Part III

Causal Inference

19

Graphical Causal Models

19.1 Causation and Counterfactuals

Take a piece of cotton, say an old rag. Apply flame to it; the cotton burns. We say the fire *caused* the cotton to burn. The flame is certainly *correlated* with the cotton burning, but, as we all know, correlation is not causation (Figure 19.1). Perhaps every time we set rags on fire we handle them with heavy protective gloves; the gloves don't make the cotton burn, but the statistical dependence is strong. So what is causation?

We do not have to settle 2500 years (or more) of argument among philosophers and scientists. For our purposes, it's enough to realize that the concept has a **counter-factual** component: if, contrary to fact, the flame had not been applied to the rag, then the rag would not have burned[1]. On the other hand, the fire makes the cotton burn whether we are wearing protective gloves or not.

To say it a somewhat different way, the distributions we observe in the world

[1] If you immediately start thinking about quibbles, like "What if we hadn't applied the flame, but the rag was struck by lightning?", then you may have what it takes to be a philosopher.

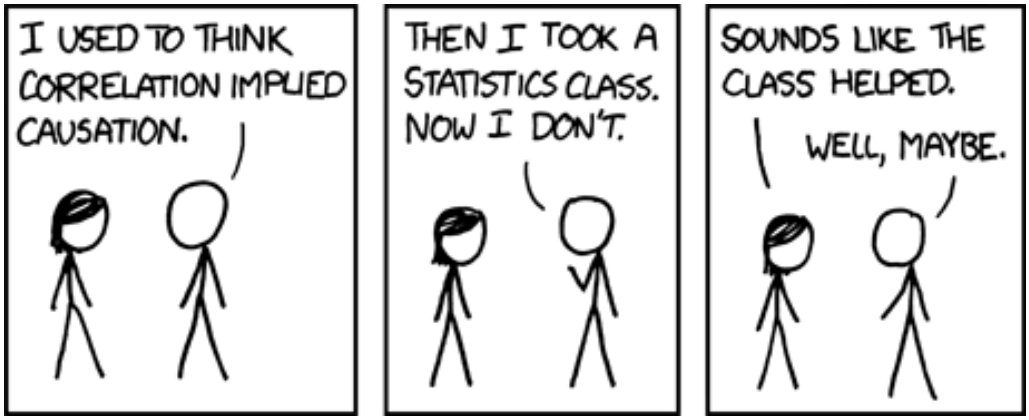


Figure 19.1 "Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there"' (Image and text copyright by Randall Munroe, used here under a Creative Commons attribution-noncommercial license; see http://xkcd.com/552/. [[TODO: Excise from the commercial version]])

are the outcome of complicated stochastic processes. The mechanisms which set the value of one variable inter-lock with those which set other variables. When we make a probabilistic prediction by conditioning — whether we predict $\mathbb{E}[Y \mid X = x]$ or $\Pr(Y \mid X = x)$ or something more complicated — we are just filtering the output of those mechanisms, picking out the cases where they happen to have set X to the value x, and looking at what goes along with that.

When we make a *causal* prediction, we want to know what would happen if the usual mechanisms controlling X were suspended and it was *set* to x. How would this change propagate to the other variables? What distribution would result for Y? This is often, perhaps even usually, what people really want to know from a data analysis, and they settle for statistical prediction either because they think it *is* causal prediction, or for lack of a better alternative.

Causal inference is the undertaking of trying to answer causal questions from empirical data. Its fundamental difficulty is that we are trying to derive counterfactual conclusions with only factual premises. As a matter of habit, we come to expect cotton to burn when we apply flames. We might even say, on the basis of purely statistical evidence, that the world has this habit. But as a matter of pure logic, no amount of evidence about what *did* happen can compel beliefs about what *would* have happened under non-existent circumstances[2]. (For all my *data* shows, all the rags I burn just so happened to be on the verge of spontaneously bursting into flames anyway.) We must supply some counter-factual or causal premise, linking what we see to what we could have seen, to derive causal conclusions.

One of our goals, then, in causal inference will be to make the causal premises as weak and general as possible, thus limiting what we take on faith.

19.2 Causal Graphical Models

We will need a formalism for representing causal relations. It will not surprise you by now to learn that these will be graphical models. We will in fact use DAG models from last time, with "parent" interpreted to mean "directly causes". These will be **causal graphical models**, or **graphical causal models**.[3]

We make the following assumptions.

1. There is some directed acyclic graph G representing the relations of causation among the our variables.

[2] The first person to really recognize this seems to have been the medieval Muslim theologian and anti-philosopher al Ghazali (1100/1997). (See Kogan (1985) for some of the history.) Very similar arguments were made centuries later by Hume (1739); whether there was some line of intellectual descent linking them — that is, any causal connection — I don't know.

[3] Because DAG models have joint distributions which factor according to the graph, we can always write them in the form of a set of equations, as $X_i = f_i(X_{\text{parents}(i)}) + \epsilon_i$, with the catch that the noise ϵ_i is not necessarily independent of X_i's parents. This is what is known, in many of the social sciences, as a **structural equation model**. So those are, strictly, a sub-class of DAG models. They are also often used to represent causal structure.

2. **The Causal Markov condition:** The joint distribution of the variables obeys the Markov property on G.
3. **Faithfulness:** The joint distribution has all of the conditional independence relations implied by the causal Markov property, and *only* those conditional independence relations.

The point of the faithfulness condition is to rule out "conspiracies among the parameters", where, say, two causes of a common effect, which would typically be dependent conditional on that effect, have their impact on the joint effect and their own distributions matched just so exactly that they remain conditionally independent.

19.2.1 Calculating the "effects of causes"

Let's fix two sub-sets of variables in the graph, X_c and X_e. (Assume they don't overlap, and call everything else X_N.) If we want to make a *probabilistic* prediction for X_e's value when X_c takes a particular value, x_c, that's the conditional distribution, $\Pr(X_e \mid X_c = x_c)$, and we saw last time how to calculate that using the graph. Conceptually, this amounts to selecting, out of the whole population or ensemble, the sub-population or sub-ensemble where $X_c = x_c$, and accepting whatever other behavior may go along with that.

Now suppose we want to ask what the effect would be, causally, of setting X_c to a particular value x_c. We represent this by "doing surgery on the graph": we (i) eliminate any arrows coming in to nodes in X_c, (ii) fix their values to x_c, and (iii) calculate the resulting distribution for X_e in the new graph. By steps (i) and (ii), we imagine suspending or switching off the mechanisms which ordinarily set X_c. The other mechanisms in the assemblage are left alone, however, and so step (iii) propagates the fixed values of X_c through them. We are not *selecting* a sub-population, but producing a new one.

If setting X_c to different values, say x_c and x_c', leads to different distributions for X_e, then we say that X_c **has an effect** on X_e — or, slightly redundantly, **has a causal effect** on X_e. Sometimes[4] "the effect of switching from x_c to x_c'" specifically refers to a change in the expected value of X_e, but since profoundly different distributions can have the same mean, this seems needlessly restrictive.[5] If one is interested in average effects of this sort, they are computed by the same procedure.

It is convenient to have a short-hand notation for this procedure of causal conditioning. One more-or-less standard idea, introduced by Judea Pearl, is to introduce a *do* operator which encloses the conditioning variable and its value. That is,

$$\Pr(X_e \mid X_c = x_c) \tag{19.1}$$

[4] Especially in economics.

[5] Economists are also fond of the horribly misleading usage of talking about "an X effect" or "the effect of X" when they mean the regression coefficient of X. Don't do this.

is probabilistic conditioning, or selecting a sub-ensemble from the old mechanisms; but

$$\Pr(X_e \mid do(X_c = x_c)) \tag{19.2}$$

is causal conditioning, or producing a new ensemble. Sometimes one sees this written as $\Pr(X_e \mid X_c \hat{=} x_c)$, or even $\Pr(X_e \mid \widehat{x_c})$. I am actually fond of the *do* notation and will use it.

Suppose that $\Pr(X_e \mid X_c = x_c) = \Pr(X_e \mid do(X_c = x_c))$. This would be extremely convenient for causal inference. The conditional distribution on the right is the causal, counter-factual distribution which tells us what would happen if x_c was imposed. The distribution on the left is the ordinary probabilistic distribution we have spent years learning how to estimate from data. When do they coincide?

One situation where they coincide is when X_c contains all the parents of X_e, and none of its descendants. Then, by the Markov property, X_e is independent of all other variables given X_c, and removing the arrows *into* X_c will not change that, or the conditional distribution of X_e given its parents. Doing causal inference for other choices of X_c will demand other conditional independence relations implied by the Markov property. This is the subject of Chapter 20.

19.2.2 Back to Teeth

Let us return to the example of Figure 18.6, and consider the relationship between exposure to asbestos and the staining of teeth. In the model depicted by that figure, the joint distribution factors as

$$\begin{aligned} & p(\text{Yellow teeth}, \text{Smoking}, \text{Asbestos}, \text{Tar in lungs}, \text{Cancer}) \\ &= p(\text{Smoking})p(\text{Asbestos}) \\ &\quad \times p(\text{Tar in lungs}|\text{Smoking}) \\ &\quad \times p(\text{Yellow teeth}|\text{Smoking}) \\ &\quad \times p(\text{Cancer}|\text{Asbestos}, \text{Tar in lungs}) \end{aligned} \tag{19.3}$$

As we saw, whether or not someone's teeth are yellow (in this model) is unconditionally independent of asbestos exposure, but conditionally *dependent* on asbestos, given whether or not they have cancer. A logistic regression of tooth color on asbestos would show a non-zero coefficient, after "controlling for" cancer. This coefficient would become significant with enough data. The usual interpretation of this coefficient would be to say that the log-odds of yellow teeth increase by so much for each one unit increase in exposure to asbestos, "other variables being held equal".[6] But to see the actual causal effect of increasing exposure to asbestos by one unit, we'd want to compare $p(\text{Yellow teeth}|do(\text{Asbestos} = a))$ to $p(\text{Yellow teeth}|do(\text{Asbestos} = a + 1))$, and it's easy to check (Exercise 19.1) that

[6] Nothing hinges on this being a logistic regression, similar interpretations are given to all the other standard models.

these two distributions have to be the same. In this case, because asbestos is exogenous, one will in fact get the same result for $p(\text{Yellow teeth}|do(\text{Asbestos} = a)$ and for $p(\text{Yellow teeth}|\text{Asbestos} = a)$.

For a more substantial example, consider Figure 19.2.[7] The question of interest here is whether regular brushing and flossing actually prevents heart disease. The mechanism by which it might do so is as follows: brushing is known to make it less likely for people to get gum disease. Gum disease, in turn, means the gums suffer from constant, low-level inflammation. Persistent inflammation (which can be measured through various messenger chemicals of the immune system) is thought to increase the risk of heart disease. Against this, people who are generally health-conscious are likely to brush regularly, and to take other actions, like regularly exercising and controlling their diets, which also make them less likely to get heart disease. In this case, if we were to manipulate whether people brush their teeth[8], we would shift the graph from Figure 19.2 to Figure 19.3, and we would have

$$p(\text{Heart disease}|\text{Brushing} = b) \neq p(\text{Heart disease}|do(\text{Brushing} = b)) \quad (19.4)$$

19.3 Conditional Independence and d-Separation Revisited

We saw in §18.3 that all distributions which conform to a common DAG share a common set of conditional independence relations. Faithful distributions have *no other* conditional independence relations. These are vital facts for causal inference.

The reason is that while *causal influence* flows one way through the graph, along the directions of arrows from parents to children, *statistical information* can flow in either direction. We can certainly make inferences about an effect from its causes, but we can equally make inferences about causes from their effects. It might be harder to actually do the calculations[9], and we might be left with more uncertainty, but we could do it. As we saw in §18.3, when conditioning on a set of variables S blocks all channels of information flow between X and Y, $X \perp\!\!\!\perp Y|S$. The faithful distributions are the ones where this implication is reversed, where $X \perp\!\!\!\perp Y|S$ implies that S blocks all paths between X and Y. In faithful graphical models, blocking information flow is exactly the same as conditional independence.

This turns out to be the single most important fact enabling causal inference. If we want to estimate the effects of causes, within a given DAG, we need to block off all non-causal channels of information flow. If we want to check whether a given DAG is correct for the variables we have, we need to be able to compare

[7] Based on de Oliveira *et al.* (2010), and the discussion of this paper by Chris Blattman (`http://chrisblattman.com/2010/06/01/does-brushing-your-teeth-lower-cardiovascular-disease/`).

[8] Hopefully, by ensuring that everyone brushes, rather than keeping people from brushing.

[9] Janzing (2007) [[TODO: update refs]] makes the very interesting suggestion that the direction of causality can be discovered by using this — roughly speaking, that if $X|Y$ is much harder to compute than is $Y|X$, we should presume that $X \to Y$ rather than the other way around.

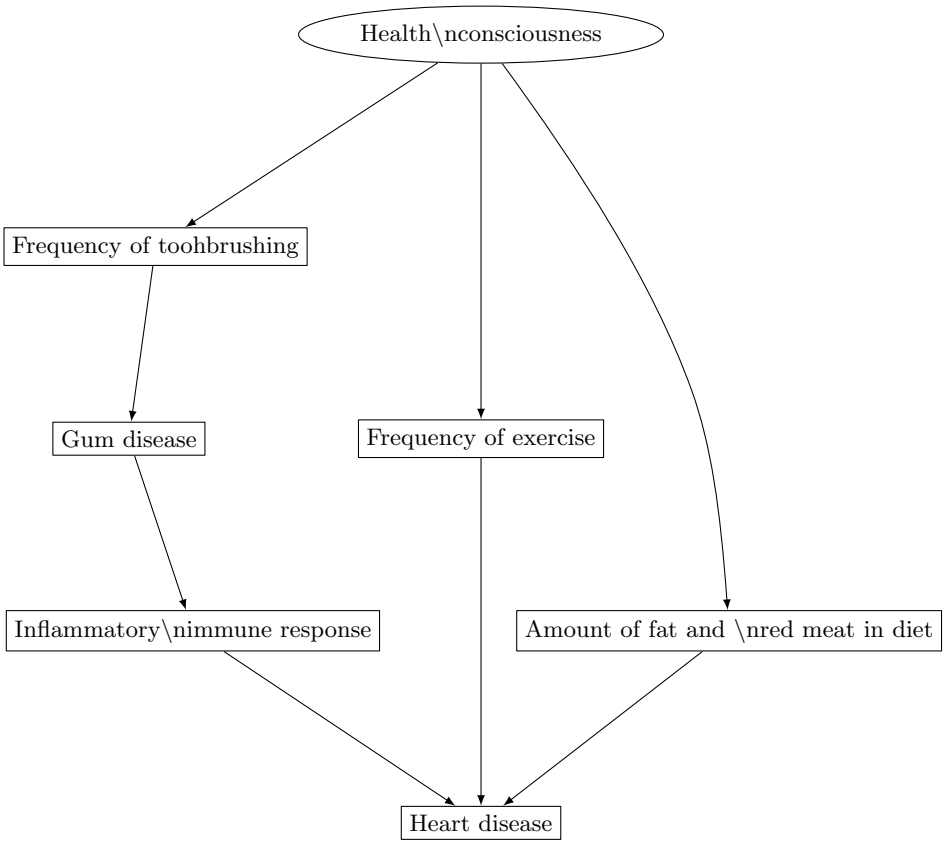


Figure 19.2 Graphical model illustrating hypothetical pathways linking brushing your teeth to not getting heart disease.

the conditional independence relations implied by the DAG to those supported by the data. If we want to discover the possible causal structures, we have to see which ones imply the conditional independencies supported by the data.

