00:22 Friday 8th April, 2016

Chapter 27

Estimating Causal Effects from Observations

Chapter 25 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.

27.1 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, for instance,

$$\Pr(Y = y|do(X = x)) = \sum_s \Pr(S = s)\Pr(Y = y|X = x, S = s) \tag{27.1}$$

Everything on the right-hand side refers to the distribution of observables, following the usual DAG without any surgery.

This is *very handy*, because we have spent the whole first part of the book learning different ways of estimating distributions like $\Pr(S = s)$ and $\Pr(Y = y|X = x, S = s)$. We can do fully non-parametric density estimation (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

$$\sum_s \widehat{\Pr}(S = s)\widehat{\Pr}(Y = y|X = x, S = s) \tag{27.2}$$

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.

27.1.1 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

$$\mathbb{E}[Y|do(X=x)] = \sum_y y \Pr(Y=y|do(X=x)) \tag{27.3}$$

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

$$\begin{aligned}
\mathbb{E}[Y|do(X=x)] &= \sum_y y \Pr(Y=y|do(X=x)) && (27.4)\\
&= \sum_y y \sum_s \Pr(Y=y|X=x,S=s)\Pr(S=s) && (27.5)\\
&= \sum_s \Pr(S=s) \sum_y y \Pr(Y=y|X=x,S=s) && (27.6)\\
&= \sum_s \Pr(S=s)\mathbb{E}[Y|X=x,S=s] && (27.7)
\end{aligned}$$

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.

27.1.2 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

$$\Pr(Y=y|do(X=x)) = \sum_s \Pr(Y=y|X=x,S=s)\Pr(S=s) \tag{27.8}$$

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,

$$\frac{1}{n}\sum_{i=1}^{n} f(s_i) \rightarrow \sum_{s} f(s)\Pr(S = s) \tag{27.9}$$

Therefore, with a large sample,

$$\Pr(Y = y|do(X = x)) \approx \frac{1}{n}\sum_{i=1}^{n} \Pr(Y = y|X = x, S = s_i) \tag{27.10}$$

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 21 (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.

27.1.3 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 $\mathbb{E}[Y|do(X = 0)]$.

[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 cultural gender) is not *quite* binary, even in mammals, but it's close enough for a good approximation. 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?); (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 appears to 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).)

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 causal effect?

$$\begin{aligned} ATE &= \sum_s \Pr(S=s)\mathbb{E}[Y|X=1,S=s] - \sum_s \Pr(S=s)\mathbb{E}[Y|X=0,S=s] && (27.11)\\ &= \sum_s \Pr(S=s)(\mathbb{E}[Y|X=1,S=s] - \mathbb{E}[Y|X=0,S=s]) && (27.12)\end{aligned}$$

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

$$\sum_s (\mu(1,s)-\mu(0,s))\Pr(S=s) = \mathbb{E}[\mu(1,S)-\mu(0,S)] \quad (27.13)$$

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

$$ATE \approx \frac{1}{n}\sum_{i=1}^{n} \mu(1,s_i) - \mu(0,s_i) \quad (27.14)$$

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 $\mathscr{T}_s$, and the set of units with $X=0$ and $S=s$ as $\mathscr{C}_s$, then

$$\begin{aligned} &\sum_s \left(\frac{1}{|\mathscr{T}_s|}\sum_{i\in\mathscr{T}_s} Y_i - \frac{1}{|\mathscr{C}_s|}\sum_{j\in\mathscr{C}_s} Y_j\right)\Pr(S=s) && (27.15)\\ &= \sum_s \left(\frac{1}{|\mathscr{T}_s|}\sum_{i\in\mathscr{T}_s} \mu(1,s)+\epsilon_i - \frac{1}{|\mathscr{C}_s|}\sum_{j\in\mathscr{C}_s} \mu(0,s)+\epsilon_j\right)\Pr(S=s) && (27.16)\\ &= \sum_s (\mu(1,s)-\mu(0,s))\Pr(S=s) + \sum_s \left(\frac{1}{|\mathscr{T}_s|}\sum_{i\in\mathscr{T}_s}\epsilon_i - \frac{1}{|\mathscr{C}_s|}\sum_{j\in\mathscr{C}_s}\epsilon_j\right)\Pr(S=s) && (27.17)\end{aligned}$$

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

$$Y_i - Y_{i^*} = \mu(1, s_i) + \epsilon_i - \mu(0, s_i) - \epsilon_{i^*} \tag{27.18}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - Y_{i^*} \tag{27.19}$$

$$= \frac{1}{n}\sum i = 1^n \mu(1, s_i) - \mu(0, s_i) + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \tag{27.20}$$

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. Explore these details.

A second direction is to remember that matching does not solve the identification problem. Computing Eq. 27.20 only gives us an estimate of the average treatment effect if S satisfies the back door criterion. If it does not, then even if matching is done perfectly, Eq. 27.20 does nothing of any particular interest. Matching is one way of estimating identified average treatment effects; it can contribute 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 arise. 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 $\mathbb{V}[\widehat{\mu}x, s] \not\to 0$ 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; therefore, so is matching[3]. It would be very nice if there was some way of lightening the curse when estimating treatment effects. We'll turn to that next.

