a job. When some of the mandatory training groups contained more workers than training slots, training opportunities were distributed by lottery. Hence, training requirements were randomly assigned conditional on the covariates used to assign workers to groups. A regression on a dummy for training plus the personal characteristics, past unemployment variables, and job history variables used to classify workers seems very likely to provide reliable estimates of the causal effect of training.[13]

In the schooling context, there is usually no lottery that directly determines whether someone will go to college or finish high school.[14] Still, we might imagine subjecting individuals of similar ability and from similar family backgrounds to an experiment that encourages school attendance. The Education Maintenance Allowance, which pays British high school students in certain areas to attend school, is one such policy experiment (Dearden, et al, 2004).

A second type of study that favors the CIA exploits detailed institutional knowledge regarding the process that determines s_i . An example is the Angrist (1998) study of the effect of voluntary military service on the later earnings of soldiers. This research asks whether men who volunteered for service in the US Armed Forces were economically better off in the long run. Since voluntary military service is not randomly assigned, we can never know for sure. Angrist therefore used matching and regression techniques to control for observed differences between veterans and nonveterans who applied to get into the all-volunteer forces between 1979 and 1982. The motivation for a control strategy in this case is the fact that the military screens soldier-applicants primarily on the basis of observable covariates like age, schooling, and test scores.

The CIA in Angrist (1998) amounts to the claim that after conditioning on all these observed characteristics veterans and nonveterans are comparable. This assumption seems worth entertaining since, conditional on X_i, variation in veteran status in the Angrist (1998) study comes solely from the fact that some qualified applicants fail to enlist at the last minute. Of course, the considerations that lead a qualified applicant to "drop out" of the enlistment process could be related to earnings potential, so the CIA is clearly not guaranteed even in this case.

3.2.3 Bad Control

We've made the point that control for covariates can make the CIA more plausible. But more control is not always better. Some variables are bad controls and should not be included in a regression model even when their inclusion might be expected to change the short regression coefficients. Bad controls are variables that are themselves outcome variables in the notional experiment at hand. That is, bad controls might just as well be dependent variables too. Good controls are variables that we can think of as having been fixed at the time the regressor of interest was determined.

The essence of the bad control problem is a version of selection bias, albeit somewhat more subtle than

[13] This program appears to raise earnings, primarily because workers in the training group went back to work more quickly.

[14] Lotteries have been used to distribute private school tuition subsidies; see, e.g., Angrist, et al. (2002).

the selection bias discussed in Chapter (2) and Section (3.2). To illustrate, suppose we are interested in the effects of a college degree on earnings and that people can work in one of two occupations, white collar and blue collar. A college degree clearly opens the door to higher-paying white collar jobs. Should occupation therefore be seen as an omitted variable in a regression of wages on schooling? After all, occupation is highly correlated with both education and pay. Perhaps it's best to look at the effect of college on wages for those within an occupation, say white collar only. The problem with this argument is that once we acknowledge the fact that college affects occupation, comparisons of wages by college degree status within an occupation are no longer apples-to-apples, *even if college degree completion is randomly assigned.*

Here is a formal illustration of the bad control problem in the college/occupation example.[15] Let W_i be a dummy variable that denotes white collar workers and let Y_i denote earnings. The realization of these variables is determined by college graduation status and potential outcomes that are indexed against C_i. We have

$$\begin{aligned} \mathrm{Y}_i &= \mathrm{C}_i\mathrm{Y}_{1i} + (1-\mathrm{C}_i)\mathrm{Y}_{0i} \\ \mathrm{W}_i &= \mathrm{C}_i\mathrm{W}_{1i} + (1-\mathrm{C}_i)\mathrm{W}_{0i} \end{aligned}$$

where $\mathrm{C}_i = 1$ for college graduates and is zero otherwise, $\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\}$ denotes potential earnings, and $\{\mathrm{W}_{1i},\mathrm{W}_{0i}\}$ denotes potential white-collar status. We assume that C_i is randomly assigned, so it is independent of all potential outcomes. We have no trouble estimating the causal effect of C_i on either Y_i or W_i since independence gives us

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{C}_i=1\right] - E\left[\mathrm{Y}_i|\mathrm{C}_i=0\right] &= E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}\right], \\ E\left[\mathrm{W}_i|\mathrm{C}_i=1\right] - E\left[\mathrm{W}_i|\mathrm{C}_i=0\right] &= E\left[\mathrm{W}_{1i}-\mathrm{W}_{0i}\right]. \end{aligned}$$

In practice, we might estimate these average treatment effects by regressing Y_i and W_i and on C_i.