19.4 Further Reading

The two foundational books on graphical causal models are Spirtes *et al.* (2001) and Pearl (2009b). Both are excellent and recommended in the strongest possible terms; but if you had to read just one, I would recommend Spirtes *et al.* (2001). If on the other hand you do not feel up to reading a book at all, then Pearl (2009a) is much shorter, and covers the high points. (Also, it's free online.) The textbook by Morgan and Winship (2007, 2015) is much less demanding mathematically, and therefore also less complete conceptually, but it does explain the crucial ideas clearly, simply, and with abundant examples.[10] Lauritzen (1996) has

[10] That textbook also discusses an alternative formalism for counterfactuals, due mainly to Donald B. Rubin and collaborators. While Rubin has done very distinguished work in causal inference, his

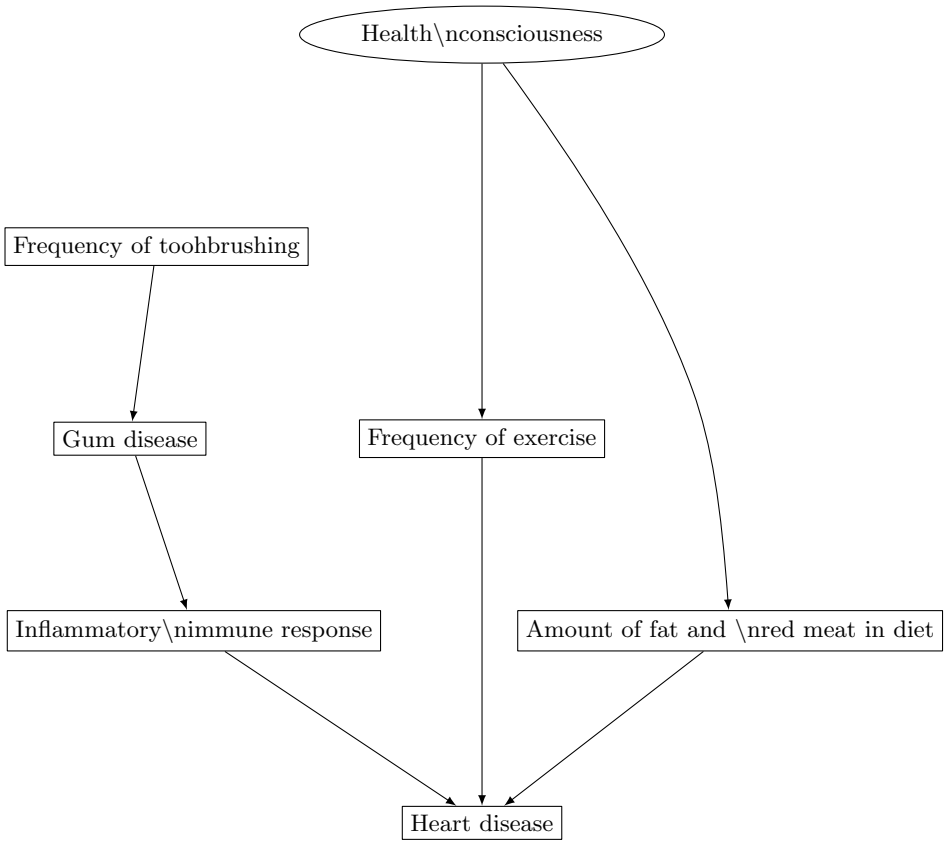


Figure 19.3 The previous graphical model, "surgically" altered to reflect a manipulation (*do*) of brushing.

a mathematically rigorous treatment of d-separation (among many other things), but de-emphasizes causality.

Many software packages for linear structural equation models and path analysis offer options to search for models; these are not, in general, reliable (Spirtes *et al.*, 2001).

Raginsky (2011) provides a fascinating information-theoretic account of graphical causal models and *do*(), in terms of the notion of directed (rather than mutual) information.

formalism is vastly harder to manipulate than are graphical models, but has no more expressive power. (Pearl (2009a) has a convincing discussion of this point, and Richardson and Robins (2013) provides a comprehensive proof that the everything expressible in the counterfactuals formalism can also be expressed with suitably-augmented graphical models.) I have thus skipped the Rubin formalism here, but there are good accounts in Morgan and Winship (2007, ch. 2), in Rubin's collected papers (Rubin, 2006), and in Imbens and Rubin (2015) (though please read Shalizi 2016 before taking any of the real-data examples in the last of these as models to imitate).

Exercises

19.1 Show, for the graphical model in Figure 18.6, that $p(\text{Yellow teeth}|do(\text{Asbestos} = a))$ is always the same as $p(\text{Yellow teeth}|do(\text{Asbestos} = a + 1))$.

20

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; we come back to the second in Chapter 22.

20.1 Causal Effects, Interventions and Experiments

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

$$\Pr(Y|do(X = x)) \tag{20.1}$$

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

$$\Pr(Y|X = x) \tag{20.2}$$

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 20.1). As we saw in Chapter 18, 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 19, 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, 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.

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 20.1 *Contrasts between ordinary probabilistic conditioning and causal conditioning. (See below on identifiability.)*

20.1.1 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 used to say, "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 $\Pr(Y|do(X=x', Z=z))$ 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 $\Pr(Y|do(X=x'))$,

[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.

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.

20.2 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 (§17.1.5). The rotation problem for factor models (§§16.5, 16.9.1) 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; 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

[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.

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

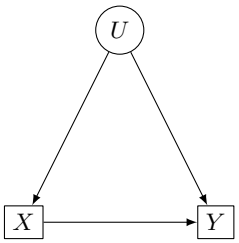


Figure 20.1 The distribution of Y given X, $\Pr(Y|X)$, **confounds** the actual causal effect of X on Y, $\Pr(Y|do(X = x))$, with the indirect dependence between X and Y created by their unobserved common cause U. (You may imagine that U is really more than one variable, with some internal sub-graph.)

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 20.1.

In Figure 20.1, 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 **deconfounds** the effect. If there is no way to de-confound, then this causal effect is unidentifiable.

The effect of X on Y in Figure 20.1 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 $P(X|U)$, $P(Y|U)$ 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, §20.3.1), we would still need some sort of adjustment to recover it.

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

20.3 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 20.2), we have that

$$\Pr(Y|do(X=x)) = \sum_t \Pr(Y|X=x, \mathrm{Pa}(X)=t) \Pr(\mathrm{Pa}(X)=t) \tag{20.3}$$

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

[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.

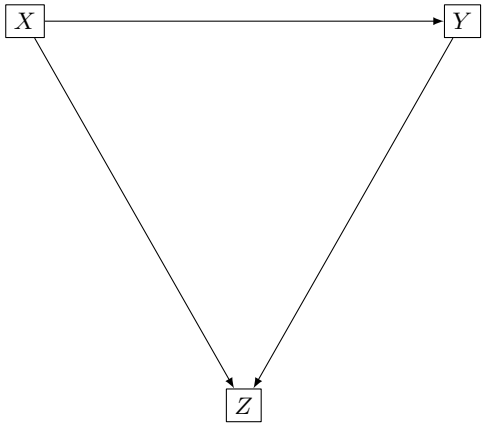


Figure 20.2 "Controlling for" additional variables can introduce bias into estimates of causal effects. Here the effect of X on Y is directly identifiable, $\Pr(Y|do(X = x)) = \Pr(Y|X = x)$. If we also condition on Z however, because it is a common *effect* of X and Y, we'd get $\Pr(Y|X = x, Z = z) \neq \Pr(Y|X = x)$. In fact, even if there were no arrow from X to Y, conditioning on Z would make Y depend on X.

"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 20.2 illustrates, and as several of the data-analysis problem sets will drive home [[CROSS-REF]], *adding* an ill-chosen covariate to a regression can create confounding.

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 influ-

encing 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.

20.3.1 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 20.3.) When S meets the back-door criterion,

$$\Pr(Y|do(X=x)) = \sum_s \Pr(Y|X=x, S=s)\Pr(S=s) \tag{20.4}$$

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. 20.3 that

$$\Pr(Y|do(X=x)) = \sum_t \Pr(\mathrm{Pa}(X)=t)\Pr(Y|X=x, \mathrm{Pa}(X)=t) \tag{20.5}$$

We can always introduce another set of conditioned variables, if we also sum over them:

$$\Pr(Y|do(X=x)) = \sum_t \Pr(\mathrm{Pa}(X)=t)\sum_s \Pr(Y, S=s|X=x, \mathrm{Pa}(X)=t) \tag{20.6}$$

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

$$\begin{aligned}\Pr(Y, S|X=x, \mathrm{Pa}(X)=t) = \\ \Pr(Y|X=x, \mathrm{Pa}(X)=t, S=s)\Pr(S=s|X=x, \mathrm{Pa}(X)=t)\end{aligned} \tag{20.7}$$

[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. 20.4.

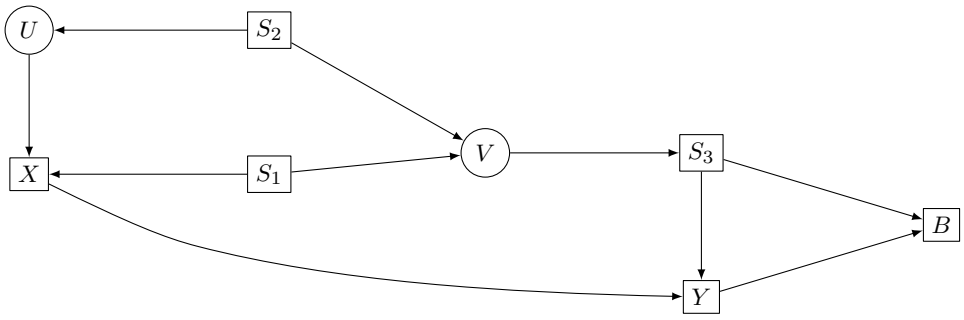


Figure 20.3 Illustration of the back-door criterion for identifying the causal effect of X on Y. Setting $S = \{S_1, S_2\}$ satisfies the criterion, but neither S_1 nor S_2 on their own would. Setting $S = \{S_3\}$, or $S = \{S_1, S_2, S_3\}$ also works. Adding B to any of the good sets makes them fail the criterion.

so

$$\Pr(Y|do(X=x)) = \sum_t \Pr(\text{Pa}(X)=t) \sum_s \Pr(Y|X=x, \text{Pa}(X)=t, S=s) \Pr(S=s|X=x, \text{Pa}(X)=t) \tag{20.8}$$

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

$$\Pr(Y|do(X=x)) = \sum_t \Pr(\text{Pa}(X)=t) \sum_s \Pr(Y|X=x, S=s) \Pr(S=s|X=x, \text{Pa}(X)=t) \tag{20.9}$$

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

$$\Pr(Y|do(X=x)) = \sum_t \Pr(\text{Pa}(X)=t) \sum_s \Pr(Y|X=x, S=s) \Pr(S=s|\text{Pa}(X)=t) \tag{20.10}$$

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

$$\Pr(Y|do(X=x)) = \sum_s \Pr(Y|X=x, S=s) \Pr(S=s) \tag{20.11}$$

as promised. □

20.3.1.1 The Entner Rules

Using the back-door criterion requires us to know the causal graph. Recently, Entner *et al.* (2013) have given a 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, or just irrelevant. 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.)

20.3.2 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

$$\Pr(Y|do(X=x)) = \sum_m \Pr(M=m|X=x) \sum_{x'} \Pr(Y|X=x', M=m) \Pr(X=x') \tag{20.12}$$

The variables M are sometimes called **mediators**.

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

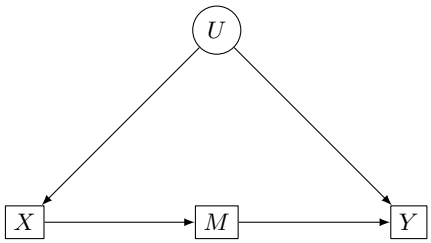


Figure 20.4 Illustration of the front-door criterion, after Pearl (2009b, Figure 3.5). X, Y and M are all observed, but U is an unobserved common cause of both X and Y. $X \leftarrow U \rightarrow Y$ is a back-door path confounding the effect of X on Y with their common cause. However, all of the effect 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 $\Pr(Y|do(X=x))$.

directed paths from X to Y, any *causal* dependence of Y on X must be mediated by a dependence of Y on M:

$$\Pr(Y|do(X=x)) = \sum_m \Pr(Y|do(M=m)) \Pr(M=m|do(X=x)) \quad (20.13)$$

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

$$\Pr(M=m|do(X=x)) = \Pr(M=m|X=x) \ . \quad (20.14)$$

Clause (iii) say that X satisfies the back-door criterion for identifying the effect of M on Y, and the inner sum in Eq. 20.12 is just the back-door computation (Eq. 20.4) of $\Pr(Y|do(M=m))$. So really we *are* using the back door criterion, twice. (See Figure 20.4.)

For example, in the "does tooth-brushing prevent heart-disease?" example of §19.2.2, 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 19.2.

20.3.2.1 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 ex-

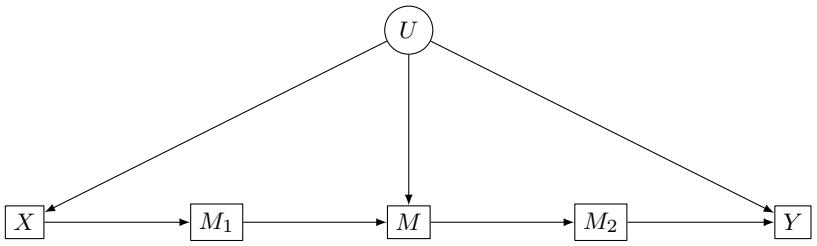


Figure 20.5 The path $X \to M \to Y$ contains all the mechanisms by which X influences Y, but is not isolated from the rest of the system ($U \to M$). The sub-mechanisms $X \to M_1 \to M$ and $M \to M_2 \to Y$ *are* isolated, and the original causal effect can be identified by composing them.

haustive 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 20.12.

