Chapter 24

Identifying Causal Effects from Observations

There are two problems which are both known as "causal inference":

1. Given the causal structure of a system, estimate the effects the variables have on each other.
2. Given data about a system, find its causal structure.

The first problem is easier, so we'll begin with it.

24.1 Causal Effects, Interventions and Experiments

As a reminder, when I talk about the causal effect of X on Y, which I write

$$\Pr(Y|do(X=x)) \tag{24.1}$$

I mean the distribution of Y which would be generated, counterfactually, were X to be set to the particular value x. This is not, in general, the same as the ordinary conditional distribution

$$\Pr(Y|X=x) \tag{24.2}$$

The reason these are different is that the latter represents taking the original population, as it is, and just filtering it to get the sub-population where $X=x$. The processes which set X to that value may also have influenced Y through other channels, and so this distribution will not, typically, really tell us what would happen if we reached in and manipulated X. We can sum up the contrast in a little table (Table 24.1). As we saw in Chapter 22, if we have the full graph for a directed acyclic graphical model, it tells us how to calculate the joint distribution of all the variables, from which of course the conditional distribution of any one variable given another follows. As we saw in Chapter 23, calculations of $\Pr(Y|do(X=x))$ use a "surgically" altered graph, in which all arrows into X are removed, and its value is pinned at x,

Probabilistic conditioning	Causal conditioning
$\Pr(Y\|X=x)$	$\Pr(Y\|do(X=x))$
Factual	Counter-factual
Select a sub-population	Generate a new population
Predicts passive observation	Predicts active manipulation
Calculate from full DAG	Calculate from surgically-altered DAG
Always identifiable when X and Y are observable	Not always identifiable even when X and Y are observable

TABLE 24.1: *Contrasts between ordinary probabilistic conditioning and causal conditioning. (See below on identifiability.)*

but the rest of the graph is as before. If we know the DAG, and we know the distribution of each variable given its parents, we can calculate any causal effect we want, by graph-surgery.

24.1.1 The Special Role of Experiment

If we want to estimate $\Pr(Y|do(X=x))$, the most reliable procedure is also the simplest: actually manipulate X to the value x, and see what happens to Y. (As my mother says, “Why think, when you can just do the experiment?”) A causal or counter-factual assumption is still required here, which is that the *next* time we repeat the manipulation, the system will respond similarly, but this is pretty weak as such assumptions go.

While this seems like obvious common sense to us now, it is worth taking a moment to reflect on the fact that systematic experimentation is a very recent thing; it only goes back to around 1600. Since then, the knowledge we have acquired by combining experiments with mathematical theories have totally transformed human life, but for the first four or five thousand years of civilization, philosophers and sages much smarter than (almost?) any scientist now alive would have dismissed experiment as something fit only for cooks, potters and blacksmiths, who didn’t *really* know what they were doing.

The major obstacle the experimentalist must navigate around is to make sure they the experiment they are doing is the one they *think* they are doing. Symbolically, when we want to know $\Pr(Y|do(X=x))$, we need to make sure that we are *only* manipulating X, and not accidentally doing $\Pr(Y|do(X=x), Z=z)$ (because we are only experimenting on a sub-population), or $\Pr(Y|do(X=x, Z=z))$ (because we are also, inadvertently, manipulating Z). There are two big main divisions about how to avoid these confusions.

1. The older strategy is to *deliberately* control or manipulate as many other variables as possible. If we find $\Pr(Y|do(X=x, Z=z))$ and $\Pr\left(Y|do(X=x', Z=z)\right)$ then we know the differences between them are indeed just due to changing X. This strategy, of actually controlling or manipulating whatever we can, is the

traditional one in the physical sciences, and more or less goes back to Galileo and the beginning of the Scientific Revolution[1].

2. The younger strategy is to *randomize* over all the other variables but X. That is, to examine the contrast between $\Pr(Y|do(X=x))$ and $\Pr\left(Y|do(X=x')\right)$, we use an independent source of random noise to decide which experimental subjects will get $do(X=x)$ and which will get $do(X=x')$. It is easy to convince yourself that this makes $\Pr(Y|do(X=x))$ equal to $\Pr(Y|X=x)$. The great advantage of the randomization approach is that we can apply it even when we cannot actually control the other causally relevant variables, or even are unsure of what they are. Unsurprisingly, it has its origins in the biological sciences, especially agriculture. If we want to credit its invention to a single culture hero, it would not be too misleading[2] to attribute it to R. A. Fisher in the early 1900s.

Experimental evidence is compelling, but experiments are often slow, expensive, and difficult. Moreover, experimenting on people is hard, both because there are many experiments we *shouldn't* do, and because there are many experiments which would just be too hard to organize. We must therefore consider how to do causal inference from non-experimental, observational data.

24.2 Identification and Confounding

For the present purposes, the most important distinction between probabilistic and causal conditioning has to do with the **identification** (or **identifiability**), of the conditional distributions. An aspect of a statistical model is **identifiable** when it cannot be changed without there also being *some* change in the distribution of the observable variables. If we can alter part of a model with no observable consequences, that part of the model is **unidentifiable**[3]. Sometimes the lack of identification is trivial: in a two-cluster mixture model, we get the same observable distribution if we swap the labels of the two clusters (§21.1.5). The rotation problem for factor models (§§19.6, 19.10.1) is a less trivial identification problem[4]. If two variables are co-linear, then their coefficients in a linear regression are unidentifiable (§2.1.1)[5]. Note that identification is about the true distribution, not about what happens with finite data. A parameter might be identifiable, but we could have so little information about it in our data that our estimates are unusable, with immensely wide confidence intervals;

[1] The anguished sound you hear as you read this is every historian of science wailing in protest as the over-simplification, but this will do as an origin myth for our purposes.

[2] See previous note.

