Morgan / Stephen L., Winship / Christopher (2007): Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press. ISBN: 978-0-521-67193-4

# Chapter 5

# Regression Estimators of Causal Effects

Regression models are perhaps the most common form of data analysis used to evaluate alternative explanations for outcomes of interest to quantitatively oriented social scientists. In the past 40 years, a remarkable variety of regression models have been developed by statisticians. Accordingly, most major data analysis software packages allow for regression estimation of the relationships between interval and categorical variables, in cross sections and longitudinal panels, and in nested and multilevel patterns. In this chapter, however, we restrict our attention to OLS regression, focusing mostly on the regression of an interval-scaled variable on a binary causal variable. As we will show, the issues are complicated enough for these models. And it is our knowledge of how least squares models work that allows us to explain the complexity.

In this chapter, we present least squares regression from three different perspectives: (1) regression as a descriptive modeling tool, (2) regression as a parametric adjustment technique for estimating causal effects, and (3) regression as a matching estimator of causal effects. We give more attention to the third of these three perspectives on regression than is customary in methodological texts because this perspective allows one to understand the others from a counterfactual perspective. At the end of the chapter, we will draw some of the connections between least squares regression and more general models, and we will discuss the estimation of causal effects for many-valued causes.

## 5.1 Regression as a Descriptive Tool

Least squares regression can be justified without reference to causality, as it can be considered nothing more than a method for obtaining a best-fitting descriptive model under entailed linearity constraints. Goldberger (1991), for example, motivates least squares regression as a technique to estimate a best-fitting linear approximation to a conditional expectation function that may be nonlinear in the population.

Consider this descriptive motivation of regression a bit more formally. If $X$ is a collection of variables that are thought to be associated with $Y$ in some way, then the conditional expectation function of $Y$, viewed as a function in $X$, is denoted $E[Y|X]$. Each particular value of the conditional expectation for a specific realization $x$ of $X$ is then denoted $E[Y|X = x]$.

Least squares regression yields a predicted surface $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is a vector of estimated coefficients from the regression of the realized values $y_i$ on $x_i$. The predicted surface, $X\hat{\beta}$, does not necessarily run through the specific points of the conditional expectation function, even for an infinite sample, because (1) the conditional expectation function may be a nonlinear function of one or more of the variables in $X$ and (2) a regression model can be fit without parameterizing all nonlinearities in $X$. An estimated regression surface simply represents a best-fitting linear approximation of $E[Y|X]$ under whatever linearity constraints are entailed by the chosen parameterization.[1]

The following demonstration of this usage of regression is simple. Most readers know this material well and can skip ahead to the next section. But, even so, it may be worthwhile to read the demonstration quickly because we will build directly on it when shifting to the consideration of regression as a causal effect estimator.

### Regression Demonstration 1

Recall the stratification example presented as Matching Demonstration 1 (see page 92 in Chapter 4). Suppose that the same data are being analyzed, as generated by the distributions presented in Tables 4.1 and 4.2; features of these distributions are reproduced in Table 5.1 in more compact form. As before, assume that well-defined causal states continue to exist and that $S$ serves as a perfect stratification of the data.[2] Accordingly, the conditional expectations in the last three panels of Table 5.1 are equal as shown.

But, for this demonstration of regression as a descriptive tool, assume that a cautious researcher does not wish to rush ahead and attempt to estimate the specific underlying causal effect of $D$ on $Y$, either averaged across all individuals or averaged across particular subsets of the population. Instead, the researcher is cautious and is willing to assert only that the variables $S$, $D$, and $Y$ constitute some portion of a larger system of causal relationships. In particular, the researcher is unwilling to assert anything about the existence or nonexistence of other variables that may also lie on the causal chain from $S$, through $D$, to $Y$. This is tantamount to doubting the claim that $S$ offers a perfect stratification of the data, even though that claim is true by construction for this example.

[1] One can fit a large variety of nonlinear surfaces with regression by artful parameterizations of the variables in $X$, but these surfaces are always generated by a linear combination of a coefficient vector and values on some well-defined coding of the variables in $X$.

[2] For this section, we will also stipulate that the conditional variances of the potential outcomes are constant across both of the potential outcomes and across levels of $S$.

Table 5.1: The Joint Probability Distribution and Conditional Population Expectations for Regression Demonstration 1

| | Joint probability distribution of $S$ and $D$ | |
|---|---|---|
| | Control group: $D = 0$ | Treatment group: $D = 1$ |
| $S = 1$ | $\Pr[S = 1, D = 0] = .36$ | $\Pr[S = 1, D = 1] = .08$ |
| $S = 2$ | $\Pr[S = 2, D = 0] = .12$ | $\Pr[S = 2, D = 1] = .12$ |
| $S = 3$ | $\Pr[S = 3, D = 0] = .12$ | $\Pr[S = 3, D = 1] = .2$ |
| | Potential outcomes under the control state | |
| $S = 1$ | $E[Y^0 \vert S = 1, D = 0] = 2$ | $E[Y^0 \vert S = 1, D = 1] = 2$ |
| $S = 2$ | $E[Y^0 \vert S = 2, D = 0] = 6$ | $E[Y^0 \vert S = 2, D = 1] = 6$ |
| $S = 3$ | $E[Y^0 \vert S = 3, D = 0] = 10$ | $E[Y^0 \vert S = 3, D = 1] = 10$ |
| | Potential outcomes under the treatment state | |
| $S = 1$ | $E[Y^1 \vert S = 1, D = 0] = 4$ | $E[Y^1 \vert S = 1, D = 1] = 4$ |
| $S = 2$ | $E[Y^1 \vert S = 2, D = 0] = 8$ | $E[Y^1 \vert S = 2, D = 1] = 8$ |
| $S = 3$ | $E[Y^1 \vert S = 3, D = 0] = 14$ | $E[Y^1 \vert S = 3, D = 1] = 14$ |
| | Observed outcomes | |
| $S = 1$ | $E[Y \vert S = 1, D = 0] = 2$ | $E[Y \vert S = 1, D = 1] = 4$ |
| $S = 2$ | $E[Y \vert S = 2, D = 0] = 6$ | $E[Y \vert S = 2, D = 1] = 8$ |
| $S = 3$ | $E[Y \vert S = 3, D = 0] = 10$ | $E[Y \vert S = 3, D = 1] = 14$ |

In this situation, suppose that the researcher simply wishes to estimate the best linear approximation to the conditional expectation $E[Y|D, S]$ and does not wish to then give a causal interpretation to any of the coefficients that define the linear approximation. The six true values of $E[Y|D, S]$ are given in the last panel of Table 5.1. Notice that the linearity of $E[Y|D, S]$ in $D$ and $S$ is present only when $S \leq 2$. The value of 14 for $E[Y|D = 1, S = 3]$ makes $E[Y|D, S]$ nonlinear in $D$ and $S$ over their full distributions.

Now consider the predicted surfaces that would result from the estimation of two alternative least squares regression equations with data from a sample of infinite size (to render sampling error zero). A regression of $Y$ on $D$ and $S$ that treats $D$ as a dummy variable and $S$ as an interval-scaled variable would yield a predictive surface of

$$\hat{Y} = -2.71 + 2.69(D) + 4.45(S). \tag{5.1}$$

This model constrains the partial association between $Y$ and $S$ to be linear. It represents a sensible predicted regression surface because it is a best-fitting, linear-in-the-parameters model of the association between $Y$ and the two variables $D$ and $S$, where "best" is defined as minimizing the average squared differences between the fitted values and the true values of the conditional expectation function.