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 20.5. 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. 20.13, 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.

20.3.3 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 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 20.6.)

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

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

[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).

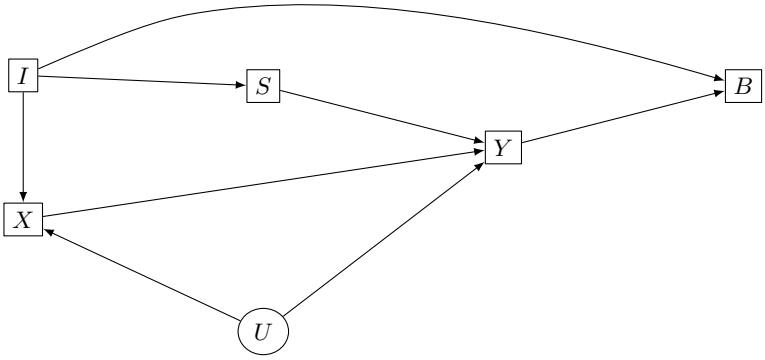


Figure 20.6 A valid instrumental variable, I, is related to the cause of interest, X, and influences Y only through its influence on X, at least once control variables block other paths. Here, to use I as an instrument, we *should* condition on S, but *should not* condition on B. (If we could condition on U, we would not need to use an instrument.)

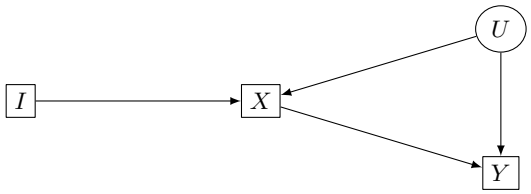


Figure 20.7 I acts as an instrument for estimating the effect of X on Y, despite the presence of the confounding, unobserved variable U.

must be channeled through X (by point (ii)), we have

$$\Pr(Y|do(I=i)) = \sum_x \Pr(Y|do(X=x)) \Pr(X=x|do(I=i)) \tag{20.15}$$

as in Eq. 20.3. We can thus identify the causal effect of X on Y whenever Eq. 20.15 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 I have bracketed in §20.3.3.3. The upshot is that while there *may* not be unique solutions, there *often* are, though they can be somewhat hard to calculate. However, in the special case where the relations between all variables are linear, we can be much more specific, fairly easily.

Let's start with the most basic possible set-up for an instrumental variable, namely that in Figure 20.7, where we just have X, Y, the instrument I, and the

[11] If X is continuous, then the analog of Eq. 20.15 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 $\Pr(Y|do(I=i))$.

unobserved confounders S. If everything is linear, identifying the causal effect of X on Y is equivalent to identifying the coefficient on the $X \to Y$ arrow. We can write

$$X = \alpha_0 + \alpha I + \delta U + \epsilon_X \tag{20.16}$$

and

$$Y = \beta_0 + \beta X + \gamma U + \epsilon_Y \tag{20.17}$$

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,

$$Y = \beta_0 + \beta\alpha_0 + \beta\alpha I + (\beta\delta + \gamma)U + \beta\epsilon_X + \epsilon_Y \tag{20.18}$$

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

$$Y = \gamma_0 + \beta\alpha I + \eta \tag{20.19}$$

where $\gamma_0 = \beta_0 + \beta\alpha_0$, and $\eta = (\beta\delta + \gamma)U + \beta\epsilon_X + \epsilon_Y$ has mean zero.

Now take the covariances:

$$\begin{aligned}
\mathrm{Cov}\left[I, X\right] &= \alpha\mathbb{V}\left[I\right] + \mathrm{Cov}\left[\epsilon_X, I\right] && (20.20)\\
\mathrm{Cov}\left[I, Y\right] &= \beta\alpha\mathbb{V}\left[I\right] + \mathrm{Cov}\left[\eta, I\right] && (20.21)\\
&= \beta\alpha\mathbb{V}\left[I\right] + (\beta\delta + \gamma)\mathrm{Cov}\left[U, I\right] && (20.22)\\
&\quad + \beta\mathrm{Cov}\left[\epsilon_X, I\right] + \mathrm{Cov}\left[\epsilon_Y, I\right]
\end{aligned}$$

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

$$\beta = \frac{\mathrm{Cov}\left[I, Y\right]}{\mathrm{Cov}\left[I, X\right]} \tag{20.23}$$

This can be estimated by substituting in the sample covariances, or any other consistent estimators of these two covariances. (§21.2 covers IV estimation in more detail.)

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

$$\begin{aligned}
\frac{\mathrm{Cov}\left[X, Y\right]}{\mathbb{V}\left[X\right]} &= \frac{\beta\mathbb{V}\left[X\right] + \gamma\mathrm{Cov}\left[U, X\right]}{\mathbb{V}\left[X\right]} && (20.24)\\
&= \beta + \gamma\frac{\mathrm{Cov}\left[U, X\right]}{\mathbb{V}\left[X\right]} && (20.25)\\
&= \beta + \gamma\frac{\delta\mathbb{V}\left[U\right]}{\alpha^2\mathbb{V}\left[I\right] + \delta^2\mathbb{V}\left[U\right] + \mathbb{V}\left[\epsilon_X\right]} && (20.26)
\end{aligned}$$

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].

[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 $\mathbb{E}\left[Y|X = x\right] = r_0 + r_1 x$, then Eq. 20.26 is *exactly* the coefficient we would want to use. OLS is doing its job.

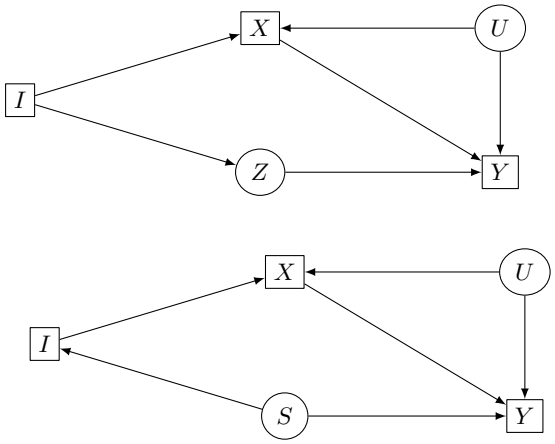


Figure 20.8 Left: I is not a valid instrument for identifying the effect of X on Y, because I can influence Y through a path not going through X. If we could control for Z, however, I would become valid. Right: I is not a valid instrument for identifying the effect of X on Y, because there is an unblocked back-door path connecting I and Y. If we could control for S, however, I would become valid.

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). §21.2 will also look at the more complicated estimation methods which can handle this, again assuming linearity.

20.3.3.1 Some Invalid Instruments

Not everything which looks like an instrument actually works. The ones which don't are called **invalid** instruments. 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 20.8, left). If there are unblocked back-door paths linking I and Y, e.g., if I and Y have common ancestors, then I is again not a valid instrument (Figure 20.8, right).

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".

20.3.3.2 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 (mathematical) life. Just as the front-door criterion relies on using our scientific knowledge, or rather theories, to find isolated and exhaustive mechanisms, finding valid instruments relies on theories about the part of the world under investigation, 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, by this point in the book, if someone tries to sell you a linear regression, you should be very skeptical, but let's leave that to one side. (It's *possible* that the problem at hand 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 often 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

[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.

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 whether the steps taken to double-blind the trial actually worked is itself a causal assumption.

More generally, any argument that a candidate instrument is valid is really an argument that other channels of information flow, apart from the favored one through X, can be ruled out. This generally cannot be done through analyzing the same variables used in the instrumental-variable estimation (see below), but involves theories about the world, and rests on the strength of the evidence for those theories. As has been pointed out multiple times — e.g., by Rosenzweig and Wolpin (2000) and 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 on the path $I \rightarrow X \leftarrow U \rightarrow Y$, so conditioning on X actually creates an indirect dependence between I and Y *even if* I is valid. So $I \not\perp\!\!\!\perp Y|X$, whether or not the instrument is valid, and the test (even if done 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. 20.23 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[15].

20.3.3.3 Instrumental Variables and Integral Equations

I said above (p. 20.3.3) that, in general, identifying causal effects through instrumental variables means solving integral equations. It's worth exploring that, because it provides some insight into how instrumental variables works, especially for non-linear systems. Since this is somewhat mathematically involved, however, you may want to skip this section on first reading.

To grasp what it means to identify causal effects by solving integral equations, let's start with the most basic set up, where the cause X, the effect Y, and the instrument I are all binary. There are then really only two numbers that

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

[15] Young (2017) re-analyzes hundreds of published papers in economics to argue that this is scenario is actually rather common.

need to be identified, $\Pr(Y = 1|do(X = 0))$ and $\Pr(Y = 1|do(X = 1))$. Eq. 20.15 becomes now a system of equations involving these effects:

$$\Pr(Y = 1|do(I = 0)) = \Pr(Y = 1|do(X = 0))\Pr(X = 0|do(I = 0)) + \Pr(Y = 1|do(X = 1))\Pr(X =$$
$$\Pr(Y = 1|do(I = 0)) = \Pr(Y = 1|do(X = 0))\Pr(X = 0|do(I = 1)) + \Pr(Y = 1|do(X = 1))\Pr(X =$$

The left-hand sides are identifiable (by the assumptions on I), as are the probabilities $\Pr(X|do(I))$. So, once we get those, we have a system of two linear equations with two unknowns, $\Pr(Y = 1|do(X = 0))$ and $\Pr(Y = 1|do(X = 1))$. Since there are as many equations as unknowns, there is a unique solution, unless the equations are redundant (Exercise 20.4).

If we put together some vectors and matrices,

$$\vec{f}_{I\to Y} \equiv \begin{bmatrix} \Pr(Y = 1|do(I = 0)) \\ \Pr(Y = 1|do(I = i)) \end{bmatrix} \tag{20.28}$$

$$\vec{f}_{X\to Y} \equiv \begin{bmatrix} \Pr(Y = 1|do(X = 0)) \\ \Pr(Y = 1|do(X = 1)) \end{bmatrix} \tag{20.29}$$

$$\mathbf{f}_{I\to X} \equiv \begin{bmatrix} \Pr(X = 0|do(I = 0)) & \Pr(X = 1|do(I = 0)) \\ \Pr(X = 0|do(I = 1)) & \Pr(X = 1|do(I = 1)) \end{bmatrix} \tag{20.30}$$

then Eq. 20.27 becomes

$$\vec{f}_{I\to Y} = \mathbf{f}_{I\to X}\vec{f}_{X\to Y} \tag{20.31}$$

and we can make the following observations:

1. The effect of the instrument I on the response Y, $\vec{f}_{I\to Y}$, is a linear transformation of the desired causal effects, $\vec{f}_{X\to Y}$.
2. Getting those desired effects requires inverting a linear operator, the matrix $\mathbf{f}_{I\to X}$.
3. That inversion is possible if, and only if, all of the eigenvalues of $\mathbf{f}_{I\to X}$ are non-zero.

There is nothing too special about the all-binary case, except that we can write everything out explicitly. If the cause, effect and instrument are all categorical, with the number of levels being c_x, c_y and c_i respectively, then there are $(c_y - 1)c_x$ parameters to identify, and Eq. 20.15 leads to a system of $(c_y - 1)c_i$ equations, so the effects will be identifiable (in general) so long as $c_i \geq c_x$. There will, once again, be a matrix form of the system of equations, and solving the system means inverting a matrix in whose entries are the effects of I on X, $\Pr(X = x|do(I = i))$. This, in turn, is something we can do so long as all of the eigenvalues are non-zero.

In the continuous case, we will replace our vectors by conditional density functions:

$$f_{I\to Y}(y|i) \equiv f(y|do(I = i)) \tag{20.32}$$
$$f_{X\to Y}(y|x) \equiv f(y|do(X = x)) \tag{20.33}$$
$$f_{I\to X}(x|i) \equiv f(x|do(I = i)) \tag{20.34}$$

Eq. 20.15 now reads

$$f_{I\to Y}(y|i) = \int f_{X\to Y}(y|x) f_{I\to X}(x|i) dx \tag{20.35}$$

This is linear in the desired function $f_{X\to Y}$, so we define the linear operator

$$\Phi h \equiv \int h(x) f_{I\to X}(x|i) dx \tag{20.36}$$

and re-write Eq. 20.15 one last time as

$$f_{I\to Y} = \Phi f_{I\to X} \tag{20.37}$$

which could be solved by

$$\Phi^{-1} f_{I\to Y} = f_{I\to X} \tag{20.38}$$

An operator like Φ is called an "integral operator", and equations like Eq. 20.35 or 20.37 are "integral equations".

If we take Eq. 20.15, multiply both sides by y, and sum (or integrate) over all possible y, we get

$$\mathbb{E}[Y|do(I=i)] = \sum_y \sum_x y \Pr(Y=y|do(X=x)) \Pr(X=x|do(I=i)) \tag{20.39}$$

$$= \sum_x \sum_y y \Pr(Y=y|do(X=x)) \Pr(X=x|do(I=i)) \tag{20.40}$$

$$= \sum_x \Pr(X=x|do(I=i)) \mathbb{E}[Y|do(X=x)] \tag{20.41}$$

$$= \Phi \mathbb{E}[Y|do(X)] \tag{20.42}$$

So, again, the conditional expectations (= average causal effects) we'd like to identify can be obtained by solving a linear integral equation. This doesn't require that either the functions $\mathbb{E}[Y|do(I=i)]$ or $\mathbb{E}[Y|do(X=x)]$ be linear (in i and x, respectively), it just follows from the Markov property[16].

20.3.4 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 20.4. 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).

[16] In fact, one reason the Markov property is important in studying dynamics is that it lets us move from studying non-linear individual trajectories to the linear evolution of probability distributions (Lasota and Mackey, 1994).