[3] More formally, divide the model's parameters into two parts, say θ and ψ. The distinction between θ_1 and θ_2 is identifiable if, for all ψ_1, ψ_2, the distribution over observables coming from (θ_1, ψ_1) is different from that coming from (θ_2, ψ_2). If the right choice of ψ_1 and ψ_2 masks the distinction between θ_1 and θ_2, then θ is unidentifiable.

[4] As this example suggests, what is identifiable depends on what is observed. If we could observe the factors directly, factor loadings would be identifiable.

[5] As that example suggests, whether one aspect of a model is identifiable or not can depend on other aspects of the model. If the co-linearity was broken, the two regression coefficients would become identifiable.

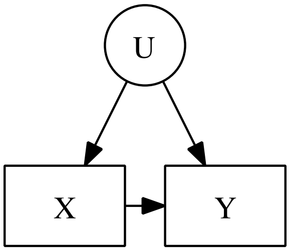

FIGURE 24.1: *The distribution of Y given X,* $\Pr(Y|X)$*,* **confounds** *the actual causal effect of X on Y,* $\Pr(Y|do(X=x))$*, with the indirect dependence between X and Y created by their unobserved common cause U. (You may imagine that U is really more than one variable, with some internal sub-graph.)*

that's unfortunate, but we just need more data. An unidentifiable parameter, however, cannot be estimated even with infinite data.[6]

When X and Y are both observable variables, $\Pr(Y|X=x)$ can't help being identifiable. (Changing this conditional distribution just *is* changing part of the distribution of observables.) Things are very different, however, for $\Pr(Y|do(X=x))$. In some models, it's entirely possible to change this drastically, and always have the same distribution of observables, by making compensating changes to other parts of the model. When this is the case, we simply cannot estimate causal effects from observational data. The basic problem is illustrated in Figure 24.1.

In Figure 24.1, X is a parent of Y. But if we analyze the dependence of Y on X, say in the form of the conditional distribution $\Pr(Y|X=x)$, we see that there are two channels by which information flows from cause to effect. One is the direct, causal path, represented by $\Pr(Y|do(X=x))$. The other is the indirect path, where X gives information about its parent U, and U gives information about its child Y. If we just observe X and Y, we cannot separate the causal effect from the indirect inference. The causal effect is **confounded** with the indirect inference. More generally, the effect of X on Y is confounded whenever $\Pr(Y|do(X=x)) \neq \Pr(Y|X=x)$. If there is some way to write $\Pr(Y|do(X=x))$ in terms of distributions of observables, we say that the confounding can be removed by an **identification strategy**, which **de-confounds** the effect. If there is no way to de-confound, then this causal effect is unidentifiable.

The effect of X on Y in Figure 24.1 is unidentifiable. Even if we erased the arrow from X to Y, we could get any joint distribution for X and Y we liked by picking

[6] For more on identifiability, and what to do with unidentifiable problems, see the great book by Manski (2007).

$P(X|U)$, $P(Y|U)$ and $P(U)$ appropriately. So we cannot even, in this situation, use observations to tell whether X is actually a cause of Y. Notice, however, that even if U was observed, it would still not be the case that $\Pr(Y|X=x)=\Pr(Y|do(X=x))$. While the effect would be identifiable (via the back door criterion; see below), we would still need some sort of adjustment to recover it.

In the next section, we will look at such identification strategies and adjustments.

24.3 Identification Strategies

To recap, we want to calculate the causal effect of X on Y, $\Pr(Y|do(X=x))$, but we cannot do an experiment, and must rely on observations. In addition to X and Y, there will generally be some **covariates** Z which we know, and we'll assume we know the causal graph, which is a DAG. Is this enough to determine $\Pr(Y|do(X=x))$? That is, does the joint distribution **identify** the causal effect?

The answer is "yes" when the covariates Z contain all the other relevant variables[7]. The inferential problem is then no worse than any other statistical estimation problem. In fact, if we know the causal graph and get to observe all the variables, then we could (in principle) just use our favorite non-parametric conditional density estimate at each node in the graph, with its parent variables as the inputs and its own variable as the response. Multiplying conditional distributions together gives the whole distribution of the graph, and we can get any causal effects we want by surgery. Equivalently (Exercise 2), we have that

$$\Pr(Y|do(X=x)) = \sum_{t} \Pr(Y|X=x, \mathrm{Pa}(X)=t)\Pr(\mathrm{Pa}(X)=t) \tag{24.3}$$

where $\mathrm{Pa}(X)$ is the complete set of parents of X.

If we're willing to assume more, we can get away with just using non-parametric regression or even just an additive model at each node. Assuming yet more, we could use parametric models at each node; the linear-Gaussian assumption is (alas) very popular.

If some variables are *not* observed, then the issue of which causal effects are observationally identifiable is considerably trickier. Apparently subtle changes in which variables are available to us and used can have profound consequences.

The basic principle underlying all considerations is that we would like to condition on adequate **control** variables, which will block paths linking X and Y *other than* those which would exist in the surgically-altered graph where all paths into X have been removed. If other unblocked paths exist, then there is some confounding of the causal effect of X on Y with their mutual dependence on other variables.

This is familiar to use from regression as the basic idea behind using additional variables in our regression, where the idea is that by introducing covariates, we "control for" other effects, until the regression coefficient for our favorite variable represents only its causal effect. Leaving aside the inadequacies of linear regression as such (Chapter 2), we need to be cautious here. Just conditioning on everything possible does *not* give us adequate control, or even necessarily bring us closer to it. As Figure 24.2 illustrates, and as several of the data-analysis problem sets will drive home, *adding* an ill-chosen covariate to a regression can create confounding.

[7] This condition is sometimes known as **causal sufficiency**. Strictly speaking, we do not have to suppose that *all* causes are included in the model and observable. What we have to assume is that all of the remaining causes have such an unsystematic relationship to the ones included in the DAG that they can be modeled as noise. (This does not mean that the noise is necessarily small.) In fact, what we really have to assume is that the relationships between the causes omitted from the DAG and those included is so intricate and convoluted that it might as well be noise, along the lines of algorithmic information theory (Li and Vitányi, 1997), whose key result might be summed up as "Any determinism distinguishable from randomness is insufficiently complex". But here we verge on philosophy.