Bad control means that a comparison of earnings conditional on W_i does not have a causal interpretation. Consider the difference in mean earnings between college graduates and others conditional on working at a white collar job. We can compute this in a regression model that includes W_i or by regressing Y_i on C_i in the sample where $\mathrm{W}_i = 1$. The estimand in the latter case is the difference in means with C_i switched off and on, conditional on $\mathrm{W}_i = 1$:

$$E\left[\mathrm{Y}_i|\mathrm{W}_i=1,\mathrm{C}_i=1\right] - E\left[\mathrm{Y}_i|\mathrm{W}_i=1,\mathrm{C}_i=0\right] = E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1,\mathrm{C}_i=1\right] - E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1,\mathrm{C}_i=0\right] \quad (3.2.12)$$

[15] The same problem arises in "conditional-on-positive" comparisons, discussed in detail in section (3.4.2), below.

By the joint independence of $\{\mathrm{Y}_{1i},\mathrm{W}_{1i},\mathrm{Y}_{0i},\mathrm{W}_{0i}\}$ and C_i, we have

$$E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1,\mathrm{C}_i=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1,\mathrm{C}_i=0\right]=E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right].$$

This expression illustrates the apples-to-oranges nature of the bad-control problem:

$$\begin{aligned}&E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right]\\&=\underbrace{E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]}_{\text{causal effect on college grads}}+\underbrace{\left\{E\left[\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right]\right\}}_{\text{selection bias}}.\end{aligned}$$

In other words, the difference in wages between those with and without a college degree conditional on working in a white collar job equals the causal effect of college on those with $\mathrm{W}_{1i}=1$ (people who work at a white collar job when they have a college degree) and a selection-bias term which reflects the fact that college changes the composition of the pool of white collar workers.

The selection-bias in this context can be positive or negative, depending on the relation between occupational choice, college attendance, and potential earnings. The main point is that even if $\mathrm{Y}_{1i}=\mathrm{Y}_{0i}$, so that there is no causal effect of college on wages, the conditional comparison in (3.2.12) will not tell us this (the regression of Y_i on W_i and C_i has exactly the same problem). It is also incorrect to say that the conditional comparison captures the part of the effect of college that is "not explained by occupation." In fact, the conditional comparison does not tell us much that is useful without a more elaborate model of the links between college, occupation, and earnings.[16]

As an empirical illustration, we see that the addition of two-digit occupation dummies indeed reduces the schooling coefficient in the NLSY models reported in Table 3.2.1, in this case from .087 to .066. However, it's hard to say what we should make of this decline. The change in schooling coefficients when we add occupation dummies may simply be an artifact of selection bias. So we would do better to control only for variables that are not themselves caused by education.

A second version of the bad control scenario involves *proxy control*, that is, the inclusion of variables that might partially control for omitted factors, but are themselves affected by the variable of interest. A simple version of the proxy-control scenario goes like this: Suppose you are interested in a long regression, similar to equation (3.2.10),

$$\mathrm{Y}_i=\alpha+\rho\mathrm{S}_i+\gamma a_i+\varepsilon_i, \tag{3.2.13}$$

where for the purposes of this discussion we've replaced the vector of controls A_i, with a scalar ability measure a_i. Think of this as an IQ score that measures innate ability in eighth grade, before any relevant

[16] In this example, selection bias is probably negative, that is $E\left[\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]<E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right]$. It seems reasonable to think that any college graduate can get a white collar job, so $E\left[\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]$ is not too far from $E[\mathrm{Y}_{0i}]$. But someone who gets a white collar without benefit of a college degree (i.e., $\mathrm{W}_{0i}=1$) is probably special, i.e., has a better than average Y_{0i}.

schooling choices are made (assuming everyone completes eighth grade). The error term in this equation satisfies $E[\text{S}_i\varepsilon_i] = E[a_i\varepsilon_i] = 0$ by definition. Since a_i is measured before S_i is determined, it is a good control.

Equation (3.2.13) is the regression of interest, but unfortunately, data on a_i are unavailable. However, you have a second ability measure collected later, after schooling is completed (say, the score on a test used to screen job applicants). Call this variable "late ability," a_{li}. In general, schooling increases late ability relative to innate ability. To be specific, suppose

$$a_{li} = \pi_0 + \pi_1\text{S}_i + \pi_2 a_i. \tag{3.2.14}$$

By this, we mean to say that both schooling and innate ability increase late or measured ability. There is almost certainly some randomness in measured ability as well, but we can make our point more simply via the deterministic link, (3.2.14).