For this example, one can offer a better descriptive fit at little interpretive cost by using a more flexible parameterization of $S$. An alternative regression that treats $S$ as a discrete variable represented in the estimation routine by dummy variables $S2$ and $S3$ (for $S$ equal to 2 and $S$ equal to 3, respectively) would yield a predictive surface of

$$\hat{Y} = 1.86 + 2.75(D) + 3.76(S2) + 8.92(S3). \tag{5.2}$$

Like the predicted surface for the model in Equation (5.1), this model is also a best linear approximation to the six values of the true conditional expectation $E[Y|D, S]$. The specific estimated values are

$$\begin{aligned}
D = 0, S = 1 &: \hat{Y} = 1.86,\\
D = 0, S = 2 &: \hat{Y} = 5.62,\\
D = 0, S = 3 &: \hat{Y} = 10.78,\\
D = 1, S = 1 &: \hat{Y} = 4.61,\\
D = 1, S = 2 &: \hat{Y} = 8.37,\\
D = 1, S = 3 &: \hat{Y} = 13.53.
\end{aligned}$$

In contrast to the model in Equation (5.1), for this model the variable $S$ is given a fully flexible coding. As a result, parameters are fit that uniquely represent all values of $S$.[3] The predicted change in $Y$ for a shift in $S$ from 1 to 2 is 3.76

[3] The difference between a model in which a variable is given a fully flexible coding and one in which it is given a more constrained coding is clearer for a simpler conditional expectation function. For $E[Y|S]$, consider the values in the cells of Table 5.1. The three values of $E[Y|S]$ can be obtained from the first and fourth panels of Table 5.1 as follows:

$$\begin{aligned}
E[Y|S = 1] &= \frac{.36}{(.36 + .08)}(2) + \frac{.08}{(.36 + .08)}(4) = 2.36,\\
E[Y|S = 2] &= \frac{.12}{(.12 + .12)}(6) + \frac{.12}{(.12 + .12)}(8) = 7,\\
E[Y|S = 3] &= \frac{.12}{(.12 + .2)}(10) + \frac{.2}{(.12 + .2)}(14) = 12.5.
\end{aligned}$$

Notice that these three values of $E[Y|S]$ do not fall on a straight line; the middle value of 7 is closer to 2.36 than it is to 12.5.

For $E[Y|S]$, a least squares regression of $Y$ on $S$, treating $S$ as an interval-scaled variable, would yield a predictive surface of

$$\hat{Y} = -2.78 + 5.05(S).$$

The three values of this estimated regression surface lie on a straight line −2.27, 7.32, and 12.37 – and they do not match the corresponding true values of 2.36, 7, and 12.5. A regression of $Y$ on $S$, treating $S$ as a discrete variable with dummy variables $S2$ and $S3$, would yield an alternative predictive surface of

$$\hat{Y} = 2.36 + 4.64(S2) + 10.14(S3).$$

(i.e., $5.62 - 1.86 = 3.76$ and $8.37 - 4.61 = 3.76$) whereas the predicted change in $Y$ for a shift in $S$ from 2 to 3 is 5.16 (i.e., $10.78 - 5.62 = 5.16$ and $13.53 - 8.37 = 5.16$).

Even so, the model in Equation (5.2) constrains the parameter for $D$ to be the same without regard to the value of $S$. And, because the level of $Y$ depends on the interaction of $S$ and $D$, specifying more than one parameter for the three values of $S$ does not bring the predicted regression surface into alignment with the six values of $E[Y|D, S]$ presented in the last panel of Table 5.1. Thus, even when $S$ is given a fully flexible coding (and even for an infinitely large sample), the fitted values do not equal the true values of $E[Y|D, S]$.[4] As we discuss later, a model that is saturated fully in both $S$ and $D$ – that is, one that adds two additional parameters for the interactions between $D$ and both $S2$ and $S3$ – would yield predicted values that would exactly match the six true values of $E[Y|D, S]$ in a dataset of sufficient size.

Recall the more general statement of the descriptive motivation of regression analysis presented earlier, in which the predicted surface $\hat{Y} = X\hat{\beta}$ is estimated for the sole purpose of obtaining a best-fitting linear approximation to the true conditional expectation function $E[Y|X]$. When the purposes of regression are so narrowly restricted, the outcome variable of interest, $Y$, is not generally thought to be a function of potential outcomes associated with well-defined causal states. Consequently, it would be inappropriate to give a causal interpretation to any of the estimated coefficients in $\hat{\beta}$.

This perspective implies that if one were to add more variables to the predictors, embedding $X$ in a more encompassing set of variables $W$, then a new set of least squares estimates $\hat{\gamma}$ could be obtained by regressing $Y$ on $W$. The estimated surface $W\hat{\gamma}$ then represents a best-fitting, linear-in-the-parameters, descriptive fit to a more encompassing conditional expectation function, $E[Y|W]$. Whether one then prefers $W\hat{\gamma}$ to $X\hat{\beta}$ as a description of the variation in $Y$ depends on whether one finds it more useful to approximate $E[Y|W]$ than $E[Y|X]$. The former regression approximation is often referred to as the long regression, with the latter representing the short regression. These labels are aptly chosen, when regression is considered nothing more than a descriptive tool, as there is no inherent reason to prefer a short to a long regression if neither is meant to

This second model uses a fully flexible coding of $S$, and each value of the conditional expectation function is a unique function of the parameters in the model (that is, $2.36 = 2.36$, $4.64 + 2.36 = 7$, and $10.14 + 2.36 = 12.5$). Thus, in this case, the regression model would, in a suitably large sample, estimate the three values of $E[Y|S]$ exactly.

[4] Why would one ever prefer a constrained regression model of this sort? Consider a conditional expectation function, $E[Y|X]$, where $Y$ is earnings and $X$ is years of education (with 21 values from 0 to 20). A fully flexible coding of $X$ would fit 20 dummy variables for the 21 values of $X$. This would allow the predicted surface to change only modestly between some years (such as between 7 and 8 and between 12 and 13) and more dramatically between other years (such as between 11 and 12 and between 15 and 16). However, one might wish to treat $X$ as an interval-scaled variable, smoothing these increases from year to year by constraining them to a best-fitting line parameterized only by an intercept and a constant slope. This constrained model would not fit the conditional expectation function as closely as the model with 20 dummy variables, but it might be preferred in some situations because it is easier to present and easier to estimate for a relatively small sample.

be interpreted as anything other than a best-fitting linear approximation to its respective true conditional expectation function.

In many applied regression textbooks, the descriptive motivation of regression receives no direct explication. And, in fact, many textbooks state that the only correct specification of a regression model is one that includes all explanatory variables. Goldberger (1991) admonishes such textbook writers, countering their claims with:

> An alternative position is less stringent and is free of causal language. Nothing in the CR [classical regression] model itself requires an exhaustive list of explanatory variables, nor any assumption about the direction of causality. (Goldberger 1991:173)

Goldberger is surely correct, but his perspective nonetheless begs an important question on the ultimate utility of descriptively motivated regression. Clearly, if one wishes to know only predicted values of the outcome $Y$ for those not originally studied but whose variables in $X$ are known, then being able to form the surface $X\hat{\beta}$ is a good first step (and perhaps a good last step). And, if one wishes to build a more elaborate regression model, allowing for an additional variable in $W$ or explicitly accounting for multilevel variability by modeling the nested structure of the data, then regression results will be useful if the aim is merely to generate descriptive reductions of the data. But, if one wishes to know the value of $Y$ that would result for any individual in the population if a variable in $X$ were shifted from a value $k$ to a value $k'$, then regression results may be uninformative.