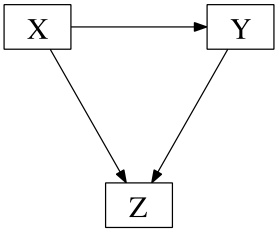

FIGURE 24.2: *"Controlling for" additional variables can introduce bias into estimates of causal effects. Here the effect of X on Y is directly identifiable,* $\Pr(Y|do(X=x)) = \Pr(Y|X=x)$. *If we also condition on Z however, because it is a common* effect *of X and Y, we'd get* $\Pr(Y|X=x, Z=z) \neq \Pr(Y|X=x)$. *In fact, even if there were no arrow from X to Y, conditioning on Z would make Y depend on X.*

There are three main ways we can find adequate controls, and so get both identifiability and appropriate adjustments:

1. We can condition on an intelligently-chosen set of covariates S, which block all the indirect paths from X to Y, but leave all the direct paths open. (That is, we can follow the regression strategy, but do it right.) To see whether a candidate set of controls S is adequate, we apply the **back-door criterion**.
2. We can find a set of variables M which **mediate** the causal influence of X on Y — all of the direct paths from X to Y pass through M. If we can identify the effect of M on Y, and of X on M, then we can combine these to get the effect of X on Y. (That is, we can just study the *mechanisms* by which X influences Y.) The test for whether we can do this combination is the **front-door criterion**.
3. We can find a variable I which affects X, and which *only* affects Y by influencing X. If we can identify the effect of I on Y, and of I on X, then we can, sometimes, "factor" them to get the effect of X on Y. (That is, I gives us variation in X which is independent of the common causes of X and Y.) I is then an **instrumental variable** for the effect of X on Y.

Let's look at these three in turn.

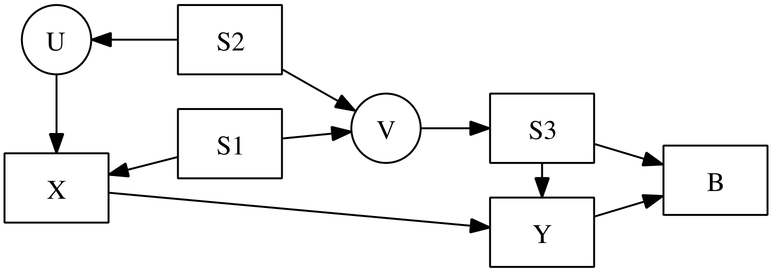

FIGURE 24.3: *Illustration of the back-door criterion for identifying the causal effect of X on Y. Setting $S = \{S_1, S_2\}$ satisfies the criterion, but neither S_1 nor S_2 on their own would. Setting $S = \{S_3\}$, or $S = \{S_1, S_2, S_3\}$ also works. Adding B to any of the good sets makes them fail the criterion.*

24.3.1 The Back-Door Criterion: Identification by Conditioning

When estimating the effect of X on Y, a **back-door path** is an undirected path between X and Y with an arrow *into* X. These are the paths which create confounding, by providing an indirect, non-causal channel along which information can flow. A set of conditioning variables or controls S satisfies the **back-door criterion** when (i) S blocks every back-door path between X and Y, and (ii) no node in S is a descendant of X. (Cf. Figure 24.3.) When S meets the back-door criterion,

$$\Pr(Y|do(X = x)) = \sum_s \Pr(Y|X = x, S = s)\Pr(S = s) \tag{24.4}$$

Notice that all the items on the right-hand side are observational conditional probabilities, not counterfactuals. Thus we have achieved identifiability, as well as having an adjustment strategy.

The motive for (i) is plain, but what about (ii)? We don't want to include descendants of X which are also ancestors of Y, because that blocks off some of the causal paths from X to Y, and we don't want to include descendants of X which are also descendants of Y, because they provide non-causal information about Y[8].

More formally, we can proceed as follows (Pearl, 2009b, §11.3.3). We know from Eq. 24.3 that

$$\Pr(Y|do(X = x)) = \sum_t \Pr(\mathrm{Pa}(X) = t)\Pr(Y|X = x, \mathrm{Pa}(X) = t) \tag{24.5}$$

[8] What about descendants of X which are neither ancestors nor descendants of Y? Conditioning on them is either creates potential colliders, if they are also descended from ancestors of Y other than X, or needlessly complicates the adjustment in Eq. 24.4.

Now suppose we can always introduce another set of conditioned variables, if we sum out over them:

$$\Pr(Y|do(X=x)) = \sum_t \Pr(\mathrm{Pa}(X)=t) \sum_s \Pr(Y, S=s|X=x, \mathrm{Pa}(X)=t) \tag{24.6}$$

We can do this for *any* set of variables S, it's just probability. It's also just probability that

$$\begin{aligned} &\Pr(Y, S|X=x, \mathrm{Pa}(X)=t) = \\ &\quad \Pr(Y|X=x, \mathrm{Pa}(X)=t, S=s)\Pr(S=s|X=x, \mathrm{Pa}(X)=t) \end{aligned} \tag{24.7}$$

so

$$\begin{aligned} &\Pr(Y|do(X=x)) = \\ &\quad \sum_t \Pr(\mathrm{Pa}(X)=t) \sum_s \Pr(Y|X=x, \mathrm{Pa}(X)=t, S=s)\Pr(S=s|X=x, \mathrm{Pa}(X)=t) \end{aligned} \tag{24.8}$$

