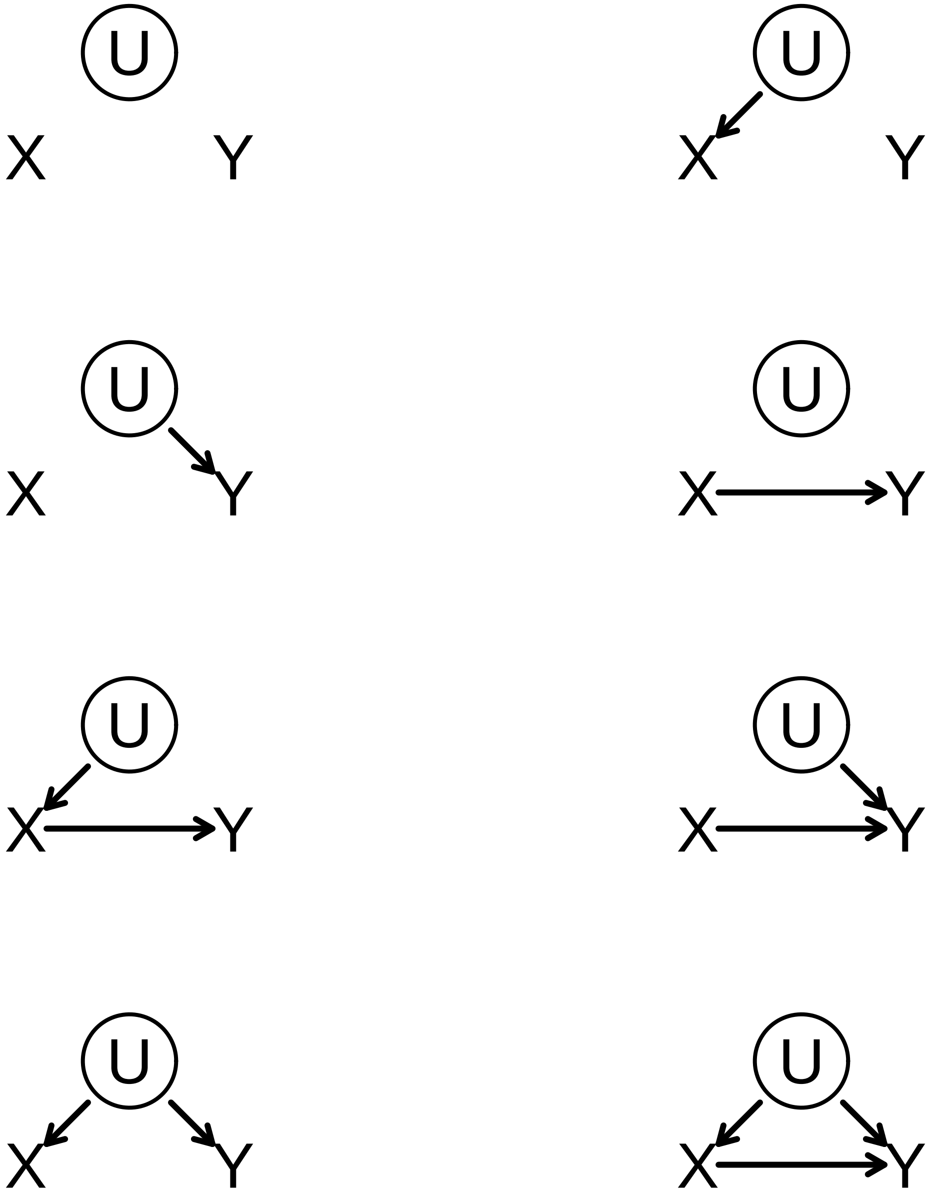

Figure 3: The eight possible causal graphs corresponding to the example.

is an arrow from X to Y. We have:

$$\begin{aligned} \alpha \neq 0 \quad &\Longrightarrow \quad \theta \text{ can be 0 } \text{ or nonzero (no conclusion)} \\ \alpha = 0 \text{ and faithfulness} \quad &\Longrightarrow \quad \theta = 0 \text{ (no causal effect).} \end{aligned}$$

Since $\widehat{\alpha}$ is a consistent estimator of α, we can substitute $\widehat{\alpha}$ for α and our conclusion is asymptotically correct. Note that if $P \in \mathcal{B}$, the relationship between α and θ breaks down. If $P \in \mathcal{B}$ then $\theta \neq 0$ but $\alpha = 0$.