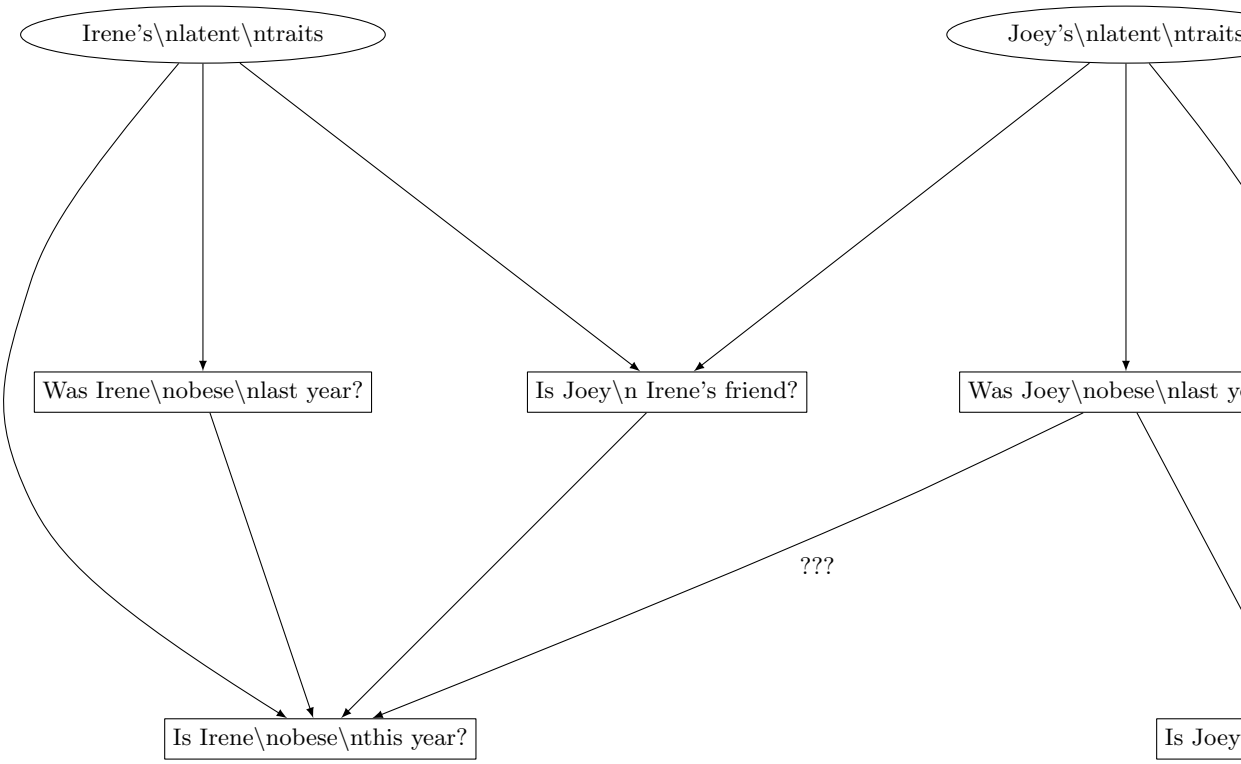


Figure 20.9 Social influence is confounded with selecting friends with similar traits, unobserved in the data.

As an example of the unidentifiable case, consider Figure 20.9. 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[17].

The difficulty arises from the latent variables for Irene and Joey (the round nodes in Figure 20.9). These include all the traits of either person which (a) influence who they become friends with, and (b) influence whether or not they

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

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[18], etc. Put simply, if Irene and Joey are friends because they spend two hours in the same bar every day drinking and eating fried 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[19].

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).

20.4 Summary

Of the four techniques I have introduced, instrumental variables are clever, but fragile and over-sold[20]. 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

[18] 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, *in that environment*, influence susceptibility to obesity. 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.

[19] Of course, the issue is not just 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.

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

a wrong guess seems better than nothing[21]. 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[22]; and indicate how different your conclusions could be if you made different assumptions.

20.4.1 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 linear 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).

Exercises

20.1 Draw a graphical model representing the situation where a causal variable X is randomized by an experimenter. Verify that $\Pr(Y|X=x)$ is then equal to $\Pr(Y|do(X=x))$. (*Hint:* Use the back door criterion.)

20.2 Prove Eq. 20.3, by using the causal Markov property of the appropriate surgically-altered graph.

1. The variable T contains all the parents of X; V contains all variables other than X, Y, and T. Explain why

$$\Pr\left(Y=y, X=x', T=t, V=v|do(X=x)\right) = \delta_{xx'} \frac{\Pr(Y=y, X=x, T=t, V=v)}{\Pr(X=x|T=t)} \tag{20.43}$$

where δ_{ij} is the "Kronecker delta", 1 when $i=j$ and 0 when $i \neq j$.
Hint: The left-hand side of the equation has to factor according to the graph we get after intervening on X, and the probability in the numerator on the right-hand side comes from the graphical model before the intervention. How do they differ?

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

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

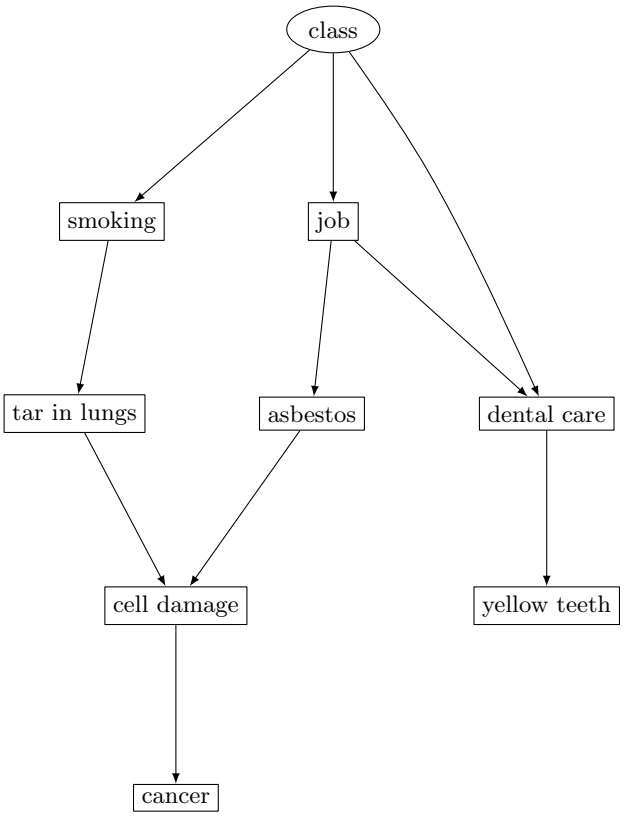


Figure 20.10 DAG for Exercise 20.3.

2. Assuming Eq. 20.43 holds, show that

$$\Pr\left(Y=y, X=x', T=t, V=v|do(X=x)\right) = \delta_{xx'} \Pr\left(Y=y, X=x, T=t, V=v|X=x, T=t\right) \Pr\left(T=t\right) \tag{20.44}$$

   Hint: $\Pr\left(A|B\right) = \Pr\left(A, B\right) / \Pr\left(B\right)$.
3. Assuming Eq. 20.44 holds, use the law of total probability to derive Eq. 20.3, i.e., to derive

$$\Pr\left(Y=y|do(X=x)\right) = \sum_t \Pr\left(Y=y|X=x, T=t\right) \Pr\left(T=t\right) \tag{20.45}$$

20.3 Refer to Figure 20.10. Can we use the front door criterion to estimate the effect of occupational prestige on cancer? If so, give a set of variables which we would use as mediators. 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?

20.4 Solve Eq. 20.27 for $\Pr\left(Y=1|do(X=0)\right)$ and $\Pr\left(Y=1|do(X=1)\right)$ in terms of the other conditional probabilities. When is the solution unique?

20.5 (*Lengthy, conceptual, open-ended*) Read Salmon (1984). When does his "statistical relevance basis" provide enough information to identify causal effects?

21

Estimating Causal Effects from Observations

Chapter 20 gave us ways of identifying causal effects, that is, of knowing when quantities like $\Pr(Y = y|do(X = x))$ are functions of the distribution of observable variables. Once we know that something is identifiable, the next question is how we can actually estimate it from data.

21.1 Estimators in the Back- and Front- Door Criteria

The back-door and front-door criteria for identification not only show us when causal effects are identifiable, they actually give us formulas for representing the causal effects in terms of ordinary conditional probabilities. When S satisfies the back-door criterion (Chapter 14), we can use parametric density models, we can model $Y|X, S = f(X, S) + \epsilon_Y$ and use regression, etc. If $\widehat{\Pr}(Y = y|X = x, S = s)$ is a consistent estimator of $\Pr(Y = y|X = x, S = s)$, and $\widehat{\Pr}(S = s)$ is a consistent estimator of $\Pr(S = s)$, then

$$\sum_s \widehat{\Pr}(S = s)\, \widehat{\Pr}(Y = y|X = x, S = s) \tag{21.1}$$

will be a consistent estimator of $\Pr(Y|do(X = x))$.

In principle, I could end this section right here, but there are some special cases and tricks which are worth knowing about. For simplicity, I will in this section only work with the back-door criterion, since estimating with the front-door criterion amounts to doing two rounds of back-door adjustment.

21.1.1 Estimating Average Causal Effects

Because $\Pr(Y|do(X = x))$ is a probability distribution, we can ask about $\mathbb{E}[Y|do(X = x)]$, when it makes sense for Y to have an expectation value; it's just

$$\mathbb{E}[Y|do(X = x)] = \sum_y y \Pr(Y = y|do(X = x)) \tag{21.2}$$

as you'd hope. This is the **average effect**, or sometimes just **the effect** of $do(X = x)$. While it is certainly not *always* the case that it summarizes all there is to know about the effect of X on Y, it is often useful.

If we identify the effect of X on Y through the back-door criterion, with control

variables S, then some algebra shows

$$\mathbb{E}[Y|do(X=x)] = \sum_y y \Pr(Y=y|do(X=x)) \tag{21.3}$$

$$= \sum_y y \sum_s \Pr(Y=y|X=x, S=s)\Pr(S=s) \tag{21.4}$$

$$= \sum_s \Pr(S=s) \sum_y y \Pr(Y=y|X=x, S=s) \tag{21.5}$$

$$= \sum_s \Pr(S=s)\mathbb{E}[Y|X=x, S=s] \tag{21.6}$$

The inner conditional expectation is just the regression function $\mu(x,s)$, for when we try to make a point-prediction of Y from X and S, so now all of the regression methods from Part I come into play. We would, however, still need to know the distribution $\Pr(S)$, so as to average appropriately. Let's turn to this.

21.1.2 Avoiding Estimating Marginal Distributions

We'll continue to focus on estimating the causal effect of X on Y using the back-door criterion, i.e., assuming we've found a set of control variables S such that

$$\Pr(Y=y|do(X=x)) = \sum_s \Pr(Y=y|X=x, S=s)\Pr(S=s) \tag{21.7}$$

S will generally contain multiple variables, so we are committed to estimating two potentially quite high-dimensional distributions, $\Pr(S)$ and $\Pr(Y|X,S)$. Even assuming that we knew all the distributions, just enumerating possible values s and summing over them would be computationally demanding. (Similarly, if S is continuous, we would need to do a high-dimensional integral.) Can we reduce these burdens?

One useful short-cut is to use the law of large numbers, rather than exhaustively enumerating all possible values of s. Notice that the left-hand side fixes y and x, so $\Pr(Y=y|X=x,S=s)$ is just some function of s. If we have an IID sample of realizations of S, say $s_1, s_2, \ldots s_n$, then the law of large numbers says that, for all well-behaved function f,

$$\frac{1}{n}\sum_{i=1}^{n} f(s_i) \to \sum_s f(s)\Pr(S=s) \tag{21.8}$$

Therefore, with a large sample,

$$\Pr(Y=y|do(X=x)) \approx \frac{1}{n}\sum_{i=1}^{n} \Pr(Y=y|X=x, S=s_i) \tag{21.9}$$

and this will still be (approximately) true when we use a consistent estimate of the conditional probability, rather than its true value.

The same reasoning applies for estimating $\mathbb{E}[Y|do(X=x)]$. Moreover, we can use the same reasoning to avoid explicitly summing over all possible s if we

do have $\Pr(S)$, by simulating from it[1]. Even if our sample (or simulation) is not completely IID, but is statistically stationary, in the sense we will cover in Chapter 23 (strictly speaking: "ergodic"), then we can still use this trick.

None of this gets us away from having to estimate $\Pr(Y|X,S)$, which is still going to be a high-dimensional object, if S has many variables.

21.1.3 Matching

Suppose that our causal variable of interest X is binary, or (almost equivalent) that we are only interested in comparing the effect of two levels, $do(X=1)$ and $do(X=0)$. Let's call these the "treatment" and "control" groups for definiteness, though nothing really hinges on one of them being in any sense a normal or default value (as "control" suggests) — for instance, we might want to know not just whether men get paid more than women, but whether they are paid more *because* of their sex[2]. In situations like this, we are often not so interested in the full distributions $\Pr(Y|do(X=1))$ and $\Pr(Y|do(X=0))$, but just in the expectations, $\mathbb{E}[Y|do(X=1)]$ and $\mathbb{E}[Y|do(X=0)]$. In fact, we are often interested just in the *difference* between these expectations, $\mathbb{E}[Y|do(X=1)] - \mathbb{E}[Y|do(X=0)]$, what is often called the **average treatment effect**, or ATE.

Suppose we are the happy possessors of a set of control variables S which satisfy the back-door criterion. How might we use them to estimate this average treatment effect?

$$\begin{aligned} ATE &= \sum_s \Pr(S=s)\,\mathbb{E}[Y|X=1,S=s] - \sum_s \Pr(S=s)\,\mathbb{E}[Y|X=0,S=s] \\ &= \sum_s \Pr(S=s)\left(\mathbb{E}[Y|X=1,S=s] - \mathbb{E}[Y|X=0,S=s]\right) \end{aligned} \tag{21.11}$$

[1] This is a "Monte Carlo" approximation to the full expectation value.

[2] The example is both imperfect and controversial. It is imperfect because biological sex (never mind socio-cultural gender) is not *quite* binary, even in mammals, though the exceptional cases are quite rare. (See Dreger 1998 for a historical perspective.) It is controversial because many statisticians insist that there is no sense in talking about causal effects unless there is some actual manipulation or intervention one could do to change X for an actually-existing "unit" — see, for instance, Holland (1986), which seems to be the source of the slogan "No causation without manipulation". I will just note that (i) this is the kind of metaphysical argument which statisticians usually avoid (if we can't talk about sex or race as causes, because changing those makes the subject a "different person", how about native language? the shape of the nose? hair color? whether they go to college? age at which they started school? grades in school?); (ii) genetic variables are highly manipulable with modern experimental techniques, though we don't use those techniques on people; (iii) real scientists routinely talk about causal effects with no feasible manipulation (e.g., "continental drift causes earthquakes"), or even imaginable manipulation (e.g., "the solar system formed because of gravitational attraction"). It may be merely coincidence that (iv) many of the statisticians who make such pronouncements work or have worked for the Educational Testing Service, an organization with an interest in asserting that, strictly speaking, sex and race cannot have any *causal* role in the score anyone gets on the SAT. (Points (i)–(iii) follow Glymour (1986); Glymour and Glymour (2014); Marcellesi (2013).)

Abbreviate $\mathbb{E}[Y|X=x,S=s]$ as $\mu(x,s)$, so that the average treatment effect is

$$\sum_s (\mu(1,s)-\mu(0,s))\Pr(S=s) = \mathbb{E}[\mu(1,S)-\mu(0,S)] \tag{21.12}$$

Suppose we got to observe μ. Then we could use the law of large numbers argument above to say

$$ATE \approx \frac{1}{n}\sum_{i=1}^{n} \mu(1,s_i)-\mu(0,s_i) \tag{21.13}$$

Of course, we don't get to see either $\mu(1,s_i)$ or $\mu(0,s_i)$. We don't even get to see $\mu(x_i,s_i)$. At best, we get to see $Y_i = \mu(x_i,s_i)+\epsilon_i$, with ϵ_i being mean-zero noise.