Now we invoke the fact that S satisfies the back-door criterion. Point (i) of the criterion, blocking back-door paths, implies that $Y \perp\!\!\!\perp \mathrm{Pa}(X)|X, S$. Thus

$$\begin{aligned} &\Pr(Y|do(X=x)) = \\ &\quad \sum_t \Pr(\mathrm{Pa}(X)=t) \sum_s \Pr(Y|X=x, S=s)\Pr(S=s|X=x, \mathrm{Pa}(X)=t) \end{aligned} \tag{24.9}$$

Point (ii) of the criterion, not containing descendants of X, means (by the Markov property) that $X \perp\!\!\!\perp S|\mathrm{Pa}(X)$. Therefore

$$\begin{aligned} &\Pr(Y|do(X=x)) = \\ &\quad \sum_t \Pr(\mathrm{Pa}(X)=t) \sum_s \Pr(Y|X=x, S=s)\Pr(S=s|\mathrm{Pa}(X)=t) \end{aligned} \tag{24.10}$$

Since $\sum_t \Pr(\mathrm{Pa}(X)=t)\Pr(S=s|\mathrm{Pa}(X)=t) = \Pr(S=s)$, we have, at last,

$$\Pr(Y|do(X=x)) = \sum_s \Pr(Y|X=x, S=s)\Pr(S=s) \tag{24.11}$$

as promised. □

24.3.1.1 The Entner Rules

Using the back-door criterion requires us to know the causal graph. Recently, Entner *et al.* (2013) have given a simple set of rules which provide *sufficient* conditions for deciding that set of variables satisfy the back-door criterion, or that X actually has no effect on Y, which can be used without knowing the graph completely.

It makes no sense to control for anything which is a descendant of either Y or X; that's either blocking a directed path or activating a collider. So let $\mathcal{W}$ be the set of all observed variables which descend neither from X nor Y.

1. If there is a set of controls S such that $X \perp\!\!\!\perp Y | S$, then X has no causal effect on Y.

   Reasoning: Y can't be a child of X if we can make them independent by conditioning on anything, and Y can't be a more remote descendant either, since S doesn't include any descendants of X. So in this situation all the paths linking X to Y must be back-door paths, and S, blocking them, shows there's no effect.

2. If there is a $W \in \mathcal{W}$ and a subset S of the $\mathcal{W}$, not including W, such that (i) $W \not\!\perp\!\!\!\perp Y | S$, but (ii) $W \perp\!\!\!\perp Y | S, X$, then X has an effect on Y, and S satisfies the back-door criterion for estimating the effect.

   Reasoning: Point (i) shows that conditioning on S leaves open path from W to Y. By point (ii), these paths must all pass through X, since conditioning on X blocks them, hence X has an effect on Y. S must block all the back-door paths between X and Y, otherwise X would be a collider on paths between W and Y, so conditioning on X would activate those paths.

3. If there is a $W \in \mathcal{W}$ and a subset S of $\mathcal{W}$, excluding W, such that (i) $W \not\!\perp\!\!\!\perp X | S$ but (ii) $W \perp\!\!\!\perp Y | S$, then X has no effect on Y.

   Reasoning: Point (i) shows that conditioning on S leaves open active paths from W to X. But by (ii), there cannot be any open paths from W to Y, so there cannot be any open paths from X to Y.

If none of these rules apply, whether X has an effect on Y, and if so what adequate controls are for finding it, will depend on the exact graph, and *cannot* be determined just from independence relations among the observables. (For proofs of everything, see the paper.)

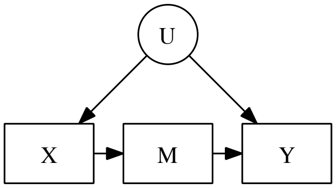

FIGURE 24.4: *Illustration of the front-door criterion, after Pearl (2009b, Figure 3.5). X, Y and M are all observed, but U is an unobserved common cause of both X and Y. $X \leftarrow U \rightarrow Y$ is a back-door path confounding the effect of X on Y with their common cause. However, all of the effect of X on Y is mediated through X's effect on M. M's effect on Y is, in turn, confounded by the back-door path $M \leftarrow X \leftarrow U \rightarrow Y$, but X blocks this path. So we can use back-door adjustment to find* $\Pr(Y|do(M=m))$, *and directly find* $\Pr(M|do(X=x)) = \Pr(M|X=x)$. *Putting these together gives* $\Pr(Y|do(X=x))$.

24.3.2 The Front-Door Criterion: Identification by Mechanisms

A set of variables M satisfies the **front-door criterion** when (i) M blocks all directed paths from X to Y, (ii) there are no unblocked back-door paths from X to M, and (iii) X blocks all back-door paths from M to Y. Then

$$\Pr(Y|do(X=x)) = \sum_{m}\Pr(M=m|X=x)\sum_{x'}\Pr\left(Y|X=x',M=m\right)\Pr\left(X=x'\right) \quad (24.12)$$

A natural reaction to the front-door criterion is "Say what?", but it becomes more comprehensible if we take it apart. Because, by clause (i), M blocks all *directed* paths from X to Y, any *causal* dependence of Y on X must be mediated by a dependence of Y on M:

$$\Pr(Y|do(X=x)) = \sum_{m}\Pr(Y|do(M=m))\Pr(M=m|do(X=x)) \quad (24.13)$$

Clause (ii) says that we can get the effect of X on M directly,

$$\Pr(M=m|do(X=x)) = \Pr(M=m|X=x)\ . \quad (24.14)$$

Clause (iii) say that X satisfies the back-door criterion for identifying the effect of M on Y, and the inner sum in Eq. 24.12 is just the back-door computation (Eq. 24.4) of $\Pr(Y|do(M=m))$. So really we *are* using the back door criterion, twice. (See Figure 24.4.)

For example, in the "does tooth-brushing prevent heart-disease?" example of §23.2.2, we have $X =$ "frequency of tooth-brushing", $Y =$ "heart disease", and we could take as the mediating M either "gum disease" or "inflammatory immune response", according to Figure 23.2.