You're worried about OVB in the regression of Y_i on S_i alone, so you propose to regress Y_i on S_i and late ability, a_{li} since the desired control, a_i, is unavailable. Using (3.2.14) to substitute for a_i in (3.2.13), the regression on S_i and a_{li} is

$$\text{Y}_i = (\alpha - \gamma\frac{\pi_0}{\pi_2}) + (\rho - \gamma\frac{\pi_1}{\pi_2})\text{S}_i + \frac{\gamma}{\pi_2}a_{li} + \varepsilon_i. \tag{3.2.15}$$

In this scenario, γ, π_1, and π_2 are all positive, so $\rho - \gamma\frac{\pi_1}{\pi_2}$ is too small unless π_1 turns out to be zero. In other words, use of a proxy control that is increased by the variable of interest generates a coefficient below the desired effect. Importantly, π_1 can be investigated to some extent: if the regression of a_{li} on S_i is zero, you might feel better about assuming that π_1 is zero in (3.2.14).

There is an interesting ambiguity in the proxy-control story that is not present in the first bad-control story. Control for outcome variables is simply misguided; you do not want to control for occupation in a schooling regression if the regression is to have a causal interpretation. In the proxy-control scenario, however, your intentions are good. And while proxy control does not generate the regression coefficient of interest, it may be an improvement on no control at all. Recall that the motivation for proxy control is equation (3.2.13). In terms of the parameters in this model, the OVB formula tells us that a regression on S_i with no controls generates a coefficient of $\rho + \gamma\delta_{as}$, where δ_{as} is slope coefficient from a regression of a_i on S_i. The schooling coefficient in (3.2.15) might be closer to ρ than the coefficient you estimate with no control at all. Moreover, assuming δ_{as} is positive, you can safely say that the causal effect of interest lies between these two.

One moral of both the bad-control and the proxy-control stories is that when thinking about controls, timing matters. Variables measured before the variable of interest was determined are generally good controls. In particular, because these variables were determined before the variable of interest, they cannot themselves

be outcomes in the causal nexus. In many cases, however, the timing is uncertain or unknown. In such cases, clear reasoning about causal channels requires explicit assumptions about what happened first, or the assertion that none of the control variables are themselves caused by the regressor of interest.[17]

3.3 Heterogeneity and Nonlinearity

As we saw in the previous section, a linear causal model in combination with the CIA leads to a linear CEF with a causal interpretation. Assuming the CEF is linear, the population regression is it. In practice, however, the assumption of a linear CEF is not really necessary for a causal interpretation of regression. For one thing, as discussed in Section 3.1.2, we can think of the regression of Y_i on X_i and S_i as providing the best linear approximation to the underlying CEF, regardless of its shape. Therefore, if the CEF is causal, the fact that regression approximates it gives regression coefficients a causal flavor. This claim is a little vague, however, and the nature of the link between regression and the CEF is worth exploring further. This exploration leads us to an understanding of regression as a computationally attractive matching estimator.

3.3.1 Regression Meets Matching

The past decade or two has seen increasing interest in matching as an empirical tool. Matching as a strategy to control for covariates is typically motivated by the CIA, as for causal regression in the previous section. For example, Angrist (1998) used matching to estimate the effects of volunteering for the military service on the later earnings of soldiers. These matching estimates have a causal interpretation assuming that, conditional on the individual characteristics the military uses to select soldiers (age, schooling, test scores), veteran status is independent of potential earnings.

An attractive feature of matching strategies is that they are typically accompanied by an explicit statement of the conditional independence assumption required to give matching estimates a causal interpretation. At the same time, we have just seen that the causal interpretation of a regression coefficient is based on exactly the same assumption. In other words, matching and regression are both control strategies. Since the core assumption underlying causal inference is the same for the two strategies, it's worth asking whether or to what extent matching really differs from regression. Our view is that regression can be motivated as a computational device for a particular sort of weighted matching estimator, and therefore the differences between regression and matching are unlikely to be of major empirical importance.

To flesh out this idea, it helps to look more deeply into the mathematical structure of the matching and regressions *estimands*, i.e., the population quantities that these methods attempt to estimate. For regression, of course, the estimand is a vector of population regression coefficients. The matching estimand is typically

[17] Griliches and Mason (1972) is a seminal exploration of the use of early and late ability controls in schooling equations. See also Chamberlain (1977, 1978) for closely related studies. Rosenbaum (1984) offers an alternative discussion of the proxy control idea using very different notation, outside of a regression framework.