Many researchers (perhaps a clear majority) who use regression models in their research are very much interested in causal effects. Knowing the interests of their readers, many textbook presentations of regression sidestep these issues artfully by, for example, discussing how biased regression coefficients result from the omission of important explanatory variables but without introducing explicit, formal notions of causality into their presentations. Draper and Smith (1998:236), for example, write of the bias that enters into estimated regression coefficients when only a subset of the variables in the "true response relationship" are included in the fitted model. Similarly, Greene (2000:334) writes of the same form of bias that results from estimating coefficients for a subset of the variables from the "correctly specified regression model."[5] And, in his presentation of regression models for social scientists, Stolzenberg (2004:188) equivocates:

> Philosophical arguments about the nature of causation notwithstanding (see Holland, 1986), in most social science uses of regression, the *effect* of an independent variable on a dependent variable is the *rate* at which differences in the independent variable are associated with (or cause) differences or changes in the dependent variable. [Italics in the original.]

[5] There are, of course, other textbooks that do present a more complete perspective, such as Berk (2004), Freedman (2005), and Gelman and Hill (2007).

We also assume that the readers of our book are interested in causal effect estimators. And thus, although we recognize the classical regression tradition, perhaps best defended by Goldberger (1991) as interpretable merely as a descriptive data reduction tool, we will consider regression primarily as a causal effect estimator in the following sections of this chapter. And we further note that, in spite of our reference to Goldberger (1991), in other writing Goldberger has made it absolutely clear that he too was very much interested in the proper usage of regression models to offer warranted causal claims. This is perhaps most clear in work in which he criticized what he regarded as unwarranted causal claims generated by others using regression techniques, such as in his robust critique of Coleman's Catholic schools research that we summarized in Subsection 1.3.2 (see Goldberger and Cain 1982). We will return to a discussion of the notion of a correct specification of a regression model at the end of the chapter, where we discuss the connections between theoretical models and regressions as all-cause perfect specifications. Until then, however, we return to the same basic scenario considered in our presentation of matching in Chapter 4: the estimation of a single causal effect that may be confounded by other variables.

## 5.2 Regression Adjustment as a Strategy to Estimate Causal Effects

In this section, we consider the estimation of causal effects in which least squares regression is used to adjust for variables thought to be correlated with both the causal and the outcome variables. We first consider the textbook treatment of the concept of omitted-variable bias, with which most readers are probably well acquainted. Thereafter, we consider the same set of ideas after specifying the potential outcome variables that the counterfactual tradition assumes lie beneath the observed data.

### 5.2.1 Regression Models and Omitted-Variable Bias

Suppose that one is interested in estimating the causal effect of a binary variable $D$ on an observed outcome $Y$. This goal can be motivated as an attempt to obtain an unbiased and consistent estimate of a coefficient $\delta$ in a generic bivariate regression equation:

$$Y = \alpha + \delta D + \varepsilon, \tag{5.3}$$

where $\alpha$ is an intercept and $\varepsilon$ is a summary random variable that represents all other causes of $Y$ (some of which may be related to the causal variable of interest, $D$). When Equation (5.3) is used to represent the causal effect of $D$ on $Y$ without any reference to individual-varying potential outcome variables, the parameter $\delta$ is implicitly cast as an invariant, structural causal effect that applies to all members of the population of interest.[6]

[6] Although this is generally the case, there are of course introductions to regression that explicitly define $\delta$ as the mean effect of $D$ on $Y$ across units in the population of interest or,

The OLS estimator of this bivariate regression coefficient is then:

$$\hat{\delta}_{\text{OLS, bivariate}} \equiv \frac{\text{Cov}_N(y_i,d_i)}{\text{Var}_N(d_i)}, \tag{5.4}$$

where $\text{Cov}_N(.)$ and $\text{Var}_N(.)$ are unbiased, sample-based estimates from a sample of size $N$ of the population-level covariance and variance of the variables that are their arguments.[7] Because $D$ is a binary variable, $\hat{\delta}_{\text{OLS, bivariate}}$ is exactly equivalent to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$, presented earlier in Equation (2.7) (i.e., sample mean of $y_i$ for those in the treatment group minus the sample mean of $y_i$ for those in the control group). Our analysis thus follows quite closely the prior discussion of the naive estimator in Subsection 2.6.3. The difference is that here we will develop the same basic claims with reference to the relationship between $D$ and $\varepsilon$ rather than the general implications of heterogeneity of the causal effect.

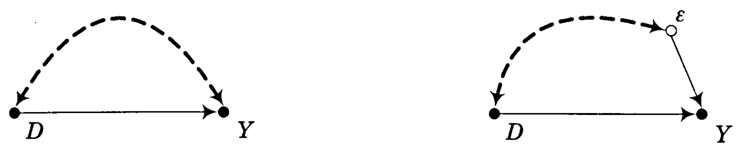
Consider first a case in which $D$ is randomly assigned, as when individuals are randomly assigned to the treatment and control groups. In this case, $D$ would be uncorrelated with $\varepsilon$ in Equation (5.3), even though there may be a chance correlation between $D$ and $\varepsilon$ in any finite set of study subjects.[8] The literature on regression, when presented as a causal effect estimator, maintains that, in this case, (1) the estimator $\hat{\delta}_{\text{OLS, bivariate}}$ is unbiased and consistent for $\delta$ in Equation (5.3) and (2) $\delta$ can be interpreted as the causal effect of $D$ on $Y$.

To understand this claim, it is best to consider a counterexample in which $D$ is correlated with $\varepsilon$ in the population because $D$ is correlated with other causes of $Y$ that are implicitly embedded in $\varepsilon$. For a familiar example, consider again the effect of education on earnings. Individuals are not randomly assigned to the treatment "completed a bachelor's degree." It is generally thought that those who complete college would be more likely to have had high levels of earnings

as was noted in the last section, without regard to causality at all.

[7] Notice that we are again focusing on the essential features of the methods, and thus we maintain our perfect measurement assumption (which allows us to avoid talking about measurement error in $D$ or in $Y$, the latter of which would be embedded in $\varepsilon$). We also ignore degree-of-freedom adjustments because we assume that the available sample is again large. To be more precise, of course, we would want to indicate that the sample variance of $D$ does not equal the population-level variance of $D$ in the absence of such a degree-of-freedom adjustment, and so on. We merely label $\text{Var}_N(.)$ as signifying such an unbiased estimate of the population-level-variance of that which is its argument. Thus, $\text{Var}_N(.)$ implicitly includes the proper degree-of-freedom adjustment, which would be $N/(N-1)$ and which would then be multiplied by the average of squared deviations from the sample mean.

[8] We will frequently refer to $D$ and $\varepsilon$ as being uncorrelated for this type of assumption, as this is the semantics that most social scientists seem to use and understand when discussing these issues. Most textbook presentations of regression discuss very specific exogeneity assumptions for $D$ that imply a correlation of 0 between $D$ and $\varepsilon$. Usually, in the social sciences, the assumption is defined either by mean independence of $D$ and $\varepsilon$ or as an assumed covariance of 0 between $D$ and $\varepsilon$. Both of these imply a correlation between $D$ and $\varepsilon$ of 0. In statistics, one often finds a stronger assumption: $D$ and $\varepsilon$ must be completely independent of each other. The argument in favor of this stronger assumption, which is convincing to statisticians, is that an inference is strongest when it holds under any transformation of $Y$ (and thus any transformation of $\varepsilon$). When full independence of $D$ and $\varepsilon$ holds, mean independence of $D$ and $\varepsilon$, a covariance of 0 between $D$ and $\varepsilon$, and a 0 correlation between $D$ and $\varepsilon$ are all implied.