Clearly, we need to estimate $\mu(1,s_i)-\mu(0,s_i)$. In principle, any consistent estimator of the regression function, $\widehat{\mu}$, would do. If, for some reason, you were scared of doing a regression, however, the following scheme might occur to you: First, find all the units in the sample with $S=s$, and compare the mean Y for those who are treated ($X=1$) to the mean Y for those who are untreated ($X=0$). Writing the the set of units with $X=1$ and $S=s$ as $\mathcal{T}_s$, and the set of units with $X=0$ and $S=s$ as $\mathcal{C}_s$, then

$$\sum_s \left(\frac{1}{|\mathcal{T}_s|}\sum_{i\in\mathcal{T}_s} Y_i - \frac{1}{|\mathcal{C}_s|}\sum_{j\in\mathcal{C}_s} Y_j\right)\Pr(S=s) \tag{21.14}$$

$$= \sum_s \left(\frac{1}{|\mathcal{T}_s|}\sum_{i\in\mathcal{T}_s} \mu(1,s)+\epsilon_i - \frac{1}{|\mathcal{C}_s|}\sum_{j\in\mathcal{C}_s} \mu(0,s)+\epsilon_j\right)\Pr(S=s) \tag{21.15}$$

$$= \sum_s (\mu(1,s)-\mu(0,s))\Pr(S=s) + \sum_s \left(\frac{1}{|\mathcal{T}_s|}\sum_{i\in\mathcal{T}_s}\epsilon_i - \frac{1}{|\mathcal{C}_s|}\sum_{j\in\mathcal{C}_s}\epsilon_j\right)\Pr(S=s) \tag{21.16}$$

The first part is what we want, and the second part is an average of noise terms, so it goes to zero as $n\to\infty$. Thus we have a consistent estimator of the average treatment effect.

We could however go further. Take any unit i where $X=1$; it has some value s_i for the covariates. Suppose we can find another unit i^* with the same value of the covariates, but with $S=0$. Then

$$Y_i - Y_{i^*} = \mu(1,s_i)+\epsilon_i-\mu(0,s_i)-\epsilon_{i^*} \tag{21.17}$$

The comparison between the response of the treated unit and this **matched** control unit is an unbiased estimate of $\mu(1,s_i)-\mu(0,s_i)$. If we can find a match i^* for every unit i, then

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - Y_{i^*} \tag{21.18}$$

$$= \frac{1}{n}\sum i = 1^n \mu(1,s_i)-\mu(0,s_i) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \tag{21.19}$$

The first average is, by the law-of-large-numbers argument, approximately the

average treatment effect, and the second is the average of noise terms, so it should be going to zero as $n \to \infty$. Thus, matching gives us a consistent estimate of the average treatment effect, without any *explicit* regression. Instead, we rely on a paired comparison, because members of the treatment group are being compared to with members of the control group with matching values of the covariates S. This often works vastly better than estimating μ through a linear model.

There are three directions to go from here. One is to deal with all of the technical problems and variations which can arise. We might match each unit against multiple other units, to get further noise reduction. If we can't find an exact match, the usual approach is to match each treated unit against the control-group unit with the closest values of the covariates. Exploring these details is important to applications, but we won't follow it up here (see further readings).

A second direction is to remember that matching does not solve the identification problem. Computing Eq. 21.19 only gives us an estimate of the average treatment effect if S satisfies the back-door criterion. If S does not, then even if matching is done perfectly, Eq. 21.19 does nothing of any particular interest. Matching is one way of estimating *identified* average treatment effects; it contributes nothing to *solving* identification problems.

Third, and finally, matching is really doing nearest neighbor regression (§1.5.1). To get the difference between the responses of treated and controlled units, we're comparing each treated unit to the control-group unit with the closest values of the covariates. When people talk about *matching* estimates of average treatment effects, they usually mean that the number of nearest neighbors we use for each treated unit is fixed as n grows.

Once we realize that matching is really just nearest-neighbor regression, it may become less compelling; at the very least many issues should come to mind. As we saw in §1.5.1, to get consistent estimates of μ out of k-nearest neighbors, we need to let k grow (slowly) with n. If k is fixed, then the bias of $\widehat{\mu}(x, s)$ is either zero or goes quickly to zero as n grows (quicker the smaller k is), but $\mathbb{V}\left[\widehat{\mu}x, s\right] \not\to 0$ as $n \to \infty$. If all we want to do is estimate the average treatment effect, this remaining asymptotic variance at each s will still average out, but it would be a problem if we wanted to look at anything more detailed. More generally, the bias-variance tradeoff is a *tradeoff*, and it's not always a good idea to prioritize low bias over anything else. Moreover, it's not exactly clear that we *should* use a fixed k, or for that matter should use nearest neighbors instead of any other consistent regression method.

Nearest neighbor regression, like every other nonparametric method, is subject to the curse of dimensionality[3]; therefore, so is matching[4]. It would be very nice

[3] An important caveat: when S is high-dimensional but all the data fall on or very near a low-dimensional sub-space, nearest neighbor regression will adapt to this low effective dimensionality (Kpotufe, 2011). Not all regression methods have this nice property.

[4] If we can could do matching easily for high-dimensional S, then we could match treated units to other treated units, and control-group units to control-group units, and do easy high-dimensional regression. Since we know high-dimensional regression is hard, and we just reduced regression to matching, high-dimensional matching must be at least as hard.

if there was some way of lightening the curse when estimating treatment effects. We'll turn to that next.

21.1.4 Propensity Scores

The problems of having to estimate high-dimensional conditional distributions and of averaging over large sets of control values are both reduced if the set of control variables has in fact only a few dimensions. If we have two sets of control variables, S and R, both of which satisfy the back-door criterion for identifying $\Pr(Y|do(X=x))$, all else being equal we should use R if it contains fewer variables than S.[5]

An important special instance of this is when we can set $R = f(S)$, for some function S, and have

$$X \perp\!\!\!\perp S|R \tag{21.20}$$

In the jargon, R is a **sufficient statistic**[6] for predicting X from S. To see why this matters, suppose now that we try to identify $\Pr(Y=y|do(X=x))$ from a back-door adjustment for R alone, not for S. We have[7]

$$\sum_r \Pr(Y|X=x, R=r)\Pr(R=r) \tag{21.21}$$

$$= \sum_{r,s} \Pr(Y, S=s|X=x, R=r)\Pr(R=r)$$

$$= \sum_{r,s} \Pr(Y|X=x, R=r, S=s)\Pr(S=s|X=x, R=r)\Pr(R=r) \tag{21.22}$$

$$= \sum_{r,s} \Pr(Y|X=x, S=s)\Pr(S=s|X=x, R=r)\Pr(R=r) \tag{21.23}$$

$$= \sum_{r,s} \Pr(Y|X=x, S=s)\Pr(S=s|R=r)\Pr(R=r) \tag{21.24}$$

$$= \sum_{s} \Pr(Y|X=x, S=s)\sum_r \Pr(S=s, R=r) \tag{21.25}$$

$$= \sum_{s} \Pr(Y|X=x, S=s)\Pr(S=s) \tag{21.26}$$

$$= \Pr(Y|do(X=x)) \tag{21.27}$$

That is to say, if S satisfies the back-door criterion, then so does R. Since R is a function of S, both the computational and the statistical problems which come from using R are no worse than those of using S, and possibly much better, if R has much lower dimension.

[5] Other things which might not be equal: the completeness of data on R and S; parametric assumptions might be more plausible for the variables in S, giving a better rate of convergence; we might be more confident that S really does satisfy the back-door criterion.

[6] This is not the same sense of the word "sufficient" as in "causal sufficiency".

[7] Going from Eq. 21.22 to Eq. 21.23 uses the fact that $R = f(S)$, so conditioning on both R and S is the same as just conditioning on S. Going from Eq. 21.23 uses the fact that $S \perp\!\!\!\perp X|R$.

It may seem far-fetched that such a summary score should exist, but really all that's required is that some combinations of the variables in S carry the same information about X as the whole of S does. Consider for instance, the set-up where

$$X \leftarrow \sum_{j=1}^{p} V_j + \epsilon_X \tag{21.28}$$

$$Y \leftarrow f(X, V_1, V_2, \ldots V_p) + \epsilon_Y \tag{21.29}$$

To identify the effect of X on Y, we need to block the back-door paths between them. Each one of the V_j provides such a back-door path, so we need to condition on *all* of them. However, if $R = \sum_{j=1}^{p} V_j$, then $X \perp\!\!\!\perp \{V_1, V_2, \ldots V_p\} | R$, so we could reduce a p-dimensional set of control variables to a one-dimensional set.

Often, as here, finding summary scores will depend on the functional form, and so not be available in the general, non-parametric case. There is, however, an important special case where, if we can use the back-door criterion at all, we can use a one-dimensional summary.

This is the case where X is binary. If we set $f(S) = \Pr(X = 1|S = s)$, and then take this as our summary R, it is not hard to convince oneself that $X \perp\!\!\!\perp S|R$ (Exercise 21.1). This $f(S)$ is called the **propensity score**. It is remarkable, and remarkably convenient, that an arbitrarily large set of control variables S, perhaps with very complicated relationships with X and Y, can always be boiled down to a single number between 0 and 1, but there it is.

That said, except in very special circumstances, there is no analytical formula for $f(S)$. This means that it must be modeled and estimated. The most common model used is logistic regression, but so far as I can see this is just because many people know no other way to model a binary outcome. Since accurate propensity scores are needed to make the method work, it would seem to be worthwhile to model R very carefully, and to consider GAM or fully non-parametric estimates. If S contains a lot of variables, then estimating $\Pr(X = 1|S = s)$ is a high-dimensional regression problem, and so itself subject to the curse of dimensionality.

21.1.5 Propensity Score Matching

If the number of covariates in S is large, the curse of dimensionality settles upon us. Many values of S will have few or no individuals at all in the data set, let alone a large number in both the treatment and the control groups. Even if the real difference $\mathbb{E}[Y|X = 1, S = s] - \mathbb{E}[Y|X = 0, S = s]$ is small, with only a few individuals in either sub-group we could easily get a large difference in sample means. And of course with continuous covariates in S, each individual will generally have no exact matches at all.

The very clever idea of Rosenbaum and Rubin (1983) is to ameliorate this by matching not on S, but on the propensity score $R = \Pr(X = 1|S)$ defined above (p. 461). We have seen already that when X is binary, adjusting for the

propensity score is just as good as adjusting for the full set of covariates S. It is easy to double-check (Exercise 21.2) that

$$\begin{aligned}&\sum_s \Pr(S=s)\left(\mathbb{E}[Y|X=1,S=s]-\mathbb{E}[Y|X=0,S=s]\right)\\&=\sum_r \Pr(R=r)\left(\mathbb{E}[Y|X=1,R=r]-\mathbb{E}[Y|X=0,R=r]\right)\end{aligned} \tag{21.30}$$

when $R = \Pr(X=1|S=s)$, so we lose nothing, for these purposes, by matching on the propensity score R rather than on the covariates S. Intuitively, we now compare each treated individual with one who was just as likely to have received the treatment, but, by chance, did not[8]. On average, the differences between such matched individuals have to be due to the treatment.

What have we gained by doing this? Since R is always a one-dimensional variable, no matter how big S is, it is going to be *much* easier to find matches on R than on S. This does not actually break the curse of dimensionality, but rather shifts its focus, from the regression of Y on X and S to the regression of X on S. Still, this can be a very real advantage.

It is important to be clear, however, that the gain here is in computational tractability and (perhaps) statistical efficiency, not in fundamental identification. With $R = \Pr(X=1|S=s)$, it will always be true that $X \perp\!\!\!\perp S|R$, *whether or not* the back-door criterion is satisfied. If the criterion is satisfied, in principle there is nothing stopping us from using matching on S to estimate the effect, except our own impatience. If the criterion is not satisfied, having a compact one-dimensional summary of the wrong set of control variables is just going to let us get the wrong answer faster.

Some confusion seems to have arisen on this point, because, conditional on the propensity score, the treated group and the control group have the same distribution of covariates. (Again, recall that $X \perp\!\!\!\perp S|R$.) Since treatment and control groups have the same distribution of covariates in a randomized experiment, some people have concluded that propensity score matching is just as good as randomization[9]. This is emphatically *not* the case.

21.2 Instrumental-Variables Estimates

§20.3.3 introduced the idea of using instrumental variables to identify causal effects. Roughly speaking, I is an instrument for identifying the effect of X on Y when I is a cause of X, but the only way I is associated with Y is through directed paths which go through X. To the extent that variation in I predicts variation in X and Y, this can only be because X has a causal influence on Y. More precisely, given some controls S, I is a valid instrument when $I \not\perp\!\!\!\perp X|S$, and every path from I to Y left open by S has an arrow into X.

[8] Methods of approximate matching often work better on propensity scores than on the full set of covariates, because the former are lower-dimensional.

[9] These people do not include Rubin and Rosenbaum, but it is easy to see how their readers could come away with this impression. See Pearl (2009b, §11.3.5), and especially Pearl (2009a).

In the simplest case, of Figure 20.7, we saw that when everything is linear, we can find the causal coefficient of Y on X as

$$\beta = \frac{\text{Cov}[I, Y]}{\text{Cov}[I, X]} \tag{21.31}$$

A one-unit change in I causes (on average) an α-unit change in X, and an $\alpha\beta$-unit change in Y, so β is, as it were, the gearing ratio or leverage of the mechanism connecting I to Y.

Estimating β by plugging in the sample values of the covariances into Eq. 21.31 is called the **Wald estimator** of β. In more complex situations, we might have multiple instruments, and be interested in the causal effects of multiple variables, and we might have to control for some covariates to block undesired paths and get valid instruments. In such situations, the Wald estimator breaks down.

There is however a more general procedure which still works, provided the linearity assumption holds. This is called **two-stage regression**, or **two-stage least squares** (2SLS).

1. Regress X on I and S. Call the fitted values $\hat{x}$.
2. Regress Y on $\hat{x}$ and S, but *not* on I. The coefficient of Y on $\hat{x}$ is a consistent estimate of β.

The logic is very much as in the Wald estimator: conditional on S, variations in I are independent of the rest of the system. The only way they can affect Y is through their effect on X. In the first stage, then, we see how much changes in the instruments affect X. In the second stage, we see how much these I-caused changes in X change Y; and this gives us what we want.

To actually prove that this works, we would need to go through some heroic linear algebra to show that the population version of the two-stage estimator is actually equal to β, and then a straight-forward argument that plugging in the appropriate sample covariance matrices is consistent. The details can be found in any econometrics textbook, so I'll skip them. (But see Exercise 21.4.)