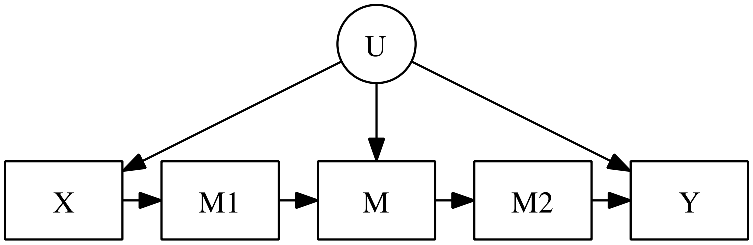

FIGURE 24.5: *The path $X \to M \to Y$ contains all the mechanisms by which X influences Y, but is not isolated from the rest of the system (U → M). The sub-mechanisms $X \to M_1 \to M$ and $M \to M_2 \to Y$* are *isolated, and the original causal effect can be identified by composing them.*

24.3.2.1 The Front-Door Criterion and Mechanistic Explanation

Morgan and Winship (2007, ch. 8) give a useful insight into the front-door criterion. Each directed path from X to Y is, or can be thought of as, a separate **mechanism** by which X influences Y. The requirement that all such paths be blocked by M, (i), is the requirement that the set of mechanisms included in M be "exhaustive". The two back-door conditions, (ii) and (iii), require that the mechanisms be "isolated", not interfered with by the rest of the data-generating process (at least once we condition on X). Once we identify an isolated and exhaustive set of mechanisms, we know all the ways in which X actually affects Y, and any indirect paths can be discounted, using the front-door adjustment 24.12.

One interesting possibility suggested by this is to elaborate mechanisms into sub-mechanisms, which could be used in some cases where the plain front-door criterion won't apply[9], such as Figure 24.5. Because U is a parent of M, we cannot use the front-door criterion to identify the effect of X on Y. (Clause (i) holds, but (ii) and (iii) both fail.) But we can use M_1 and the front-door criterion to find $\Pr(M|do(X=x))$, and we can use M_2 to find $\Pr(Y|do(M=m))$. Chaining those together, as in Eq. 24.13, would given $\Pr(Y|do(X=x))$. So even though the whole mechanism from X to Y is not isolated, we can still identify effects by breaking it into sub-mechanisms which *are* isolated. This suggests a natural point at which to stop refining our account of the mechanism into sub-sub-sub- mechanisms: when we can identify the causal effects we're concerned with.

[9] The ideas in this paragraph come from conversation Prof. Winship; see Morgan and Winship (2015, ch. 10).

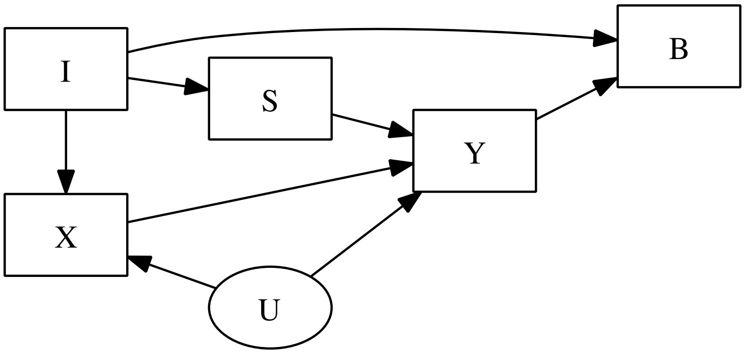

FIGURE 24.6: *A valid instrumental variable, I, is related to the cause of interest, X, and influences Y only through its influence on X, at least once control variables block other paths. Here, to use I as an instrument, we* should *condition on S, but* should not *condition on B. (If we could condition on U, we would not need to use an instrument.)*

24.3.3 Instrumental Variables

A variable I is an **instrument**[10] for identifying the effect of X on Y when there is a set of controls S such that (i) $I \not\perp\!\!\!\perp X|S$, and (ii) every unblocked path from I to Y has an arrow pointing into to X. Another way to say (ii) is that $I \perp\!\!\!\perp Y|S, do(X)$. Colloquially, I influences Y, but only through first influencing X (at least once we control for S). (See Figure 24.6.)

How is this useful? By making back-door adjustments for S, we can identify $\Pr(Y|do(I=i))$ and $\Pr(X|do(I=i))$. Since all the causal influence of I on Y must be channeled through X (by point (ii)), we have

$$\Pr(Y|do(I=i)) = \sum_x \Pr(Y|do(X=x))\Pr(X=x|do(I=i)) \tag{24.15}$$

as in Eq. 24.3. We can thus identify the causal effect of X on Y whenever Eq. 24.15 can be solved for $\Pr(Y|do(X=x))$ in terms of $\Pr(Y|do(I=i))$ and $\Pr(X|do(I=i))$. Figuring out when this is possible in general requires an excursion into the theory of integral equations[11], which is beyond the scope of this class; the upshot is that, in gen-

[10] The term "instrumental variables" comes from econometrics, where they were originally used, in the 1940s, to identify parameters in simultaneous equation models. (The metaphor was that I is a measuring instrument for the otherwise inaccessible parameters.) Definitions of instrumental variables are surprisingly murky and controversial outside of extremely simple linear systems; this one is taken from Galles and Pearl (1997), via Pearl (2009b, §7.4.5).

[11] If X is continuous, then the analog of Eq. 24.15 is $\Pr(Y|do(I=i)) = \int p(Y|do(X=x))p(X=x|do(I=i))dx$, where the "integral operator" $\int \cdot p(X=x|do(I=i))dx$ is known, as is $\Pr(Y|do(I=i))$.

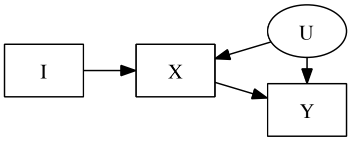

FIGURE 24.7: *I acts as an instrument for estimating the effect of X on Y, despite the presence of the confounding, unobserved variable U.*

eral, there are no solutions. However, in the special case where the relations between all variables are linear, we can do better.