(a) A graph in which the causal effect of $D$ on $Y$ is unidentified

(b) An equivalent regression representation of the same graph

Figure 5.1: Graphs for a regression equation of the causal effect of $D$ on $Y$.

in the absence of a college education. If this is true, $D$ and the population-level error term $\varepsilon$ are correlated because those who have a 1 on $D$ are more likely to have high values rather than low values for $\varepsilon$. For this example, the least squares regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) would not yield a consistent and unbiased estimate of $\delta$ that can be regarded as an unbiased and consistent estimate of the causal effect of $D$ on $Y$. Instead, $\hat{\delta}_{\text{OLS,bivariate}}$ must be interpreted as an upwardly biased and inconsistent estimate of the causal effect of $D$ on $Y$. In the substance of the college-degree example, $\hat{\delta}_{\text{OLS, bivariate}}$ would be a poor estimate of the causal effect of a college degree on earnings, as it would suggest that the effect of obtaining a college degree is larger than it really is.[9]

Figure 5.1 presents two causal graphs. In panel (a), $D$ and $Y$ are connected by two types of paths, the direct causal effect $D \rightarrow Y$ and an unspecified number of back-door paths signified by $D \dashleftarrow\dashrightarrow Y$. (Recall that bidirected edges $\dashleftarrow\dashrightarrow$ represent an unspecified number of common causes of the two variables that they connect.) For the graph in panel (a), the causal effect of $D$ on $Y$ is unidentified because no observable variables block the back-door paths represented by $D \dashleftarrow\dashrightarrow Y$.

The graph in panel (b) is the regression analog to the causal graph in panel (a). It contains three edges: $D \rightarrow Y$, $\varepsilon \rightarrow Y$, and $D \dashleftarrow\dashrightarrow \varepsilon$, where the node for $\varepsilon$ is represented by a hollow circle ○ rather than a solid circle ● in order to indicate that $\varepsilon$ is an unobserved variable. The back-door paths from $D$ to $Y$ now run through the error term $\varepsilon$, and the dependence represented by the bidirected edge

[9] Consider for one last time the alternative and permissible descriptive interpretation: The least squares regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) could be interpreted as an unbiased and consistent estimate of $\delta$, in which the regression surface generated by the estimation of $\delta$ in Equation (5.3) can be interpreted as only a descriptively motivated, best linear prediction of the conditional expectation function, $E[Y|D]$ (i.e., where $\hat{\alpha}$ is an unbiased and consistent estimate of $E[Y|D=0]$ and $\hat{\alpha}+\hat{\delta}$ *is* an unbiased and consistent estimate of $E[Y|D=1]$). And, in the substance of the college-degree example, it could be regarded as an efficient estimate of the mean difference between the earnings of those who have obtained a college degree and those who have not. For this second type of interpretation, see the last section of this chapter.

contaminates the bivariate least squares regression coefficient for the regression of $Y$ on $D$. Bivariate regression results, when interpreted as warranted causal effect estimates, assume that there are no unblocked back-door paths from the causal variable to the outcome variable.

For many applications in the social sciences, a correlation between $D$ and $\varepsilon$ is conceptualized as a problem of omitted variables. For the example in this section, a bivariate OLS estimate of the effect of a college degree on labor market earnings would be said to be biased because intelligence is unobserved but is correlated with both education and earnings. Its omission from Equation (5.3) leads the estimate of the effect of a college degree on earnings from that equation to be larger than it would have been if a variable for intelligence were instead included in the equation.

This perspective, however, has led to much confusion, especially in cases in which a correlation between $D$ and $\varepsilon$ emerges because subjects choose different levels of $D$ based on their expectations about the variability of $Y$, and hence their own expectations of the causal effect itself. For example, those who attend college may be more likely to benefit from college than those who do not, even independent of the unobserved ability factor. Although this latent form of anticipation can be labeled an omitted variable, it is generally not. Instead, the language of research shifts toward notions such as self-selection bias, and this is less comfortable territory for the typical applied researcher.

To clarify the connections between omitted-variable bias and self-selection bias within a more general presentation, we draw on the counterfactual model in the next section. We break the error term in Equation (5.3) into component pieces defined by underlying potential outcome variables and allow for the more general forms of causal effect heterogeneity that are implicitly ruled out by constant-coefficient models.

### 5.2.2 Potential Outcomes and Omitted-Variable Bias

Consider the same set of ideas but now use the potential outcome framework to define the observable variables. We build directly on the variant of the counterfactual model presented in Subsection 3.2.2. From that presentation, recall Equation (3.5), which we reintroduce here as

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{v^0 + D(v^1 - v^0)\}, \tag{5.5}$$

where $\mu^0 \equiv E[Y^0]$, $\mu^1 \equiv E[Y^1]$, $v^0 \equiv Y^0 - E[Y^0]$, and $v^1 \equiv Y^1 - E[Y^1]$. We could rewrite this equation to bring it into closer alignment with Equation (5.3) by stipulating that $\alpha = \mu^0$, $\delta = (\mu^1 - \mu^0)$, and $\varepsilon = v^0 + D(v^1 - v^0)$. But note that this is not what is typically meant by the terms $\alpha$, $\delta$, and $\varepsilon$ in Equation (5.3). The parameters $\alpha$ and $\delta$ in Equation (5.3) are not considered to be equal to $E[Y^0]$ or $E[\delta]$ for two reasons: (1) models are usually asserted in the regression tradition (e.g., in Draper and Smith 1998) without any reference to underlying causal states tied to potential outcomes and (2) the parameters $\alpha$ and $\delta$ are usually implicitly held to be constant structural effects that do not vary over individuals in the population. Similarly, the error term, $\varepsilon$, in Equation (5.3) is not separated into two pieces as a function of the definition of potential outcomes and their relationship to $D$. For these reasons, Equation (5.5) is quite different from the traditional bivariate regression in Equation (5.3), in the sense that it is more finely articulated but also irretrievably tied to a particular formalization of a causal effect.

Suppose that we are interested in estimating the average treatment effect, denoted $(\mu^1 - \mu^0)$ here. $D$ could be correlated with the population-level variant of the error term $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ in Equation (5.5) in two ways. First, suppose that there is a net baseline difference in the hypothetical no-treatment state that is correlated with membership in the treatment group, but the size of the individual-level treatment effect does not differ on average between those in the treatment group and those in the control group. In this case, $\upsilon^0$ would be correlated with $D$, generating a correlation between $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ and $D$, even though the $D(\upsilon^1 - \upsilon^0)$ term in $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ would be equal to zero on average because $\upsilon^1 - \upsilon^0$ does not vary with $D$. Second, suppose there is a net treatment effect difference that is correlated with membership in the treatment group, but there is no net baseline difference in the absence of treatment. Now, $D(\upsilon^1 - \upsilon^0)$ would be correlated with $D$, even though $\upsilon^0$ is not, because the average difference in $\upsilon^1 - \upsilon^0$ varies across those in the treatment group and those in the control group. In either case, an OLS regression of the realized values of $Y$ on $D$ would yield a biased and inconsistent estimate of $(\mu^1 - \mu^0)$.

It may be helpful to see precisely how these sorts of bias come about with reference to the potential outcomes of individuals. Table 5.2 presents three simple two-person examples in which the least squares bivariate regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) is biased. Each panel presents the potential outcome values for two individuals and then the implied observed data and error term in the braces from Equation (5.5). Assume for convenience that there are only two types of individuals in the population, both of which are homogeneous with respect to the outcomes under study and both of which comprise one half of the population. For the three examples in Table 5.2, we have sampled one of each of these two types of individuals for study.

For the example in the first panel, the true average treatment effect is 15, because for the individual in the treatment group $\delta_i$ is 10 whereas for the individual in the control group $\delta_i$ is 20. The values of $\upsilon_i^1$ and $\upsilon_i^0$ are deviations of the values of $y_i^1$ and $y_i^0$ from $E[Y^1]$ and $E[Y^0]$, respectively. Because these expectations are equal to 20 and 5, the values of $\upsilon_i^1$ are both equal to 0 because each individual's value of $y_i^1$ is equal to 20. In contrast, the values of $\upsilon_i^0$ are equal to 5 and $-5$ for the individuals in the treatment and control groups, respectively, because their two values of $y_i^0$ are 10 and 0.