As mentioned in §21.2, there are circumstances where it is possible to use instrumental variables in nonlinear and even nonparametric models. The technique becomes far more complicated, however, because finding $\Pr(Y = y|do(X = x))$ requires solving Eq. 20.15,

$$\Pr(Y|do(I = i)) = \sum_x \Pr(Y|do(X = x)) \Pr(X = x|do(I = i))$$

and likewise finding $\mathbb{E}[Y|do(X = x)]$ means solving

$$\mathbb{E}[Y|do(I = i)] = \sum_x \mathbb{E}[Y|do(X = x)] \Pr(X = x|do(I = i)) \tag{21.32}$$

When, as is generally the case, x is continuous, we have rather an integral equation,

$$\mathbb{E}[Y|do(I = i)] = \int \mathbb{E}[Y|do(X = x)] \, p(x|do(I = i))dx \tag{21.33}$$

Solving such integral equations is not (in general) impossible, but it is hard, and the techniques needed are much more complicated than even two-stage least squares. I will not go over them here, but see Li and Racine (2007, chs. 16–17).

21.3 Uncertainty and Inference

The point of the identification strategies from Chapter 20 is to reduce the problem of causal inference to that of ordinary statistical inference. Having done so, we can assess our uncertainty about any of our estimates of causal effects the same way we would assess any other statistical inference. If we want confidence intervals or standard errors for $\mathbb{E}[Y|do(X=1)] - \mathbb{E}[Y|do(X=0)]$, for instance, we can treat our estimate of this like any other point estimate, and proceed accordingly. In particular, we can use the bootstrap (Chapter 6), if analytical formulas are unavailable or unappealing.

The one wrinkle to the use of analytical formulas comes from two-stage least-squares. Taking standard errors, confidence intervals, etc., for β from the usual formulas for the second regression neglects the fact that this estimate of β comes from regressing Y on $\hat{x}$, which is itself an estimate and so uncertain. Even if this is handled with some care, two-stage least squares is extraordinarily vulnerable to any violations in the usual assumptions about IID Gaussian errors. Young (2017), reviewing over 1000 (!) instrumental-variable regressions from top economics journals, shows that this is not merely a theoretical concern, but undermines a huge amount of the published literature.

21.4 Recommendations

Instrumental variables are a very clever idea, but they need to be treated with caution. They only work if the instruments are valid, and that validity rests just on assumptions about the causal structure. The crucial point, after all, is that the instrument is an indirect cause of Y, but *only* through X, with no other (unblocked) paths connecting I to Y. This can only too easily fail, if some indirect path has been neglected. They also require great care in their statistical inference (Young, 2017).

Matching, especially propensity score matching, is just as ingenious, and just as much at the mercy of the correctness of the DAG. Whether we match directly on covariates, or indirectly through the propensity score, what matters is whether the covariates really block off the back-door pathways between X and Y. If the covariates block those pathways, well and good; any consistent form of regression will work, including one called "matching" because "nonparametric nearest-neighbor smoothing" sounds too scary. If the covariates do not block the back-door pathways, then no amount of statistical ingenuity is going to help you.

There is a curious divide, among practitioners, between those who lean mostly on instrumental variables, and those who lean mostly on matching. The former

tend to suspect that (in our terms) the covariates used in matching are not enough to block all the back-door paths[10], *and* to think that the work is more or less over once an exogenous variable has been found. The matchers, for their part, think the instrumentalists are too quick to discount the possibility that their instruments are connected to Y through unmeasured pathways[11], but that if you match on enough variables, you've got to block the back-door paths. (They don't often worry that in doing so they might be activating colliders, or blocking front-door paths.) As is often the case in science, there is much truth to each faction's criticism of the other side. *You* are now in a position to think more clearly about these matters, and to act more intelligently, than many practitioners.

Throughout these chapters, we have been assuming that we know the correct DAG. Without such assumptions, or ones equivalent to them, none of these ideas can be used. In the next chapter, then, we will look at how to actually begin *discovering* causal structure from data.

21.5 Further Reading

The material in §21.1 is largely "folklore", though see Morgan and Winship (2007), which also treats instrumental variable estimation, and a number of other, more specialized techniques, like "regression discontinuity designs" and "difference in differences". It does not, however, consider nonparametric regression methods.

On matching, Stuart (2010) is another good review, including software as well as methods. For some of the asymptotic theory, including the connection to nearest neighbor methods, see Abadie and Imbens (2006).

The propensity score matching method has become incredibly popular since

[10] As an example for their side, Arceneaux *et al.* (2010) applied matching methods to an actual experiment, where the real causal relations could be worked out straightforwardly. Well-conduced propensity-score "matching suggests that [a] pre-election phone call that encouraged people to wear their seat belts also generated huge increases in voter turnout". The paper gives a convincing explanation of where this illusory effect comes from, i.e., of what the unblocked back-door path is, which I will not spoil for you.

[11] For instance, a widely-promoted preprint by three economists argued that watching television caused autism in children. (I leave tracking down the manuscript as an exercise for the reader.) The economists used the variation in how much it rains across different locations in the states of California, Oregon and Washington as an instrument (I) to predict average TV-watching (X) and its affects on the prevalence of autism (Y). It is certainly plausible that kids watch more TV when it rains, and that neither TV-watching nor autism causes rain. But this leaves open the question of whether rain and the prevalence of autism might not have some common cause, and for the west coast of the US in particular it is easy to find one. It is well-established that the risk of autism is higher among children of older parents, and that more-educated people tend to have children later in life. All three states have, of course, a striking contrast between large, rainy cities full of educated people (San Francisco, Portland, Seattle), and very dry, very rural locations on the other side of the mountains. Thus there is a (potential) uncontrolled common cause of rain and autism, namely geographic location, and the situation is as in Figure 20.8. — For a rather more convincing effort to apply ideas about causal inference to understanding the changing prevalence of autism, see Liu *et al.* (2010).

Rosenbaum and Rubin (1983), and there are a huge number of implementations of various versions of it. The `optmatch` package in R is notable for doing the actual matching in an extremely flexible and efficient way, but leaves defining matching criteria largely to the user (Hansen and Klopfer, 2006). The `MatchIt` package (Ho *et al.*, 2011) includes more tools for actually calculating propensity scores or other mesures of similarity, and then doing the matching. Stuart (2010) is also good on relevant software in R and other languages.

Rubin and Waterman (2006) is an extremely clear and easy-to-follow introduction to propensity score matching as a method of causal inference; Imbens and Rubin (2015) is a more comprehensive presentation of the work done by Rubin, Imbens and collaborators on estimating causal effects by matching, propensity scores, and instrumental variables. (Many of the original papers are reprinted in Rubin 2006.) While sound on theory, that book's worked examples cannot be recommended as examples of statistical craft (Shalizi, 2016).

King and Nielsen (2016) is an interesting argument against matching on propensity scores, in favor of matching on the full set of covariates, related to the extra variance of estimating the propensity scores.

Exercises

21.1 Suppose X is binary, and define $R = \Pr(X = 1|S)$. Show that $X \perp\!\!\!\perp S|R$. .

21.2 Prove Eq. 21.30.

21.3 Suppose that X has three levels, say $0, 1, 2$. Let R be the vector $(\Pr(X = 0|S = s), \Pr(X = 1|S = s))$. Prove that $X \perp\!\!\!\perp S|R$. (This is how to generalize propensity scores to non-binary X.)

21.4 For the situation in Figure 20.7, prove that the two-stage least-squares estimate of β is the same as the Wald estimate.

22

Discovering Causal Structure from Observations

[[ATTN: Further examples]]

The last few chapters have, hopefully, convinced you that when you want to do causal inference, it would help to know the causal graph. We have seen how the graph would let us calculate the effects of actual or hypothetical manipulations of the variables in the system. Furthermore, the graph tells us about what effects we can and cannot identify, and estimate, from observational data. But everything has posited that we know the graph somehow. This chapter finally deals with where the graph comes from.

There are fundamentally three ways to get the DAG:

- Prior knowledge
- Guessing-and-testing
- Discovery algorithms

There is only a little to say about the first, because, while it's important, it's not very statistical. As functioning adult human beings, you have a lot of everyday causal knowledge, which does not disappear the moment you start doing data analysis. Moreover, you are the inheritor of a vast scientific tradition which has, through patient observation, toilsome experiments, ingenious theorizing and intricate debate, acquired even more causal knowledge. You can and should use this. Someone's sex or race or caste at birth might be causes of the job they get or their income at age 30, but not the other way around. Running an electric current through a wire produces heat at a rate proportional to the square of the current. Malaria is due to a parasite transmitted by mosquitoes, and spraying mosquitoes with insecticides makes the survivors more resistant to those chemicals. All of these sorts of ideas can be expressed graphically, or at least as constraints on graphs.

We can, and should, also use graphs to represent scientific ideas which are not as secure as Ohm's law or the epidemiology of malaria. The ideas people work with in areas like psychology or economics, are really quite tentative, but they are ideas about the causal structure of parts of the world, and so graphical models are implicit in them.

All of which said, even if we think we know very well what's going on, we will often still want to check it, and that brings us the guess-and-test route.

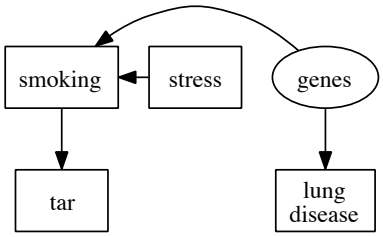


Figure 22.1 A hypothetical causal model in which smoking is associated with lung disease, but does not cause it. Rather, both smoking and lung disease are caused by common genetic variants. (This idea was due to R. A. Fisher.) Smoking is also caused, in this model, by stress.

22.1 Testing DAGs

A graphical causal model makes two kinds of qualitative claims. One is about direct causation. If the model says X is a parent of Y, then it says that changing X will change the (distribution of) Y. If we experiment on X (alone), moving it back and forth, and yet Y is unaltered, we know the model is wrong and can throw it out.

The other kind of claim a DAG model makes is about probabilistic conditional independence. If S d-separates X from Y, then $X \perp\!\!\!\perp Y|S$. If we observed X, Y and S, and see that $X \not\perp\!\!\!\perp Y|S$, then we know the model is wrong and can throw it out. (More: we know that there is a path linking X and Y which isn't blocked by S.) Thus in the model of Figure 22.1, lungdisease $\perp\!\!\!\perp$ tar|smoking. If lung disease and tar turn out to be dependent when conditioning on smoking, the model must be wrong.

This then is the basis for the guess-and-test approach to getting the DAG:

- Start with an initial guess about the DAG.
- Deduce conditional independence relations from d-separation.
- Test these, and reject the DAG if variables which ought to be conditionally independent turn out to be dependent.

This is a distillation of primary-school scientific method: formulate a hypotheses (the DAG), work out what the hypothesis implies, test those predictions, reject hypotheses which make wrong predictions.

It may happen that there are only a few competing, scientifically-plausible models, and so only a few, competing DAGs. Then it is usually a good idea to focus on checking predictions which *differ* between them. So in both Figure 22.1 and in Figure 22.2, stress $\perp\!\!\!\perp$ tar|smoking. Checking that independence thus does nothing to help us distinguish between the two graphs. In particular, confirming that stress and tar are independent given smoking really doesn't give us evidence

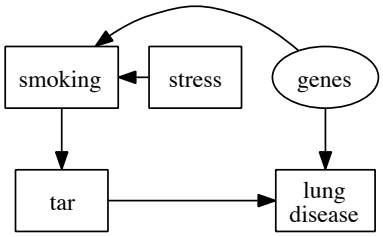


Figure 22.2 As in Figure 22.1, but now tar in the lungs does cause lung disease.

for the model from Figure 22.1, since it equally follows from the other model. If we want such evidence, we have to look for something they *disagree* about.

In any case, testing a DAG means testing conditional independence, so let's turn to that next.

22.2 Testing Conditional Independence

Recall from §18.4 that conditional independence is equivalent to zero conditional information: $X \perp\!\!\!\perp Y|Z$ if and only if $I[X;Y|Z] = 0$. In principle, this solves the problem. In practice, estimating mutual information is non-trivial, and in particular the sample mutual information often has a very complicated distribution. You *could* always bootstrap it, but often something more tractable is desirable. Completely general conditional independence testing is actually an active area of research. Some of this work is still quite mathematical (Sriperumbudur *et al.*, 2010), but it has already led to practical tests (Székely and Rizzo, 2009; Gretton *et al.*, 2012; Zhang *et al.*, 2011) and no doubt more are coming soon.

If all the variables are discrete, one just has a big contingency table problem, and could use a G^2 or χ^2 test. If everything is linear and multivariate Gaussian, $X \perp\!\!\!\perp Y|Z$ is equivalent to zero partial correlation[1]. Nonlinearly, if $X \perp\!\!\!\perp Y|Z$, then $\mathbb{E}[Y \mid Z] = \mathbb{E}[Y \mid X, Z]$, so if smoothing Y on X and Z leads to different predictions than just smoothing on Z, conditional independence fails. To reverse this, and go from $\mathbb{E}[Y \mid Z] = \mathbb{E}[Y \mid X, Z]$ to $X \perp\!\!\!\perp Y|Z$, requires the extra assumption that Y doesn't depend on X through its variance or any other moment. (This is weaker than the linear-and-Gaussian assumption, of course.)

The conditional independence relation $X \perp\!\!\!\perp Y|Z$ is fully equivalent to $\Pr(Y \mid X, Z) = \Pr(Y \mid Z)$. We could check this using non-parametric density estimation, though we would have to bootstrap the distribution of the test statistic. A more automatic, if slightly less rigorous, procedure comes from the idea mentioned in §14.5:

[1] Recall that the partial correlation between X and Y given Z is the correlation between X and Y, after linearly regressing each of them on Z separately. That is, it is the correlation of their residuals.

If X is in fact useless for predicting Y given Z, then an adaptive bandwidth selection procedure (like cross-validation) should realize that giving any finite bandwidth to X just leads to over-fitting. The bandwidth given to X should tend to the maximum allowed, smoothing X away altogether. This argument can be made more formal, and made into the basis of a test (Hall *et al.*, 2004; Li and Racine, 2007).

22.3 Faithfulness and Equivalence

In graphical models, d-separation implies conditional independence: if S blocks all paths from U to V, then $U \perp\!\!\!\perp V|S$. To reverse this, and conclude that if $U \perp\!\!\!\perp V|S$ then S must d-separate U and V, we need an additional assumption, already referred to in §19.2, called **faithfulness**. More exactly, if the distribution is faithful to the graph, then if S does not d-separate U from V, $U \not\perp\!\!\!\perp V|S$. The combination of faithfulness and the Markov property means that $U \perp\!\!\!\perp V|S$ if and only if S d-separates U and V.