27.1.4 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[4]

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

$$X \perp\!\!\!\perp S|R \tag{27.21}$$

In the jargon, R is a **sufficient statistic**[5] 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[6]

$$\begin{aligned}
&\sum_r \Pr(Y|X=x,R=r)\Pr(R=r) && (27.22)\\
&= \sum_{r,s} \Pr(Y,S=s|X=x,R=r)\Pr(R=r) &&\\
&= \sum_{r,s} \Pr(Y|X=x,R=r,S=s)\Pr(S=s|X=x,R=r)\Pr(R=r) && (27.23)\\
&= \sum_{r,s} \Pr(Y|X=x,S=s)\Pr(S=s|X=x,R=r)\Pr(R=r) && (27.24)\\
&= \sum_{r,s} \Pr(Y|X=x,S=s)\Pr(S=s|R=r)\Pr(R=r) && (27.25)\\
&= \sum_{s} \Pr(Y|X=x,S=s)\sum_r \Pr(S=s,R=r) && (27.26)\\
&= \sum_{s} \Pr(Y|X=x,S=s)\Pr(S=s) && (27.27)\\
&= \Pr(Y|do(X=x)) && (27.28)
\end{aligned}$$

[3]To see this, observe that 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 can reduce regression to matching, high-dimensional matching must be at least as hard.

[4]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.

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

[6]Going from Eq. 27.23 to Eq. 27.24 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. 27.24 uses the fact that $S \perp\!\!\!\perp X|R$.

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.

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

$$X \leftarrow \sum_{j=1}^{p} V_j + \epsilon_X \tag{27.29}$$

$$Y \leftarrow f(X, V_1, V_2, \ldots V_p) + \epsilon_Y \tag{27.30}$$

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 $X \perp\!\!\!\perp \left\{V_1, V_2, \ldots V_p\right\} | R$, 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 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.

27.1.5 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, 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 solve this by matching not on S, but on the propensity score defined in the last section. 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 2) that

$$\begin{aligned}\sum_s \Pr(S=s)(\mathbb{E}[Y|X=1,S=s]-\mathbb{E}[Y|X=0,S=s]) \\ = \sum_r \Pr(R=r)(\mathbb{E}[Y|X=1,R=r]-\mathbb{E}[Y|X=0,R=r]) \end{aligned} \quad (27.31)$$

when $R = \Pr(X=1|S=s)$, so we lose no essential information 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[7]. 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[8]. This is emphatically *not* the case.

The propensity score matching method has become incredibly popular since 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. See Stuart (2010) for a fairly recent listing of relevant software in R and other languages.

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

[8] 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).

27.2 Instrumental-Variables Estimates

§25.3.3 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 $I \not\perp X|S$, and every path from I to Y left open by S has an arrow into X.

In the simplest case, of Figure 25.7, we saw that when everything is linear, we can find the causal coefficient of Y on X as

$$\beta = \frac{\mathrm{Cov}[I,Y]}{\mathrm{Cov}[I,X]} \tag{27.32}$$

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. 27.32 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 4.)

As mentioned in §27.2, 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. 25.15,

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

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

$$\mathbb{E}[Y|do(I=i)] = \sum_x \mathbb{E}[Y|do(X=x)] \Pr(X=x|do(I=i)) \tag{27.33}$$

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

$$\mathbb{E}[Y|do(I=i)] = \int \mathbb{E}[Y|do(X=x)]\, p(x|do(I=i))dx \tag{27.34}$$

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).

27.3 Uncertainty and Inference

The point of the identification strategies from Chapter 25 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.

27.4 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 is rests just as much on assumptions about the underlying DAG as any of the other identification strategies. 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.

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[9], and to think that the business 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[10], but that if you match on enough variables, you've got to block the back-door paths. (They don't often worry that they might be conditioning on colliders, or blocking front-door paths, as they do so.) 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.

27.5 Further Reading

The material in §27.1 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. For some of the asymptotic theory, including the connection to nearest neighbor methods, see Abadie and Imbens (2006).

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

[9] 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 for comparison. 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.

[10] [[TODO: Mention the rainfall-as-instrument paper?]] For instance, a recent and widely-promoted preprint by three economists argued that watching television caused autism in children. (I leave tracking down the paper as an exercise for the reader.) The economists used the variation in how much it rains across different locations in California, Oregon and Washington as an instrument 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 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 25.8. — For a rather more convincing effort to apply ideas about causal inference to understanding the changing prevalence of autism, see Liu *et al.* (2010).

(2015) is a more comprehensive presentation of the estimation work done by Rubin, Imbens and collaborators on estimating causal effects by matching, propensity scores, and instrumental variables. (Much of the original work is reprinted in Rubin 2006.)

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.

27.6 Exercises

1. Suppose X is binary, and define $R = \Pr(X = 1|S)$. Show that $X \perp\!\!\!\perp S|R$. .
2. Prove Eq. 27.31.
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.)
4. For the situation in Figure 25.7, prove that the two-stage least-squares estimate of β is the same as the Wald estimate.