As noted earlier, the bivariate regression estimate of the coefficient on $D$ is equal to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$. Accordingly, a regression of the values for $y_i$ on $d_i$ would yield a value of 0 for the intercept and a value of 20 for the coefficient on $D$. This estimated value of 20 is an upwardly biased estimate for the true average causal effect because the values of $d_i$ are positively correlated with the values of the error term $\upsilon_i^0 + d_i(\upsilon_i^1 - \upsilon_i^0)$. In this

Table 5.2: Examples of the Two Basic Forms of Bias for Least Squares Regression

Differential baseline bias only

| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
|---|---|---|---|---|---|---|---|
| In treatment group | 20 | 10 | 0 | 5 | 20 | 1 | 0 |
| In control group | 20 | 0 | 0 | −5 | 0 | 0 | −5 |

Differential treatment effect bias only

| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
|---|---|---|---|---|---|---|---|
| In treatment group | 20 | 10 | 2.5 | 0 | 20 | 1 | 2.5 |
| In control group | 15 | 10 | −2.5 | 0 | 10 | 0 | 0 |

Both types of bias

| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
|---|---|---|---|---|---|---|---|
| In treatment group | 25 | 5 | 5 | −2.5 | 25 | 1 | 5 |
| In control group | 15 | 10 | −5 | 2.5 | 10 | 0 | 2.5 |

case, the individual with a value of 1 for $d_i$ has a value of 0 for the error term whereas the individual with a value of 0 for $d_i$ has a value of −5 for the error term.

For the example in the second panel, the relevant difference between the individual in the treatment group and the individual in the control group is in the value of $y_i^1$ rather than $y_i^0$. In this variant, both individuals would have had the same outcome if they were both in the control state, but the individual in the treatment group would benefit relatively more from being in the treatment state. Consequently, the values of $d_i$ are correlated with the values of the error term in the last column because the true treatment effect is larger for the individual in the treatment group than for the individual in the control group. A bivariate regression would yield an estimate of 10 for the average causal effect, even though the true average causal effect is only 7.5 in this case.

Finally, in the third panel of the table, both forms of baseline and net treatment effect bias are present, and in opposite directions. In combination, however, they still generate a positive correlation between the values of $d_i$ and the error term in the last column. This pattern results in a bivariate regression estimate of 15, which is upwardly biased for the true average causal effect of 12.5.

For symmetry, and some additional insight, now consider two additional two-person examples in which regression gives an unbiased estimate of the average causal effect. For the first panel of Table 5.3, the potential outcomes are independent of $D$, and as a result a bivariate regression of the values $y_i$ on $d_i$ would

Table 5.3: Two-Person Examples in Which Least Squares Regression Estimates are Unbiased

| | Independence of $(Y^1, Y^0)$ from $D$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
| In treatment group | 20 | 10 | 0 | 0 | 20 | 1 | 0 |
| In control group | 20 | 10 | 0 | 0 | 10 | 0 | 0 |
| | Offsetting dependence of $Y^1$and $Y^0$ on $D$ | | | | | | |
| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
| In treatment group | 20 | 10 | 5 | $-5$ | 20 | 1 | 5 |
| In control group | 10 | 20 | $-5$ | 5 | 20 | 0 | 5 |

yield an unbiased estimate of 10 for the true average causal effect. But the example in the second panel is quite different. Here, the values of $v_i^1$ and $v_i^0$ are each correlated with the values of $d_i$, but they cancel each other out when they jointly constitute the error term in the final column. Thus, a bivariate regression yields an unbiased estimate of 0 for the true average causal effect of 0. And, yet, with knowledge of the values for $y_i^1$ and $y_i^0$, it is clear that these results mask important heterogeneity of the causal effect. Even though the average causal effect is indeed 0, the individual-level causal effects are equal to 10 and $-10$ for the individuals in the treatment group and control group, respectively. Thus, regression gives the right answer, but it hides the underlying heterogeneity that one would almost certainly wish to know.

Having considered these examples, we are now in a position to answer, from within the counterfactual framework, the question that so often confounds students when first introduced to regression as a causal effect estimator: What is the error term of a regression equation? Compare the third and fourth columns with the final column in Tables 5.2 and 5.3. The regression error term, $v^0 + D(v^1 - v^0)$, is equal to $v^0$ for those in the control group and $v^1$ for those in the treatment group. This can be seen without reference to the examples in the tables. Simply rearrange $v^0 + D(v^1 - v^0)$ as $Dv^1 + (1 - D)v^0$ and then rewrite Equation (5.5) as

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{Dv^1 + (1 - D)v^0\}. \tag{5.6}$$

It should be clear that the error term now appears very much like the observability of $Y$ definition presented earlier as $DY^1 + (1 - D)Y^0$ in Equation (2.2). Just as $Y$ switches between $Y^1$ and $Y^0$ as a function of $D$, the error term switches between $v^1$ and $v^0$ as a function of $D$. Given that $v^1$ and $v^0$ can be interpreted as $Y^1$ and $Y^0$ centered around their respective population-level expectations $E[Y^1]$ and $E[Y^0]$, this should not be surprising.

Even so, few presentations of regression characterize the error term of a bivariate regression in this way. Some notable exceptions do exist. The connection is made to the counterfactual tradition by specifying Equation (5.3) as

$$Y = \alpha + \delta D + \varepsilon_{(D)}, \tag{5.7}$$

where the error term $\varepsilon_{(D)}$ is considered to be an entirely different random variable for each value of $D$ (see Pratt and Schlaifer 1988). Consequently, the error term $\varepsilon$ in Equation (5.3) switches between $\varepsilon_{(1)}$ and $\varepsilon_{(0)}$ in Equation (5.7) depending on whether $D$ is equal to 1 or 0.[10]

Before moving on to adjustment techniques, it seems proper to ask one final question. If both $\upsilon^1$ and $\upsilon^0$ are uncorrelated with $D$, will the bivariate least squares regression coefficient for $D$ be an unbiased and consistent estimate of the average causal effect? Yes, but two qualifications should be noted, both of which were revealed in the second example in Table 5.3. First, bivariate regression can yield an unbiased and consistent estimate in other cases, as when the nonzero correlations that $\upsilon^1$ and $\upsilon^0$ have with $D$ "cancel out" in the construction of the combined error term $D\upsilon^1 + (1 - D)\upsilon^0$. Second, an unbiased and consistent regression estimate of the average causal effect may still mask important heterogeneity of causal effects. The first of these qualifications would rarely apply to real-world applications, but the second qualification, we suspect, obtains widely and is less frequently recognized than it should be.

### 5.2.3 Regression as Adjustment for Otherwise Omitted Variables

How well can regression adjust for an omitted variable if that variable is observed and included in an expanded regression equation? The basic strategy behind regression analysis as an adjustment technique to estimate a causal effect is to add a sufficient set of "control variables" to the bivariate regression in Equation (5.3) in order to break a correlation between the treatment variable $D$ and the error term $\varepsilon$, as in

$$Y = \alpha + \delta D + X\beta + \varepsilon^*, \tag{5.8}$$

where $X$ represents one or more variables, $\beta$ is a coefficient (or a conformable vector of coefficients if $X$ represents more than one variable), $\varepsilon^*$ is a residualized version of the original error term $\varepsilon$ from Equation (5.3), and all else is as defined for Equation (5.3).

For the multiple regression analog to the least squares bivariate regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4), the observed data values $d_i$ and $x_i$ are embedded in an all-encompassing $\mathbf{Q}$ matrix, which is $N \times K$, where $N$ is the number of respondents and $K$ is the number of variables in $X$ plus 2 (one

[10] This is the same approach taken by Freedman (see Berk 2004, Freedman 2005), and he refers to Equation (5.7) as a response schedule. See also the discussion of Sobel (1995). For a continuous variable, Garen (1984) notes that there would be an infinite number of error terms (see discussion of Garen's Equation 10).