Unfortunately, this reasoning is invalid. Let $\mathcal{P}$ be a set of distributions for (X, Y, U). Our model is

$$\mathcal{P}' = \mathcal{P} - \mathcal{B}$$

where $\mathcal{B}$ is the set of unfaithful distributions. The problem is that we can explain $\widehat{\alpha} \approx 0$ by graph 1 or by a P that is close to $\mathcal{B}$. We can always find a distribution P is that is faithful but arbitrarily close to unfaithful. We can never tell if $\widehat{\alpha} \approx 0$ is due to "no arrow from X to Y" or from P being very close to unfaithful. No matter how large n is, we can find a P that is so close to unfaithful that it could result in $\widehat{\alpha} \approx 0$.

By the way, keep in mind that U is very high dimensional. The set $\mathcal{B}$ might be "small" in some sense, but it is very complex. It is like a spider web.

To simplify matters, consider the linear case. The model for the DAG is

$$\begin{aligned} U &= \epsilon_1 \\ X &= aU + \epsilon_2 \\ Y &= bX + cU + \epsilon_3. \end{aligned}$$

Here, the ϵ_i's are mean 0 error terms. The causal effect is b. But all we observe is (X, Y). The correlation between X and Y is $\rho = a^2 + ac + b$. The problem is:

It is easy to construct cases where b is huge but $\rho \approx 0$. Ruling out the case when b is large and $\rho = 0$ (unfaithfulness) isn't enough but we can stll have b large and $\rho \approx 0$.

To make all this more precise, let $\psi = 1$ if there is an arrow from X to Y and let $\psi = 0$ if there is no arrow from X to Y. Let $\widehat{\psi}$ be the output of any causal discovery procedure (which can be set-valued). Suppose that $\widehat{\psi}$ is non-trivial, meaning that $1 \in \widehat{\psi}$ with increasing probability when $b \neq 0$. Let $\mathcal{P}_0$ be the set of faithful distributions with zero causal effect.

Theorem 6 *For any non-trivial procedure,*

$$\sup_{P \in \mathcal{P}_0} P(\widehat{\psi} = \psi) \to 1$$

as $n \to \infty$. In other words, if the procedure is non-trivil, we cannot control the type I error.

This result follows since there are infinitely many distributions in $\mathcal{P}'$ that are arbitrarily close to $\mathcal{B}$ and the procedure breaks down at $\mathcal{B}$. **The problem is that asymptotics have to be uniform over $\mathcal{P}$.** This is a point I have emphasized many times in this course. Uniformity is critical for sound statistical reasoning.

There is another way to see the problem. Consider the causal effect

$$\theta(x) = E[Y(x)] = \int \mathbb{E}[Y|X = x, U = u]p(u)du = \int m(x,u)p(u)du.$$

Discovering the graph involves implicitly estimating (or testing) $\theta(x)$. But it is clear that $\theta(x)$ is not estimable. It depends on $\mathbb{E}[Y|X = x, U = u]$ and $p(u)$. But we never observe U. **We can't estimate $m(x,u)$ if we don't observe u. Hence we can't estimate the causal effect.** We can't estimate parameters that are functions of unobserved random variables! The causal parameter is not identified. It is easy to show that the only valid confidence interval for $\theta(x)$ is the entire real line. In other words, if we want

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}'} P(\theta(x) \in C_n) \geq 1 - \alpha$$

then $C_n = \mathbb{R}$ with high probability. This shows that the causal effect cannot be estimated.