Let's start with the most basic possible set-up for an instrumental variable, namely that in Figure 24.7, where we just have X, Y, the instrument I, and the unobserved confounders S. If everything is linear, identifying the causal effect of X on Y is equivalent to identifying the coefficient on the $X \rightarrow Y$ arrow. We can write

$$X = \alpha_0 + \alpha I + \delta U + \epsilon_X \tag{24.16}$$

and

$$Y = \beta_0 + \beta X + \gamma U + \epsilon_Y \tag{24.17}$$

where ϵ_X and ϵ_Y are mean-zero noise terms, independent of each other and of the other variables, and we can, without loss of generality, assume U has mean zero as well. We want to find β. Substituting,

$$Y = \beta_0 + \beta\alpha_0 + \beta\alpha I + (\beta\delta + \gamma)U + \beta\epsilon_X + \epsilon_Y \tag{24.18}$$

Since U, ϵ_X and ϵ_Y are all unobserved, we can re-write this as

$$Y = \gamma_0 + \beta\alpha I + \eta \tag{24.19}$$

where $\eta = (\beta\delta + \gamma)U + \beta\epsilon_X + \epsilon_Y$ has mean zero.

Now take the covariances:

$$\begin{aligned}
\mathrm{Cov}[I,X] &= \alpha\mathrm{Var}[I] + \mathrm{Cov}[\epsilon_X, I] && (24.20)\\
\mathrm{Cov}[I,Y] &= \beta\alpha\mathrm{Var}[I] + \mathrm{Cov}[\eta, I] && (24.21)\\
&= \beta\alpha\mathrm{Var}[I] + (\beta\delta + \gamma)\mathrm{Cov}[U,I] && (24.22)\\
&\quad + \beta\mathrm{Cov}[\epsilon_X, I] + \mathrm{Cov}[\epsilon_Y, I]
\end{aligned}$$

By condition (ii), however, we must have $\mathrm{Cov}[U,I] = 0$, and of course $\mathrm{Cov}[\epsilon_X, I] = \mathrm{Cov}[\epsilon_Y, I] = 0$. Therefore $\mathrm{Cov}[I,Y] = \beta\alpha\mathrm{Var}[I]$. Solving,

$$\beta = \frac{\mathrm{Cov}[I,Y]}{\mathrm{Cov}[I,X]} \tag{24.23}$$

This can be estimated by substituting in the sample covariances, or any other consistent estimators of these two covariances.

On the other hand, the (true or population-level) coefficient for linearly regressing Y on X is

$$\begin{aligned}
\frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]} &= \frac{\beta\mathrm{Var}[X]+\gamma\mathrm{Cov}[U,X]}{\mathrm{Var}[X]} && (24.24)\\
&= \beta+\gamma\frac{\mathrm{Cov}[U,X]}{\mathrm{Var}[X]} && (24.25)\\
&= \beta+\gamma\frac{\delta\mathrm{Var}[U]}{\alpha^2\mathrm{Var}[I]+\delta^2\mathrm{Var}[U]+\mathrm{Var}[\epsilon_X]} && (24.26)
\end{aligned}$$

That is, "OLS is biased for the causal effect when X is correlated with the noise". In other words, simple regression is misleading in the presence of confounding[12].

The instrumental variable I provides a source of variation in X which is uncorrelated with the other common ancestors of X and Y. By seeing how both X and Y respond to these perturbations, and using the fact that I only influences Y through X, we can deduce something about how X influences Y, though linearity is very important to our ability to do so.

The simple line of reasoning above runs into trouble if we have multiple instruments, or need to include controls (as the definition of an instrument allows). In §25.2 we'll look at the more complicated estimation methods which can handle this, again assuming linearity.

24.3.3.1 Some Invalid Instruments

Not everything which looks like an instrument actually works. If Y is indeed a descendant of I, but there is a line of descent that doesn't go through X, then I is not a valid instrument for X (Figure 24.8). If there are unblocked back-door paths linking I and Y — if I and Y have common ancestors, for instance — then I is not a valid instrument (Figure 24.9).

Economists sometimes refer to both sets of problems with instruments as "violations of exclusion restrictions". The second sort of problem, in particular, is a "failure of exogeneity".

24.3.3.2 Critique of Instrumental Variables

By this point, you may well be thinking that instrumental variable estimation is very much like using the front-door criterion. There, the extra variable M came between X and Y; here, X comes between I and Y. It is, perhaps, surprising (if not annoying) that using an instrument only lets us identify causal effects under extra assumptions, but that's life. Just as the front-door criterion relies on using our scientific knowledge, or rather theories, to find isolated and exhaustive mechanisms, finding valid

[12] But observe that if we want to make a linear prediction of Y and only have X available, i.e., to find the best r_1 in $\mathbf{E}[Y|X=x]=r_0+r_1x$, then Eq. 24.26 is *exactly* the coefficient we would want to use. OLS is doing its job.

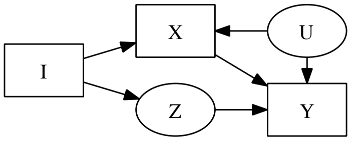

FIGURE 24.8: *I is not a valid instrument for identifying the effect of X on Y, because I can influence Y through a path not going through X. If we could control for Z, however, I would become valid.*

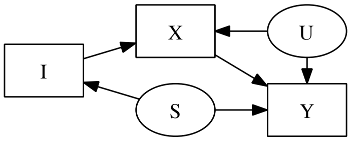

FIGURE 24.9: *I is not a valid instrument for identifying the effect of X on Y, because there is an unblocked back-door path connecting I and Y. If we could control for S, however, I would become valid.*

instruments relies on theories about the world (or the part of it under study), and one would want to try to check those theories.