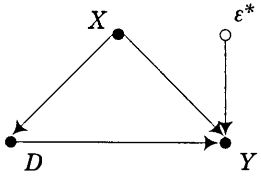


Figure 5.2: A causal graph for a regression equation in which the causal effect of $D$ on $Y$ is identified by conditioning on $X$.

for the constant and one for the treatment variable $D$). The OLS estimator for the parameters in Equation (5.8) is then written in matrix notation as

$$\hat{\delta}_{\text{OLS, multiple}} \equiv (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}, \tag{5.9}$$

where $\mathbf{y}$ is an $N \times 1$ vector for the observed outcomes $y_i$. As all regression textbooks show, there is nothing magical about these least squares computations, even though the matrix representation may appear unfamiliar to some readers. OLS regression is equivalent to the following three-step regression procedure with reference to Equation (5.8) [and without reference to the perhaps overly compact Equation (5.9)]: (1) Regress $y_i$ on $x_i$ and calculate $y_i^* = y_i - \hat{y}_i$; (2) regress $d_i$ on $x_i$ and calculate $d_i^* = d_i - \hat{d}_i$; (3) regress $y_i^*$ on $d_i^*$. The regression coefficient on $d_i^*$ yielded by step (3) is the OLS estimate of $\delta$, which is typically declared unbiased and consistent for $\delta$ in Equation (5.8) if the true correlation between $D$ and $\varepsilon^*$ is assumed to be equal to zero. Thus, in this simple example, OLS regression is equivalent to estimating the relationship between residualized versions of $Y$ and $D$ from which their common dependence on other variables in $X$ has been "subtracted out."

Even though the variables in $X$ might be labeled control variables in a regression analysis of a causal effect, this label expresses the intent rather than the outcome of their utilization. The goal of such a regression adjustment strategy is to find variables in $X$ that can be used to redraw the causal graph in panel (b) of Figure 5.1 as the DAG in Figure 5.2. If this can be done, then one can condition on $X$ in order to consistently estimate the causal effect of $D$ on $Y$ because $X$ blocks the only back-door path between $D$ and $Y$.

If $D$ is uncorrelated with $\varepsilon^*$ (i.e., the error term net of adjustment for $X$), then least squares regression yields an estimate that is ostensibly freed of the bias generated by the correlation of the treatment $D$ with the error term $\varepsilon$ in Equation (5.3). However, even in this case some complications remain when one invokes the potential outcome model.

First, if one assumes that $\delta$ is truly constant across individuals (i.e., that $y_i^1 - y_i^0$ is equal to the same constant for all individuals $i$), then the OLS estimate is unbiased and consistent for $\delta$ and for $(\mu^1 - \mu^0)$. If, however, $y_i^1 - y_i^0$ is not constant, then the OLS estimate represents a conditional-variance-weighted estimate of the underlying causal effects of individuals, $\delta_i$, in which the weights are a function of the conditional variance of $D$ (see Angrist 1998, as well as our explanation of this result in the next section). Under these conditions, the OLS estimate is unbiased and consistent for this particular weighted average, which is usually not a causal parameter of interest.

Second, note that the residualized error term, $\varepsilon^*$, in Equation (5.8) is not equivalent to either $\varepsilon$ from Equation (5.3) or to the multipart error term $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ from Equation (5.5). Rather, it is defined by whatever adjustment occurs within Equation (5.8), as represented by the term $X\beta$. Consequently, the residualized error term $\varepsilon^*$ cannot be interpreted independently of decisions about how to specify the vector of adjustment variables in $X$, and this can make it difficult to define when a net covariance between $D$ and $\varepsilon^*$ can be assumed to be zero.

We explain these two complications and their important implications in the following sections of this chapter, where we consider a variety of examples that demonstrate the connections between matching and regression estimators of causal effects. Before developing these explanations, however, we conclude this section with two final small-$N$ examples that demonstrate how the regression adjustment strategy does and does not work.

Table 5.4 presents two six-person examples. For both examples, a regression of $Y$ on $D$ yields a biased estimate of the true average treatment effect. And, in fact, both examples yield the same biased estimate because the observed values $y_i$ and $d_i$ are the same for both examples. Moreover, an adjustment variable $X$ is also available for both examples, and its observed values $x_i$ have the same associations with the observed values $y_i$ and $d_i$ for both examples. But the underlying potential outcomes differ substantially between the two examples. These differences render regression adjustment by $X$ effective for only the first example.

For the example in the first panel, a regression of $Y$ on $D$ would yield an estimate of the coefficient for $D$ of 11.67, which is an upwardly biased estimate of the true average causal effect of 10. The bias arises because the correlation between the error term in the last column and the realized values for $d_i$ is not zero but is instead .33.

For the example in the second panel, a regression of $Y$ on $D$ would yield an estimate of the coefficient for $D$ of 11.67 because the values for $y_i$ and $d_i$ are exactly the same as for the example in the first panel. Moreover, this estimate is also upwardly biased because the error term in the last column is positively correlated with the realized values of $d_i$. However, here the patterns are more complex. The underlying potential outcomes are different, and individual-level heterogeneity of the causal effect is now present. One member of the control group has an individual-level treatment effect of only 8, and as a result the true average treatment effect is only 9.67. Consequently, the same bivariate regression coefficient of 11.67 has a larger upward bias in this second example, and the correlation between the values of $d_i$ and the error term in the last column is now .39 rather than .33.[11]

[11] Moreover, the correlation between the values of $d_i$ and both $\upsilon_i^1$ and $\upsilon_i^0$ differs, with the former generating a correlation coefficient of .51 and the latter generating a correlation coefficient of .33.

Table 5.4: Two Six-Person Examples in Which Regression Adjustment is Differentially Effective

Regression adjustment with $X$ generates an unbiased estimate for $D$

| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $x_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
|---|---|---|---|---|---|---|---|---|
| In treatment group | 20 | 10 | 2.5 | 2.5 | 20 | 1 | 1 | 2.5 |
| In treatment group | 20 | 10 | 2.5 | 2.5 | 20 | 1 | 1 | 2.5 |
| In treatment group | 15 | 5 | −2.5 | −2.5 | 15 | 1 | 0 | −2.5 |
| In control group | 20 | 10 | 2.5 | 2.5 | 10 | 0 | 1 | 2.5 |
| In control group | 15 | 5 | −2.5 | −2.5 | 5 | 0 | 0 | −2.5 |
| In control group | 15 | 5 | −2.5 | −2.5 | 5 | 0 | 0 | −2.5 |

Regression adjustment with $X$ does not generate an unbiased estimate for $D$

| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $x_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
|---|---|---|---|---|---|---|---|---|
| In treatment group | 20 | 10 | 2.83 | 2.5 | 20 | 1 | 1 | 2.83 |
| In treatment group | 20 | 10 | 2.83 | 2.5 | 20 | 1 | 1 | 2.83 |
| In treatment group | 15 | 5 | −2.17 | −2.5 | 15 | 1 | 0 | −2.17 |
| In control group | 18 | 10 | .83 | 2.5 | 10 | 0 | 1 | 2.5 |
| In control group | 15 | 5 | −2.17 | −2.5 | 5 | 0 | 0 | −2.5 |
| In control group | 15 | 5 | −2.17 | −2.5 | 5 | 0 | 0 | −2.5 |

This underlying difference in potential outcomes also has consequences for the capacity of regression adjustment to effectively generate unbiased estimates of the average treatment effect. This is easiest to see by rearranging the rows in Table 5.4 for each of the two examples based on the values of $X$ for each individual, as in Table 5.5. For the first example, the values of $d_i$ are uncorrelated with the error term within subsets of individuals defined by the two values of $X$. In contrast, for the second example, the values of $d_i$ remain positively correlated with the error term within subsets of individuals defined by the two values of $X$. Thus, conditioning on $X$ breaks the correlation between $D$ and the error term in the first example but not in the second example. Because the observed data are the same for both examples, this difference is entirely a function of the underlying potential outcomes that generate the data.