This seems extremely promising. We can test whether $U \perp\!\!\!\perp V|S$ for any sets of variables we like. We could in particular test whether each pair of variables is independent, given all sorts of conditioning variable sets S. If we assume faithfulness, when we find that $X \perp\!\!\!\perp Y|S$, we know that S blocks all paths linking X and Y, so we learn something about the graph. If $X \not\perp\!\!\!\perp Y|S$ for all S, we would seem to have little choice but to conclude that X and Y are directly connected. Might it not be possible to reconstruct or discover the right DAG from knowing all the conditional independence and dependence relations?

This is on the right track, but too hasty. Start with just two variables:

$$X \rightarrow Y \Rightarrow X \not\perp\!\!\!\perp Y \tag{22.1}$$

$$X \leftarrow Y \Rightarrow X \not\perp\!\!\!\perp Y \tag{22.2}$$

With only two variables, there is only one independence (or dependence) relation to worry about, and it's the same no matter which way the arrow points.

Similarly, consider these arrangements of three variables:

$$X \rightarrow Y \rightarrow Z \tag{22.3}$$

$$X \leftarrow Y \leftarrow Z \tag{22.4}$$

$$X \leftarrow Y \rightarrow Z \tag{22.5}$$

$$X \rightarrow Y \leftarrow Z \tag{22.6}$$

The first two are chains, the third is a fork, the last is a collider. It is not hard to check (Exercise 22.1) that the first three DAGs all imply exactly the same set of conditional independence relations, which are different from those implied by the fourth[2]

[2] In all of the first three, $X \not\perp\!\!\!\perp Z$ but $X \perp\!\!\!\perp Z|Y$, while in the collider, $X \perp\!\!\!\perp Z$ but $X \not\perp\!\!\!\perp Z|Y$. Remarkably enough, the work which introduced the notion of forks and colliders, Reichenbach (1956), missed this — he thought that $X \perp\!\!\!\perp Z|Y$ in a collider as well as a fork. Arguably, this highly uncharacteristic mistake by a great scholar delayed the development of causal inference by thirty

These examples illustrate a general problem. There may be multiple graphs which imply the same independence relations, even when we assume faithfulness. When this happens, the exact same distribution of observables can factor according to, and be faithful to, all of those graphs. The graphs are thus said to be **equivalent**, or **Markov equivalent**. Observations alone cannot distinguish between equivalent DAGs. Experiment can, of course — changing Y alters both X and Z in a fork, but not a chain — which shows that there really is a difference between the DAGs, just not one *observational* data can track.

22.3.1 Partial Identification of Effects

Chapters 20–21 considered the identification and estimation of causal effects under the assumption that there was a single known graph. If there are multiple equivalent DAGs, then, as mentioned above, no amount of purely observational data can select a single graph. Background knowledge lets us rule out some equivalent DAGs[3], but it may not narrow the set of possibilities to a single graph. How then are we to actually do our causal estimation?

We *could* just pick one of the equivalent graphs, and do all of our calculations as though it were the only possible graph. This is often what people seem to do. The kindest thing one can say about it is that it shows confidence; phrases like "lying by omission" also come to mind.

A more principled alternative is to admit that the uncertainty about the DAG means that causal effects are only *partially* identified. Simply put, one does the estimation in each of the equivalent graphs, and reports the range of results[4]. If each estimate is consistent, then this gives a consistent estimate of the range of possible effects. Because the effects are not fully identified, this range will not narrow to a single point, even in the limit of infinite data, but admitting this, rather than claiming a non-existent precision, is simple scientific honesty.

22.4 Causal Discovery with Known Variables

Section 22.1 talks about how we can test a DAG, once we have it. This lets us eliminate some DAGs, but still leaves mysterious where they come from in the first place. While in principle there is nothing wrong which deriving your DAG from a vision of serpents biting each others' tails, so long as you test it, it would be nice to have a systematic way of finding good models. This is the problem of model discovery, and especially of causal discovery.

years or more, and is one of the reasons why, as Dean Eckles once put it, formal causal inference is an "idea behind its time" (http://www.deaneckles.com/blog/429_ideas-behind-their-time-formal-causal-inference/).

[3] If we know that X, Y and Z have to be in either a chain or a fork, with Y in the middle, and we know that X comes before Y in time, then we can rule out the fork and the chain $X \leftarrow Y \rightarrow Z$.

[4] Sometimes the different graphs will gave the same estimates of certain effects. For example, the chain $X \rightarrow Y \rightarrow Z$ and the fork $X \leftarrow Y \rightarrow Z$ will agree on the effect of Y on Z.

Causal discovery is silly with just one variable, and too hard for us with just two.[5]

With three or more variables, we have however a very basic principle. If there is no edge between X and Y, in either direction, then X is neither Y's parent nor its child. But any variable is independent of its non-descendants given its parents. Thus, for some set[6] of variables S, $X \perp\!\!\!\perp Y|S$ (Exercise 22.2). If we assume faithfulness, then the converse holds: if $X \perp\!\!\!\perp Y|S$, then there cannot be an edge between X and Y. Thus, there is no edge between X and Y if and only if we can make X and Y independent by conditioning on some S. Said another way, there is an edge between X and Y if and only if we cannot make the dependence between them go away, no matter what we condition on[7].

So let's start with three variables, X, Y and Z. By testing for independence and conditional independence, we could learn that there had to be edges between X and Y and Y and Z, but not between X and Z. But conditional independence is a symmetric relationship, so how could we **orient** those edges, give them direction? Well, to rehearse a point from the last section, there are only four possible directed graphs corresponding to that undirected graph:

- $X \rightarrow Y \rightarrow Z$ (a chain);
- $X \leftarrow Y \leftarrow Z$ (the other chain);
- $X \leftarrow Y \rightarrow Z$ (a fork on Y);
- $X \rightarrow Y \leftarrow Z$ (a collision at Y)

With the fork or either chain, we have $X \perp\!\!\!\perp Z|Y$. On the other hand, with the collider we have $X \not\perp\!\!\!\perp Z|Y$. Thus $X \not\perp\!\!\!\perp Z|Y$ if and only if there is a collision at Y. By testing for *this* conditional dependence, we can either definitely orient the edges, or rule out an orientation. If $X - Y - Z$ is just a subgraph of a larger graph, we can still identify it as a collider if $X \not\perp\!\!\!\perp Z|\{Y, S\}$ for *all* collections of nodes S (not including X and Z themselves, of course).

With more nodes and edges, we can **induce** more orientations of edges by consistency with orientations we get by identifying colliders. For example, suppose we know that X, Y, Z is either a chain or a fork on Y. If we learn that $X \rightarrow Y$, then the triple *cannot* be a fork, and must be the chain $X \rightarrow Y \rightarrow Z$. So orienting the $X - Y$ edge induces an orientation of the $Y - Z$ edge. We can also sometimes orient edges through background knowledge; for instance we might know that Y comes later in time than X, so if there is an edge between them it *cannot* run from Y to X.[8] We can eliminate other edges based on similar sorts of background

[5] But see Janzing (2007); Hoyer *et al.* (2009) for some ideas on how you could do it if you're willing to make some extra assumptions. The basic idea of these papers is that the distribution of effects given causes should be simpler, in some sense, than the distribution of causes given effects.

[6] Possibly empty: conditioning on the empty set of variables is the same as not conditioning at all.

[7] "No causation without association", as it were.

[8] Some have argued, or at least entertained the idea, that the logic here is backwards: rather than order in time constraining causal relations, causal order *defines* time order. (Versions of this idea are discussed by, inter alia, Russell (1927); Wiener (1961); Reichenbach (1956); Pearl (2009b); Janzing (2007) makes a related suggestion). Arguably then using order in time to orient edges in a causal graph begs the question, or commits the fallacy of *petitio principii*. But of course every syllogism

knowledge: men tend to be heavier than women, but changing weight does not change sex, so there can't be an edge (or even a directed path!) from weight to sex, though there could be one the other way around.

To sum up, we can rule out an edge between X and Y whenever we can make them independent by conditioning on other variables; and when we have an $X - Y - Z$ pattern, we can identify colliders by testing whether X and Z are dependent given Y. Having oriented the arrows going into colliders, we induce more orientations of other edges.

Putting these three things — edge elimination by testing, collider finding, and inducing orientations — gives the most basic causal discovery procedure, the SGS (Spirtes-Glymour-Scheines) algorithm (Spirtes *et al.*, 2001, §5.4.1, p. 82). This assumes:

1. The data-generating distribution has the causal Markov property on a graph G.
2. The data-generating distribution is faithful to G.
3. Every member of the population has the same distribution.
4. All relevant variables are in G.
5. There is only *one* graph G to which the distribution is faithful.

Abstractly, the algorithm works as follows:

- Start with a complete undirected graph on all p variables, with edges between all nodes.
- For each pair of variables X and Y, and each set of other variables S, see if $X \perp\!\!\!\perp Y | S$; if so, remove the edge between X and Y.
- Find colliders by checking for conditional dependence; orient the edges of colliders.
- Try to orient undirected edges by consistency with already-oriented edges; do this recursively until no more edges can be oriented.

Pseudo-code is in §22.7.

Call the result of the SGS algorithm $\widehat{G}$. If all of the assumptions above hold, and the algorithm is correct in its guesses about when variables are conditionally independent, then $\widehat{G} = G$. In practice, of course, conditional independence guesses are really statistical tests based on finite data, so we should write the output as $\widehat{G}_n$, to indicate that it is based on only n samples. If the conditional independence test is consistent, then

$$\lim_{n \to \infty} \Pr\left(\widehat{G}_n \neq G\right) = 0 \tag{22.7}$$

In other words, the SGS algorithm converges in probability on the correct causal

does, so this isn't a distinctively *statistical* issue. (Take the classic: "All men are mortal; Socrates is a man; therefore Socrates is mortal." How can we know that *all* men are mortal until we know about the mortality of this particular man, Socrates? Isn't this just like asserting that tomatoes and peppers must be poisonous, because they belong to the nightshade family of plants, all of which are poisonous?) While these philosophical issues are genuinely fascinating, this footnote has gone on long enough, and it is time to return to the main text.

structure; it is consistent for all graphs G. Of course, at finite n, the probability of error — of having the wrong structure — is (generally!) not zero, but this just means that, like any statistical procedure, we cannot be absolutely certain that it's not making a mistake.

One consequence of the independence tests making errors on finite data can be that we fail to orient some edges — perhaps we missed some colliders. These unoriented edges in $\widehat{G}_n$ can be thought of as something like a confidence region — they have *some* orientation, but multiple orientations are all compatible with the data.[9] As more and more edges get oriented, the confidence region shrinks.

If the fifth assumption above fails to hold, then there are multiple graphs G to which the distribution is faithful. This is just a more complicated version of the difficulty of distinguishing between the graphs $X \rightarrow Y$ and $X \leftarrow Y$. All the graphs in the equivalence class may have some arrows in common; in that case the SGS algorithm will identify those arrows. If some edges differ in orientation across the equivalence class, SGS will not orient them, even in the limit. In terms of the previous paragraph, the confidence region never shrinks to a single point, just because the data doesn't provide the information needed to do this. The graph is only partially identified.

If there *are* unmeasured relevant variables, we can get not just unoriented edges, but actually arrows pointing in both directions. This is an excellent sign that some basic assumption is being violated.

22.4.1 The PC Algorithm

The SGS algorithm is statistically consistent, but very computationally inefficient; the number of tests it does grows exponentially in the number of variables p. This is the worst-case complexity for *any* consistent causal-discovery procedure, but this algorithm just proceeds immediately to the worst case, not taking advantage of any possible short-cuts.

Since it's enough to find *one* S making X and Y independent to remove their edge, one obvious short-cut is to do the tests in some order, and skip unnecessary tests. On the principle of doing the easy work first, the revised edge-removal step would look something like this:

- For each X and Y, see if $X \perp\!\!\!\perp Y$; if so, remove their edge.
- For each X and Y which are still connected, and each third variable Z connected to X or Y, see if $X \perp\!\!\!\perp Y|Z$; if so, remove the edge between X and Y.
- For each X and Y which are still connected, and each third and fourth variables Z_1 and Z_2 both connected to X or both connected to Y, see if $X \perp\!\!\!\perp Y|Z_1, Z_2$; if so, remove the edge between X and Y.
- ...

[9] I say "multiple orientations" rather than "all orientations", because picking a direction for one edge might induce an orientation for others.

- For each X and Y which are still connected at the k^{th} stage, see if there are k variables $Z_1, Z_2, \ldots Z_k$ all connected to X or all connected to Y where $X \perp\!\!\!\perp Y | \{Z_1, \ldots Z_k\}$; if so, remove, the edge between X and Y.
- ...
- Stop when $k = p - 2$.

If all the tests are done correctly, this will give the same result as the SGS procedure (Exercise 22.4). And if some of the tests give erroneous results, conditioning on a small number of variables will tend to be more reliable than conditioning on more (why?).

We can be even more efficient, however. If $X \perp\!\!\!\perp Y|S$ for any S at all, then $X \perp\!\!\!\perp Y|S'$, where all the variables in S' are adjacent to X or Y (or both) (Exercise 22.3). To see the sense of this, suppose that there is a single long directed path running from X to Y. If we condition on any of the variables along the chain, we make X and Y independent, but we could always move the point where we block the chain to be either right next to X or right next to Y. So when we are trying to remove edges and make X and Y independent, we only need to condition on variables which are still connected to X and Y, not ones in totally different parts of the graph.

This then gives us the PC[10] algorithm (Spirtes *et al.* 2001, §5.4.2, pp. 84–88; see also §22.7). It works exactly like the SGS algorithm, except for the edge-removal step, where it tries to condition on as few variables as possible (as above), and only conditions on adjacent variables. The PC algorithm has the same assumptions as the SGS algorithm, and the same consistency properties, but generally runs much faster, and does many fewer statistical tests. It should be the default algorithm for attempting causal discovery.

22.4.2 Causal Discovery with Hidden Variables

Suppose that the set of variables we measure is *not* causally sufficient. Could we at least discover this? Could we possibly get hold of *some* of the causal relationships? Algorithms which can do this exist (e.g., the CI and FCI algorithms of Spirtes *et al.* (2001, ch. 6)), but they require considerably more graph-fu. (The RFCI algorithm (Colombo *et al.*, 2012) is a modern, fast successor to FCI.) The results of these algorithms can succeed in removing *some* edges between observable variables, and definitely orienting some of the remaining edges. If there are actually no latent common causes, they end up acting like the SGS or PC algorithms.

Partial identification of effects

When all relevant variables are observed, all effects are identified within one graph; partial identification happens because multiple graphs are equivalent. When some variables are not observed, we may have to use the identification strategies to get at the same effect. In fact, the same effect may be identified in

[10] Peter-Clark

one graph and not identified in another, equivalent graph. This is, again, unfortunate, but when it happens it needs to be admitted.