For yet another perspective, let us suppose that we model the whole distribution. The distribution is

$$p(u,x,y) = p(u)p(x|u)p(y|x,u) = p(u_1,\ldots,u_k)p(x|u_1,\ldots,u_k)p(y|x,u_1,\ldots,u_k).$$

The unknown parameters are the three functions $p(u_1,\ldots,u_k)$, $p(x|u_1,\ldots,u_k)$ and $p(y|x,u_1,\ldots,u_k)$. Suppose we take a chance and assume these distributions are Normal. We can then get the mle for the parameters and hence for $\theta(x)$. But again, we don't observe any $U's$. It's easy to see that the mle is for $\theta(x)$ is not defined. That is, every value of $\theta(x)$ is an mle.

To have reliable inference we need uniformly consistent estimates and we need valid confidence sets. There are no consistent estimators or valid confidence sets for causal parameters when there is unobserved confounding. The only solutions are: measure the confounders or do a randomized study.

Things get even worse when there are more than two variables. Let's consider another example. Suppose that we have the time ordered variables X, Y, Z. There are potential (unobserved) confounders U and V. See Figure 4. Again, the causal effects are not identified. There is nothing we can do here. But let's follow the causal discovery logic.

Suppose we observe a large sample and find that (i) X and Y are dependent, (ii) Y and Z are dependent and (iii) X and Z are conditionally independent given Y. I will explain in class how we can use the logic of causal discovery to conclude that:
(a) X causes Y
(b) Y causes Z
(c) there are no confounding variables in the Universe.

The last conclusion is astounding and should be be a hint that something is wrong.

Summary: Here is the bottom line:

1. In any real example based on observational data, we have to allow for the possibility that there are unobserved confounding variables.