This example demonstrates an important conceptual point. Recall that the basic strategy behind regression analysis as an adjustment technique is to estimate

$$Y = \alpha + \delta D + X\beta + \varepsilon^*,$$

Table 5.5: A Rearrangement to Show How Regression Adjustment is Differentially Effective

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Regression adjustment with $X$ generates an unbiased estimate for $D$ | | | | | | | |
| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $x_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
| | For those with $X=1$ | | | | | | | |
| In treatment group | 20 | 10 | 2.5 | 2.5 | 20 | 1 | 1 | 2.5 |
| In treatment group | 20 | 10 | 2.5 | 2.5 | 20 | 1 | 1 | 2.5 |
| In control group | 20 | 10 | 2.5 | 2.5 | 10 | 0 | 1 | 2.5 |
| | For those with $X=0$ | | | | | | | |
| In treatment group | 15 | 5 | −2.5 | −2.5 | 15 | 1 | 0 | −2.5 |
| In control group | 15 | 5 | −2.5 | −2.5 | 5 | 0 | 0 | −2.5 |
| In control group | 15 | 5 | −2.5 | −2.5 | 5 | 0 | 0 | −2.5 |
| | Regression adjustment with $X$ does not generate an unbiased estimate for $D$ | | | | | | | |
| | $y_i^1$ | $y_i^0$ | $v_i^1$ | $v_i^0$ | $y_i$ | $d_i$ | $x_i$ | $v_i^0+d_i(v_i^1-v_i^0)$ |
| | For those with $X=1$ | | | | | | | |
| In treatment group | 20 | 10 | 2.83 | 2.5 | 20 | 1 | 1 | 2.83 |
| In treatment group | 20 | 10 | 2.83 | 2.5 | 20 | 1 | 1 | 2.83 |
| In control group | 18 | 10 | .83 | 2.5 | 10 | 0 | 1 | 2.5 |
| | For those with $X=0$ | | | | | | | |
| In treatment group | 15 | 5 | −2.17 | −2.5 | 15 | 1 | 0 | −2.17 |
| In control group | 15 | 5 | −2.17 | −2.5 | 5 | 0 | 0 | −2.5 |
| In control group | 15 | 5 | −2.17 | −2.5 | 5 | 0 | 0 | −2.5 |

where $X$ represents one or more control variables, $\beta$ is a coefficient (or a conformable vector of coefficients if $X$ represents more than one variable), and $\varepsilon^*$ is a residualized version of the original error term $\varepsilon$ from Equation (5.3) [see our earlier presentation of Equation (5.8)]. The literature on regression often states that an estimated coefficient $\hat{\delta}$ from this regression equation is unbiased and consistent for the average causal effect if $\varepsilon^*$ is uncorrelated with $D$. But, because the specific definition of $\varepsilon^*$ is conditional on the specification of $X$, many researchers find this requirement of a zero correlation difficult to interpret and hence difficult to evaluate.

The crux of the idea, however, can be understood without reference to the error term $\varepsilon^*$ but rather with reference to the simpler (and, as we have argued earlier) more clearly defined error term $v^0 + D(v^1 - v^0)$ from Equation (5.5) [or, equivalently, $Dv^1 + (1-D)v^0$ from Equation (5.6)]. Regression adjustment by $X$ in Equation (5.8) will yield an unbiased and consistent estimate of the average causal effect of $D$ when

1. $D$ is uncorrelated with $v^0 + D(v^1 - v^0)$ for each subset of respondents identified by distinct values on the variables in $X$,

2. the causal effect of $D$ does not vary with $X$, and

3. a fully flexible parameterization of $X$ is used.[12]

Consider the relationship between this set of conditions and what was described earlier in Subsection 3.2.1 as an assumption that treatment assignment is ignorable. Switching notation from $S$ to $X$ in Equation (3.3) results in

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid X, \tag{5.10}$$

where, again, the symbol $\perp\!\!\!\perp$ denotes independence. Now, rewrite the assumption, deviating $Y^0$ and $Y^1$ from their population-level expectations:

$$(v^0, v^1) \perp\!\!\!\perp D \mid X. \tag{5.11}$$

This switch from $(Y^0, Y^1)$ to $(v^0, v^1)$ does not change the assumption, at least insofar as is relevant here (because we have defined the individual-level causal effect as a linear difference, because the expectation operator is linear, and because $E[Y^0]$ and $E[Y^1]$ do not depend on who is in the treatment state and who is in the control state). Consequently, ignorability of treatment assignment can be defined only with respect to individual-level departures from the true average potential outcomes across all members of the population under the assumptions already introduced.

Given that an assumption of ignorable treatment assignment can be written as Equation (5.11), the connections between this assumption and the set of conditions that justify a regression estimator as unbiased and consistent for the effect of $D$ on $Y$ should now be clear. If treatment assignment is ignorable as defined in Equation (5.11), then a regression equation that conditions fully on all values of $X$ by including a fully flexible coding of $X$ as a set of dummy variables will yield unbiased and consistent regression estimates of the average causal effect of $D$ on $Y$. Even so, ignorability is not equivalent to the set of conditions just laid out. Instead, $v^0$ and $v^1$ [as well as functions of them, such as $v^0 + D(v^1 - v^0)$] must only be mean independent of $D$ conditional on $X$, not fully independent of $D$ conditional on $X$.

Stepping back from this correspondence, we should note that this is not the only set of conditions that would establish least squares estimation unbiased and consistent for the average causal effect, but it is the most common

[12] Here again, we use the word uncorrelated to characterize the necessary association between $D$ and $v^0 + D\,(v^1 - v^0)$. More formally, it would be best to state that $D$ and $v^0 + D(v^1 - v^0)$ must be mean independent, so that a 0 covariance of $D$ and $v^0 + D(v^1 - v^0)$ is implied.

set of conditions that would apply in most research situations.[13] Our point in laying it out is not to provide a rigid guideline applicable to all types of regression models but instead to show why the earlier statement that "$\varepsilon^*$ must be uncorrelated with $D$" is insufficiently articulated from a counterfactual perspective.

A larger point of this section, however, is that much of the received wisdom on regression modeling breaks down in the presence of individual-level heterogeneity of a causal effect, as would be present in general when causal effects are defined with reference to underlying potential outcomes tied to well-defined causal states. In the next section, we begin to explain these complications more systematically, starting from the assumption, as in prior chapters, that causal effects are inherently heterogeneous and likely to vary systematically between those in the treatment and control groups. We then present the connections among regression, matching, and stratification, building directly on our presentation of matching as conditioning by stratification in Chapter 4.

## 5.3 The Connections Between Regression and Matching

In this section, we return to the demonstrations utilized to motivate matching estimators in Chapter 4. Our goal is to establish when matching and regression yield different results, even though a researcher is attempting to adjust for the same set of variables. We then show how regression estimators can be reformulated to yield the same results as matching estimators – as a full parameterization of a perfect stratification of the data and then as weighted least squares estimators in which the weights are a function of the propensity score. In these cases, we show that regression is an effective estimator of causal effects defined by potential outcomes.

### 5.3.1 Regression as Conditional-Variance-Weighted Matching

We first show why least squares regression can yield misleading causal effect estimates in the presence of individual-level heterogeneity of causal effects, even though the only variable that needs to be adjusted for is given a fully flexible coding (i.e., when the adjustment variable is parameterized with a dummy variable for each of its values, save one for the reference category).[14] When

[13] For example, the second condition can be dropped if the heterogeneity of the causal effect is modeled as a function of $X$ (i.e., the parameterization is fully saturated in both $D$ and $X$). In this case, however, regression then becomes a way of enacting a stratification of the data, as for the matching techniques presented in the last chapter.