In fact, instrumental variable estimates of causal effects are often presented as more or less unquestionable, and free of theoretical assumptions; economists, and other social scientists influenced by them, are especially apt to do this. As the economist Daniel Davies puts it[13], devotees of this approach

> have a really bad habit of saying:
> "Whichever way you look at the numbers, X".
> when all they can really justify is:
> "Whichever way **I** look at the numbers, X".
> but in fact, I should have said that they could only really support:
> "Whichever way **I** look at **these** numbers, X".

(Emphasis in the original.) It will not surprise you to learn that I think this is very wrong.

I hope that, after four months of nonlinear models, if someone tries to sell you a linear regression, you should be very skeptical, but let's leave that to one side. (It's not *impossible* that everything really is linear.) The clue that instrumental variable estimation is a creature of theoretical assumptions is point (ii) in the definition of an instrument: $I \perp\!\!\!\perp Y | S, do(X)$. This says that if we eliminate all the arrows into X, the control variables S block all the other paths between I and Y. This is *exactly* as much an assertion about mechanisms as what we have to do with the front-door criterion. In fact it doesn't just say that every mechanism by which I influences Y is mediated by X, it also says that there are no common causes of I and Y (other than those blocked by S).

This assumption is most easily defended when I is genuinely random, For instance, if we do a randomized experiment, I might be a coin-toss which assigns each subject to be in either the treatment or control group, each with a different value of X. If "compliance" is not perfect (if some of those in the treatment group don't actually get the treatment, or some in the control group do), it is nonetheless plausible that the only route by which I influences the outcome is through X, so an instrumental variable regression is appropriate. (I here is sometimes called "intent to treat".)

Even here, we must be careful. If we are evaluating a new medicine, whether people *think* they are getting a medicine or not could change how they act, and medical outcomes. Knowing whether they were assigned to the treatment or the control group would thus create another path from I to Y, not going through X. This is why randomized clinical trials are generally "double-blinded" (neither patients nor medical personnel know who is in the control group); but how effective the double-blinding is itself a theoretical assumption.

More generally, any argument that a candidate instrument is valid is really an argument that other channels of influence, apart from the favored one through X, can be ruled out. This generally cannot be done through analyzing the same variables

[13] In part four of his epic and insightful review of *Freakonomics*; see `http://d-squareddigest.blogspot.com/2007/09/freakiology-yes-folks-its-part-4-of.html`.

used in the instrumental-variable estimation (see below), but involves some theory about the world, and rests on the strength of the evidence for that theory. As has been pointed out multiple times — for instance, by Rosenzweig and Wolpin (2000), and by Deaton (2010) — the theories needed to support instrumental variable estimates in particular concrete cases are often *not* very well-supported, and plausible rival theories can produce very different conclusions from the same data.

Many people have thought that one *can* test for the validity of an instrument, by looking at whether $I \perp\!\!\!\perp Y|X$ — the idea being that, if influence flows from I through X to Y, conditioning on X should block the channel. The problem is that, in the instrumental-variable set-up, X is a collider, so conditioning on X actually creates an indirect dependence *even if* I is valid. So $I \not\!\perp\!\!\!\perp Y|X$, whether or not the instrument is valid, and the test (even if performed perfectly with infinite data) tells us nothing[14].

A final, more or less technical, issue with instrumental variable estimation is that many instruments are (even if valid) **weak** — they only have a little influence on X, and a small covariance with it. This means that the denominator in Eq. 24.23 is a number close to zero. Error in estimating the denominator, then, results in a much larger error in estimating the ratio. Weak instruments lead to noisy and imprecise estimates of causal effects. It is not hard to construct scenarios where, at reasonable sample sizes, one is actually better off using the biased OLS estimate than the unbiased but high-variance instrumental estimate.

[14] However, see Pearl (2009b, §8.4) for a different approach which can "screen out very bad would-be instruments".

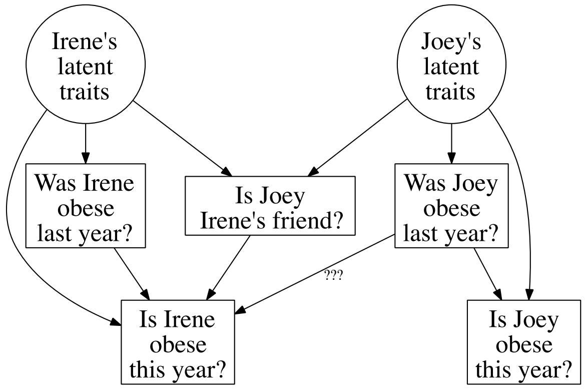

FIGURE 24.10: *Social influence is confounded with selecting friends with similar traits, unobserved in the data.*

24.3.4 Failures of Identification

The back-door and front-door criteria, and instrumental variables, are all *sufficient* for estimating causal effects from probabilistic distributions, but are not *necessary*. A necessary condition for *un*-identifiability is the presence of an unblockable back-door path from X to Y. However, this is not sufficient for lack of identification — we might, for instance, be able to use the front door criterion, as in Figure 24.4. There are necessary and sufficient conditions for the identifiability of causal effects in terms of the graph, and so for un-identifiability, but they are rather complex and I will not go over them (see Shpitser and Pearl (2008), and Pearl (2009b, §§3.4–3.5) for an overview).

As an example of the unidentifiable case, consider Figure 24.10. This DAG depicts the situation analyzed in Christakis and Fowler (2007), a famous paper claiming to show that obesity is contagious in social networks (at least in the suburb of Boston where the data was collected). At each observation, participants in the study get their weight taken, and so their obesity status is known over time. They also provide the name of a friend. This friend is often in the study. Christakis and Fowler were interested in the possibility that obesity is contagious, perhaps through some process of behavioral influence. If this is so, then Irene's obesity status in year 2 should depend on Joey's obesity status in year one, but *only* if Irene and Joey are friends — not if they are just random, unconnected people. It is indeed the case that if Joey becomes obese, this predicts a substantial increase in the odds of Joey's friend Irene becoming obese, even controlling for Irene's previous history of obesity[15].