22.4.3 On Conditional Independence Tests

The abstract algorithms for causal discovery assume the existence of consistent tests for conditional independence. The implementations known to me mostly assume either that variables are discrete (so that one can basically use the χ^2 test), or that they are continuous, Gaussian, and linearly related (so that one can test for vanishing partial correlations), though the `pcalg` package does allow users to provide their own conditional independence tests as arguments. It bears emphasizing that these restrictions are *not* essential. As soon as you have a consistent independence test, you are, in principle, in business. In particular, consistent *non-parametric* tests of conditional independence would work perfectly well. An interesting example of this is the paper by Chu and Glymour (2008), on finding causal models for the time series, assuming additive but non-linear models.

22.5 Software and Examples

The PC and FCI algorithms are implemented in the stand-alone Java program `Tetrad` (`http://www.phil.cmu.edu/projects/tetrad/`). They are also implemented in the `pcalg` package on CRAN (Kalisch *et al.*, 2010, 2012). This package also includes functions for calculating the effects of interventions from fitted graphs, assuming linear models. The documentation for the package is somewhat confusing; rather see Kalisch *et al.* (2012) for a tutorial introduction.

[[TODO: Cleanup output from the package]]

It's worth going through how `pcalg` works[11]. The code is designed to take advantage of the modularity and abstraction of the PC algorithm itself; it separates actually finding the graph completely from performing the conditional independence test, which is rather a function the user supplies. (Some common ones are built in.) For reasons of computational efficiency, in turn, the conditional independence tests are set up so that the user can just supply a set of sufficient statistics, rather than the raw data.

Let's walk through an example[12], using the `mathmarks` data set. This contains grades ("marks") from 88 university students in five mathematical subjects, algebra, analysis, mechanics, statistics and vectors. All five variables are positively correlated with each other.

[11] A word about installing the package: you'll need the package `Rgraphviz` for drawing graphs, which is hosted not on CRAN (like `pcalg`) but on BioConductor. Try installing it, and its dependencies, before installing `pcalg`. See `http://www.bioconductor.org/packages/release/bioc/html/Rgraphviz.html` for help on installing `Rgraphviz`.

[12] After Spirtes *et al.* (2001, §6.12, pp. 152–154).

```
library(pcalg)
library(SMPracticals)
data(mathmarks)
suffStat <- list(C=cor(mathmarks),n=nrow(mathmarks))
pc.fit <- pc(suffStat, indepTest=gaussCItest, p=ncol(mathmarks),alpha=0.005)
```

This uses a Gaussian (-and-linear) test for conditional independence, `gaussCItest`, which is built into the `pcalg` package. Basically, it hopes to test whether $X \perp\!\!\!\perp Y|Z$ by testing whether the partial correlation of X and Y given Z is close to zero. These partial correlations can all be calculated from the correlation matrix, so the line before creates the sufficient statistics needed by `gaussCItest` — the matrix of correlations and the number of data points. We also have to tell `pc` how many variables there are, and what significance level to use in the test (here, 0.5%).

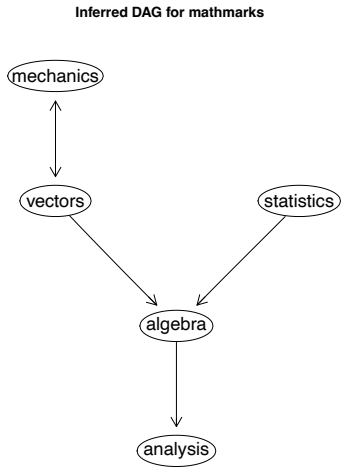
Before going on, I encourage you to run `pc` as above, but with `verbose=TRUE`, and to study the output.

Figure 22.3 shows the resulting DAG. If we take it seriously, it says that grades in analysis are driven by grades in algebra, while algebra in turn is driven by statistics and vectors. While one could make up stories for why this would be so (perhaps something about the curriculum?), it seems safer to regard this as a warning against *blindly* trusting any algorithm —- a key assumption of the PC algorithm, after all, is that there are no unmeasured but causally-relevant variables, and it is easy to believe these are violated. For instance, while *knowledge of* different mathematical fields may be causally linked (it would indeed be hard to learn much mechanics without knowing about vectors), test scores are only imperfect measurements of knowledge.

The size of the test may seem low, but remember we are doing a lot of tests:

```
summary(pc.fit)
## Object of class 'pcAlgo', from Call:
## pc(suffStat = suffStat, indepTest = gaussCItest, alpha = 0.005,
##     p = ncol(mathmarks))
##
## Nmb. edgetests during skeleton estimation:
## ===========================================
## Max. order of algorithm:  3
## Number of edgetests from m = 0 up to m = 3 :  20 38 10 0
##
## Graphical properties of skeleton:
## ================================
## Max. number of neighbours:  2 at node(s) 2
## Avg. number of neighbours:  1
##
## Adjacency Matrix G:
##   1 2 3 4 5
## 1 . 1 . . .
## 2 1 . 1 . .
## 3 . . . 1 .
## 4 . . . . .
## 5 . . 1 . .
```

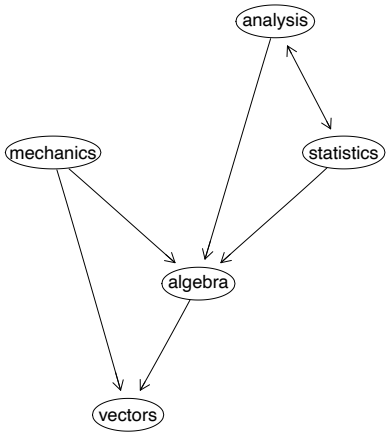
This tells us that it considered going up to conditioning on three variables (the




```
library(Rgraphviz)
plot(pc.fit,labels=colnames(mathmarks),main="Inferred DAG for mathmarks")
```

Figure 22.3 DAG inferred by the PC algorithm from the `mathmarks` data. Two-headed arrows, like undirected edges, indicate that the algorithm was unable to orient the edge. (It is obscure why `pcalg` sometimes gives an edge it cannot orient no heads and sometimes two.)

maximum possible, since there are only five variables), that it did twenty tests of unconditional independence, 31 tests where it conditioned on one variable, four tests where it conditioned on two, and none where it conditioned on three. This 55 tests in all, so a simple Bonferroni correction suggests the over-all size is $55 \times 0.005 = 0.275$. This is probably pessimistic (the Bonferroni correction typically is). Setting $\alpha = 0.05$ gives a somewhat different graph (Figure 22.4).




```
plot(pc(suffStat, indepTest=gaussCItest, p=ncol(mathmarks),alpha=0.05),
  labels=colnames(mathmarks),main="")
```

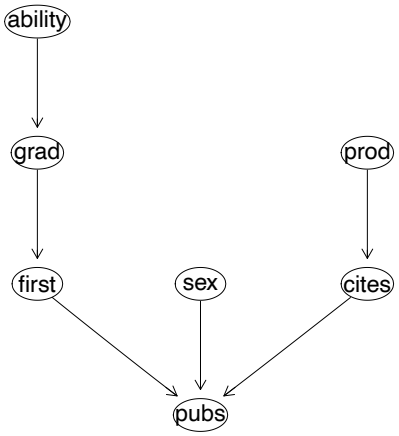
Figure 22.4 Inferred DAG when the size of the test is 0.05.

For a second example[13], let's use some data on academic productivity among psychologists. The two variables of ultimate interest were the publication (`pubs`) and citation (`cites`) rates, with possible measured causes including `ability` (basically, standardized test scores), graduate program quality `grad` (basically, the program's national rank), the quality of the psychologist's first job, `first`, a measure of productivity `prod`, and sex. There were 162 subjects, and while the actual data isn't reported, the correlation matrix is.

```
psychs
##         ability grad prod first   sex cites pubs
## ability    1.00 0.62 0.25  0.16 -0.10  0.29 0.18
## grad       0.62 1.00 0.09  0.28  0.00  0.25 0.15
## prod       0.25 0.09 1.00  0.07  0.03  0.34 0.19
## first      0.16 0.28 0.07  1.00  0.10  0.37 0.41
## sex       -0.10 0.00 0.03  0.10  1.00  0.13 0.43
## cites      0.29 0.25 0.34  0.37  0.13  1.00 0.55
## pubs       0.18 0.15 0.19  0.41  0.43  0.55 1.00
```

The model found by `pcalg` is fairly reasonable-looking (Figure 22.5). Of course, the linear-and-Gaussian assumption has no particular support here, and there is at least one variable for which it must be wrong (which?), but unfortunately with just the correlation matrix we cannot go further.

[13] Following Spirtes *et al.* (2001, §5.8.1, pp. 98–102).




```
plot(pc(list(C=psychs,n=162),indepTest=gaussCItest,p=7,alpha=0.01),
  labels=colnames(psychs),main="")
```

Figure 22.5 Causes of academic success among psychologists. The arrow from citations *to* publications is a bit odd, but not impossible — people who get cited more might get more opportunities to do research and so to publish.

22.6 Limitations on Consistency of Causal Discovery

There are some important limitations to causal discovery algorithms (Spirtes *et al.*, 2001, §12.4). They are *universally* consistent: for all causal graphs G,[14]

$$\lim_{n\to\infty} \Pr\left(\widehat{G}_n \neq G\right) = 0 \tag{22.8}$$

The probability of getting the graph wrong can be made arbitrarily small by using enough data. However, this says nothing about *how much* data we need to achieve a given level of confidence, i.e., the *rate* of convergence. *Uniform* consistency would mean that we could put a bound on the probability of error as a function of n which did not depend on the true graph G. Robins *et al.* (2003) proved that *no* uniformly-consistent causal discovery algorithm can exist. The issue, basically, is that the Adversary could make the convergence in Eq. 22.8 arbitrarily slow by selecting a distribution which, while faithful to G, came *very close* to being unfaithful, making some of the dependencies implied by the graph arbitrarily small. For any given dependence strength, there's some amount of data which will let us recognize it with high confidence, but the Adversary can make the required data size as large as he likes by weakening the dependence, without ever setting it to zero[15].

The upshot is that so *uniform*, *universal* consistency is out of the question; we can be *universally* consistent, but without a uniform rate of convergence; or we can converge *uniformly*, but only on some less-than-universal class of distributions. These might be ones where all the dependencies which do exist are not too weak (and so not too hard to learn reliably from data), or the number of true edges is not too large (so that if we haven't seen edges yet they probably don't exist; Janzing and Herrmann, 2003; Kalisch and Bühlmnann, 2007).

It's worth emphasizing that the Robins *et al.* (2003) no-uniform-consistency result applies to *any* method of discovering causal structure from data. Invoking human judgment, Bayesian prior distributions over possible causal structures, etc., etc., won't get you out of it.

22.7 Pseudo-code for the SGS Algorithm[16]

When you see a loop, assume that it gets entered at least once. "Replace" in the sub-functions always refers to the input graph.

```
SGS = function(set of variables V) {
    Ĝ = colliders(prune( complete undirected graph on V))
    until (Ĝ == G′) {
        Ĝ = G′
```

[14] If the true distribution is faithful to multiple graphs, then we should read G as their equivalence class, which has some undirected edges.

[15] See §18.4 for a more quantitative statement of how the required sample size relates to non-parametric measures of the strength of dependence.

[16] This section may be omitted on first (and maybe even second) reading.

```
            G' = orient(Ĝ)
        }
        return(Ĝ)
}

prune = function(G) {
        for each A, B ∈ V {
            for each S ⊆ V \ {A, B} {
                if A ⫫ B|S { G = G \ (A − B) }
            }
        }
        return(G)
}

collliders = function(G) {
        for each (A − B) ∈ G {
            for each (B − C) ∈ G {
                if (A − C) ∉ G {
                    collision = TRUE
                    for each S ⊂ B ∩ V \ {A, C} {
                        if A ⫫ C|S { collision = FALSE }
                    }
                    if (collision) { replace (A − B) with (A → B), (B − C) with (B ← C) }
                }
            }
        }
        return(G)
}

orient = function(G) {
        if ((A → B) ∈ G & (B − C) ∈ G & (A − C) ∉ G) { replace (B − C) with (B → C) }
        if ((directed path from A to B)∈ G & (A − B) ∈ G) { replace (A − B) with (A → B) }
        return(G)
}
```

22.8 Further Reading

The best single reference on causal discovery algorithms remains Spirtes *et al.* (2001). A lot of work has been done in recent years by the group centered around ETH-Zürich, beginning with Kalisch and Bühlmnann (2007), connecting this to modern statistical concerns about sparse effects and high-dimensional modeling.

As already mentioned, the best reference on partial identification is Manski (2007). Partial identification of causal effects due to multiple equivalent DAGs is considered in Maathuis *et al.* (2009), along with efficient algorithms for linear

systems, which are applied in Maathuis *et al.* (2010), and implemented in the `pcalg` package as `ida`.

Discovery is possible for directed cyclic graphs, though since it's harder to understand what such models mean, it is less well-developed. Important papers on this topic include Richardson (1996) and Lacerda *et al.* (2008).

Exercises

22.1 Prove that, assuming faithfulness, a three-variable chain and a three-variable fork imply exactly the same set of dependence and independence relations, but that these are different from those implied by a three-variable collider. Are any implications common to chains, forks, and colliders? Could colliders be distinguished from chains and forks without assuming faithfulness?

22.2 Prove that if X and Y are not parent and child, then either $X \perp\!\!\!\perp Y$, or there exists a set of variables S such that $X \perp\!\!\!\perp Y|S$. *Hint:* start with the Markov property, that any X is independent of all its non-descendants given its parents, and consider separately the cases where Y a descendant of X and those where it is not.

22.3 Prove that if $X \perp\!\!\!\perp Y|S$ for some set of variables S, then $X \perp\!\!\!\perp Y|S'$, where every variable in S' is a neighbor of X or Y.

22.4 Prove that the graph produced by the edge-removal step of the PC algorithm is exactly the same as the graph produced by the edge-removal step of the SGS algorithm. *Hint:* SGS removes the edge between X and Y when $X \perp\!\!\!\perp Y|S$ for *even one* set S.

22.5 When, exactly, does $\mathbb{E}[Y \mid X, Z] = \mathbb{E}[Y \mid Z]$ imply $Y \perp\!\!\!\perp X|Z$?

22.6 Would the SGS algorithm work on a non-causal, merely-probabilistic DAG? If so, in what sense is it a *causal* discovery algorithm? If not, why not?

22.7 Describe how to use bandwidth selection as a conditional independence test.

22.8 Read Kalisch *et al.* (2012) and write a conditional independence test function based on bandwidth selection (§14.5). Check that your test gives the right size when run on simulated cases where you know the variables are conditionally independent. Check that your test function works with `pcalg::pc`.