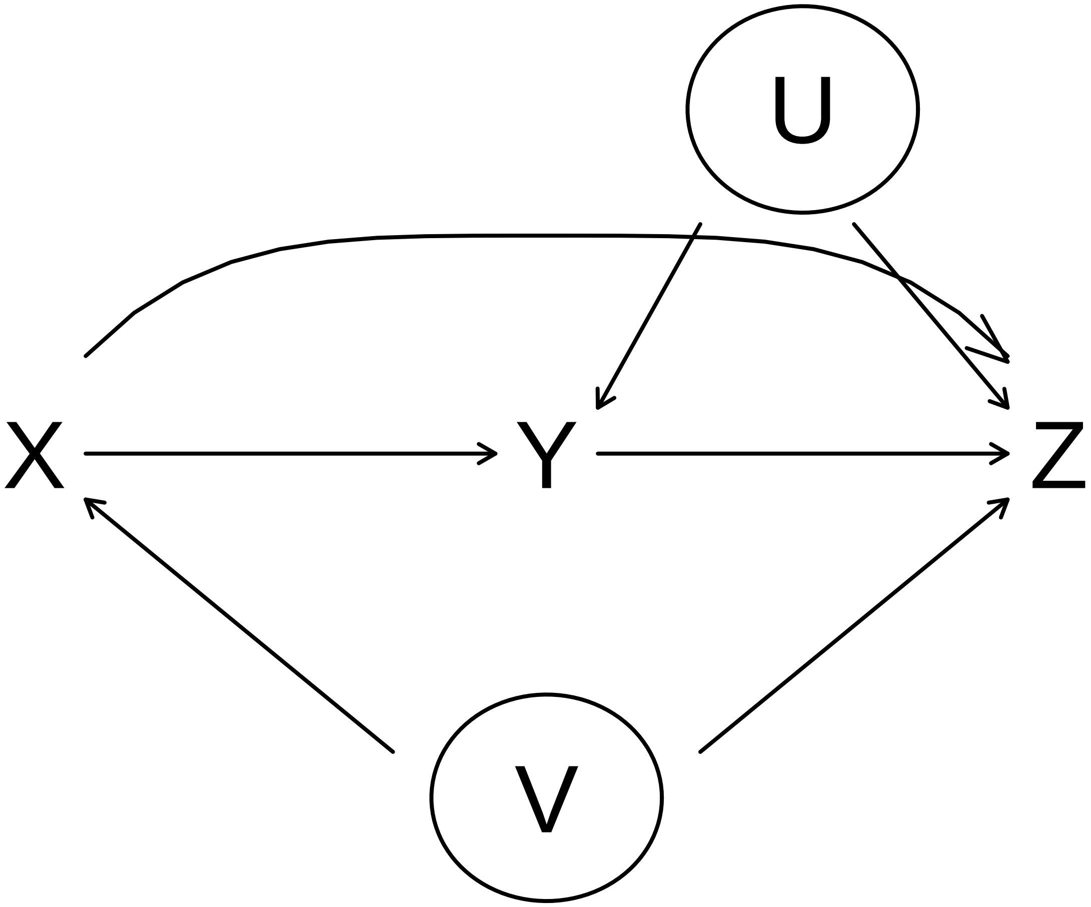

Figure 4: The full causal graph for the second example.

2. Causal quantities are functions of these unobserved variables.
3. It is impossible to estimate anything that is a function of unobserved variables.
4. Therefore, causal discovery is impossible.

Further Reading: A good tutorial with lost of good references is:

E. Kennedy (2015). Semiparametric Theory and Empirical Processes in Causal Inference. arXiv:1510.04740

Also, there is a very good, free book here:

`https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`

Appendix: More on Graphical Interventions

If you are having difficulty understanding the difference between $p(y|x)$ and $p(y|\text{set}\, x)$, then this section will provide additional explanation. It is helpful to consider two different computer programs. Consider the DAG in Figure 2. The probability function for a distribution consistent with this DAG has the form $p(x,y,z) = p(x)p(y|x)p(z|x,y)$. The following is pseudocode for generating from this distribution.

```
For $i = 1,\ldots,n$:
    $x_i$ <- $p_X(x_i)$
    $y_i$ <- $p_{Y|X}(y_i|x_i)$
    $z_i$ <- $p_{Z|X,Y}(z_i|x_i,y_i)$
```

Suppose we run this code, yielding data (x_1,y_1,z_1), $\ldots$, (x_n,y_n,z_n). Among all the times that we observe $Y=y$, how often is $Z=z$? The answer to this question is given by the conditional distribution of $Z|Y$. Specifically,

$$\begin{aligned}
\mathbb{P}(Z=z|Y=y) &= \frac{\mathbb{P}(Y=y,Z=z)}{\mathbb{P}(Y=y)} = \frac{p(y,z)}{p(y)} \\
&= \frac{\sum_x p(x,y,z)}{p(y)} = \frac{\sum_x p(x)\,p(y|x)\,p(z|x,y)}{p(y)} \\
&= \sum_x p(z|x,y)\frac{p(y|x)\,p(x)}{p(y)} = \sum_x p(z|x,y)\frac{p(x,y)}{p(y)} \\
&= \sum_x p(z|x,y)\,p(x|y).
\end{aligned}$$

Now suppose we **intervene** by changing the computer code. Specifically, suppose we fix Y at the value y. The code now looks like this:

```
set $Y = y$
for $i = 1, \dots, n$
    $x_i <- p_X(x_i)$
    $z_i <- p_{Z|X,Y}(z_i|x_i, y)$
```

Having **set** $Y = y$, how often was $Z = z$? To answer, note that the intervention has changed the joint probability to be

$$p^*(x, z) = p(x)p(z|x, y).$$

The answer to our question is given by the marginal distribution

$$p^*(z) = \sum_x p^*(x, z) = \sum_x p(x)p(z|x, y).$$

This is $p(z|\mathsf{set}\ Y = y)$.