The difficulty arises from the latent variables for Irene and Joey (the round nodes

[15] The actual analysis was a bit more convoluted than that, but this is the general idea.

in Figure 24.10). These include all the traits of either person which (a) influence who they become friends with, and (b) influence whether or not they become obese. A very partial list of these would include: taste for recreational exercise, opportunity for recreational exercise, taste for alcohol, ability to consume alcohol, tastes in food, occupation and how physically demanding it is, ethnic background[16], etc. Put simply, if Irene and Joey are friends because they spend two hours in the same bar every day drinking and eating chicken wings with ranch dressing, it's less surprising that both of them have an elevated chance of becoming obese, and likewise if they became friends because they both belong to the decathlete's club, they are both unusually unlikely to become obese. Irene's status is predictable from Joey's, then, not (or not just) because Joey influences Irene, but because seeing what kind of person Irene's friends are tells us about what kind of person Irene is. It is not too hard to convince oneself that there is just no way, in this DAG, to get at the causal effect of Joey's behavior on Irene's that isn't confounded with their latent traits (Shalizi and Thomas, 2011). To de-confound, we would need to actual measure those latent traits, which may not be impossible but is certainly was not done here[17].

When identification is not possible — when we can't de-confound — it may still be possible to *bound* causal effects. That is, even if we can't say exactly that $\Pr(Y|do(X=x))$ must be, we can still say it has to fall within a certain (non-trivial!) range of possibilities. The development of bounds for non-identifiable quantities, what's sometimes called **partial identification**, is an active area of research, which I think is very likely to become more and more important in data analysis; the best introduction I know is Manski (2007).

[16] Friendships often run within ethnic communities. On the one hand, this means that friends tend to be more *genetically* similar than random members of the same town, so they will be usually apt to share genes which influence susceptibility to obesity (in that environment). On the other hand, ethnic communities transmit, non-genetically, traditions regarding food, alcohol, sports, exercise, etc., and (again non-genetically: Tilly (1998)) influence employment and housing opportunities.

[17] Of course, the issue is not really about obesity. Studies of "viral marketing", and of social influence more broadly, all generically have the same problem. Predicting someone's behavior from that of their friend means conditioning on the existence of a social tie between them, but that social tie is a collider, and activating the collider creates confounding.

24.4 Summary

Of the four techniques I have introduced, instrumental variables are clever, but fragile and over-sold[18]. Experimentation is ideal, but often unavailable. The back-door and front-door criteria are, I think, the best observational approaches, when they can be made to work.

Often, nothing can be made to work. Many interesting causal effects are just not identifiable from observational data. More exactly, they only become identifiable under very strong modeling assumptions, typically ones which cannot be tested from the same data, and sometimes ones which cannot be tested by any sort of empirical data whatsoever. Sometimes, we have good reasons (from other parts of our scientific knowledge) to make such assumptions. Sometimes, we make such assumptions because we have a pressing need for *some* basis on which to act, and a wrong guess is better than nothing[19]. If you do make such assumptions, you need to make clear that you are doing so, and what they are; explain your reasons for making those assumptions, and not others[20]; and indicate how different your conclusions could be if you made different assumptions.

24.4.1 Further Reading

My presentation of the three major criteria is heavily indebted to Morgan and Winship (2007), but I hope not a complete rip-off. Pearl (2009b) is also essential reading on this topic. Berk (2004) provides an excellent critique of naive (that is, overwhelmingly common) uses of regression for estimating causal effects.

Most econometrics texts devote considerable space to instrumental variables. Didelez *et al.* (2010) is a very good discussion of instrumental variable methods, with less-standard applications. There is some work on non-parametric versions of instrumental variables (e.g., Newey and Powell 2003), but the form of the models must be restricted or they are unidentifiable. On the limitations of instrumental variables, Rosenzweig and Wolpin (2000) and Deaton (2010) are particularly recommended; the latter reviews the issue in connection with important recent work in development economics and the alleviation of extreme poverty, an area where statistical estimates really do matter.

There is a large literature in the philosophy of science and in methodology on the notion of "mechanisms". References I have found useful include, in general, Salmon (1984), and, specifically on social processes, Elster (1989), Hedström and Swedberg (1998) (especially Boudon 1998), Hedström (2005), Tilly (1984, 2008), and DeLanda (2006).

[18] I confess that I would probably not be so down on them if others did not push them up so excessively.

[19] As I once heard a distinguished public health expert put it, "This problem is too important to worry about getting it right."

[20] "My boss/textbook says so" and "so I can estimate β" are not good reasons

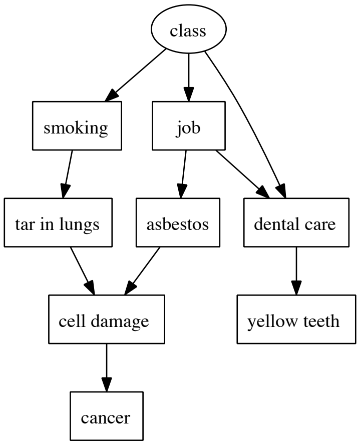

FIGURE 24.11: *DAG for Exercise 3.*

24.5 Exercises

1. Draw a graphical model representing the situation where a causal variable X is set at random. Verify that $\Pr(Y|X=x)$ is then equal to $\Pr(Y|do(X=x))$. (*Hint:* Use the back door criterion.)

2. Prove Eq. 24.3, by using the causal Markov property of the appropriate surgically-altered graph.

3. Refer to Figure 24.11. Can we use the front door criterion to estimate the effect of occupational prestige on cancer? If so, give a set S of variables that we would adjust for in the front-door method. Is there more than one such set? If so, can you find them all? Are there variables we could add to this set (or sets) which would violate the front-door criterion?

4. Read Salmon (1984). When does his "statistical relevance basis" provide enough information to identify causal effects?