[14] When we write of a fully flexible coding of a variable, we are referring to a dummy variable coding of that variable only. As we will discuss later, a saturated model entails a fully flexible coding of each variable *as well as all interactions between them.* For the models discussed

a single parameter is calculated for the causal effect of $D$ on $Y$, least squares estimators implicitly invoke conditional-variance weighting of individual-level causal effects. This weighting scheme generates a conditional-variance-weighted estimate of the average causal effect, which is not an average causal effect that is often of any inherent interest to a researcher.[15] Angrist (1998) provides a more formal explanation of the following results, which is then placed in the context of a larger class of models in Angrist and Krueger (1999).

### Regression Demonstration 2

Reconsider Regression Demonstration 1, beginning on page 124. But now step back from the cautious mindset of the fictitious descriptively oriented researcher. Suppose that a causality-oriented researcher had performed the same exercise and obtained, in particular, the results for the regression model reported in Equation (5.2):

$$\hat{Y} = 1.86 + 2.75(D) + 3.76(S2) + 8.92(S3). \tag{5.12}$$

We know from Matching Demonstration 1 (beginning on page 92), on which Regression Demonstration 1 is based, that for this hypothetical example the average treatment effect among the treated is 3, the average treatment effect among the untreated is 2.4, and the unconditional average treatment effect is 2.64. If the causality-oriented researcher were to declare that the coefficient on $D$ of 2.75 in Equation (5.12) is a good estimate of the causal effect of $D$ on $Y$, then the researcher would be incautious but not appreciably incorrect. The value of 2.75 is indeed close to the true average treatment effect of 2.64, and we know from the setup of Regression Demonstration 1 that the variable $S$ continues to serve as a perfect stratifying variable. Thus, if the researcher were to state that the regression model in Equation (5.2) statistically controls for the common effect of $S$ on both $D$ and $Y$, as in Equation (5.8), where $S$ is specified as the sole element of $X$ but as two dummy variables $S2$ and $S3$, then the researcher is not horribly off the mark. The researcher has offered an adjustment for $S$, and gotten close to the true average treatment effect.

Unfortunately, the closeness of the estimate to the true average treatment effect is not a general feature of this type of a regression estimator. Under this particular specification of the regression equation, the OLS estimator yields

here, a saturated model would include interactions between the causal variable $D$ and each dummy variable for all but one of the values of $S$. For a model with only a fully flexible coding of $S$, these interactions are left out.

[15] It could be of interest to a researcher who seeks a minimum-variance estimate and who has reason to believe that the bias of the regression estimate is modest. We discuss this point later, but we hope to show that most applied researchers have good reason to want unbiased and consistent estimates rather than minimum mean-squared-error estimates of their causal parameters of interest.

precisely the value of 2.75 in an infinite sample as the sum of sample analogs to three terms:

$$\begin{aligned}&\frac{\text{Var}[D|S=1]\,\Pr[S=1]}{\sum_S \text{Var}[D|S=s]\,\Pr[S=s]}\{E[Y|D=1,S=1]-E[Y|D=0,S=1]\} \qquad (5.13)\\ &+\frac{\text{Var}[D|S=2]\,\Pr[S=2]}{\sum_S \text{Var}[D|S=s]\,\Pr[S=s]}\{E[Y|D=1,S=2]-E[Y|D=0,S=2]\}\\ &+\frac{\text{Var}[D|S=3]\,\Pr[S=3]}{\sum_S \text{Var}[D|S=s]\,\Pr[S=s]}\{E[Y|D=1,S=3]-E[Y|D=0,S=3]\}.\end{aligned}$$

These three terms are not as complicated as they may appear. First, note that the differences in the braces on the right-hand side of each term are simply the stratum-specific differences in the outcomes, which in this case are

$$\begin{aligned}E[Y|D=1,S=1]-E[Y|D=0,S=1] &= 4-2, \qquad (5.14)\\ E[Y|D=1,S=2]-E[Y|D=0,S=2] &= 8-6, \qquad (5.15)\\ E[Y|D=1,S=3]-E[Y|D=0,S=3] &= 14-10. \qquad (5.16)\end{aligned}$$

The left-hand portion of each term is then just a weight, exactly analogous to the stratum-specific weights that were used for Matching Demonstration 1 to average the stratum-specific causal effect estimates in various ways to obtain unbiased and consistent estimates of the average treatment effect, the average treatment effect for the treated, and the average treatment effect for the untreated. But, rather than use the marginal distribution of $S$, $\Pr[S]$, or the two conditional distributions of $S$, $\Pr[S|D=1]$ and $\Pr[S|D=0]$, a different set of weights is implicitly invoked by the least squares operation. In this case, the weights are composed of three pieces: (1) the variance of the treatment variable within each stratum, $\text{Var}[D|S=s]$, (2) the marginal probability of $S$ for each stratum, $\Pr[S=s]$, and (3) a summation of the product of these two terms across $S$ so that the three weights sum to 1.

Accordingly, the only new piece of this estimator that was not introduced and examined for Matching Demonstration 1 is the conditional variance of the treatment, $\text{Var}[D|S=s]$. Recall that the treatment variable is distributed within each stratum solely as a function of the stratum-specific propensity score, $\Pr[D|S=s]$. Thus, the treatment variable is a Bernoulli distributed random variable within each stratum. As can be found in any handbook of statistics, the variance of a Bernoulli distributed random variable is $p(1-p)$, where $p$ is the Bernoulli probability of success (in this case $D$ equal to 1) instead of failure (in this case $D$ equal to 0). Accordingly, the expected variance of the within-stratum treatment variable $D$ is simply $(\Pr[D|S=s])\,(1-\Pr[D|S=s])$.

For this example, the conditional variances $\text{Var}[D|S = s]$ contribute to the numerator of each weight as follows:

$$\text{Var}[D|S=1]\Pr[S=1] = \left[\left(\frac{.08}{.08+.36}\right)\left(1-\frac{.08}{.08+.36}\right)\right](.08+.36), \quad (5.17)$$

$$\text{Var}[D|S=2]\Pr[S=2] = \left[\left(\frac{.12}{.12+.12}\right)\left(1-\frac{.12}{.12+.12}\right)\right](.12+.12), \quad (5.18)$$

$$\text{Var}[D|S=3]\Pr[S=3] = \left[\left(\frac{.2}{.2+.12}\right)\left(1-\frac{.2}{.2+.12}\right)\right](.2+.12). \quad (5.19)$$

The terms in brackets on the right-hand sides of Equations (5.17)–(5.19) are $\text{Var}[D|S = 1]$, $\text{Var}[D|S = 2]$, and $\text{Var}[D|S = 3]$. The terms in parentheses on the right-hand sides of Equations (5.17)–(5.19) are the marginal probability of $S$ for each stratum, $\Pr[S = 1]$, $\Pr[S = 2]$, and $\Pr[S = 3]$. For example, for the stratum with $S = 1$, $\text{Var}[D|S = 1] = \left(\frac{.08}{.08+.36}\right)\left(1 - \frac{.08}{.08+.36}\right)$ and $\Pr[S = 1] = (.08 + .36)$. Finally, the denominator of each of the three stratum-specific weights in Equation (5.13) for this example is the sum of Equations (5.17)–(5.19). The denominator is constant across all three weights and simply scales the weights so that they sum to 1.

With an understanding of the implicit stratum-specific weights of least squares regression, the regression estimator can be seen clearly as an estimator for the average treatment effect but with supplemental conditional-variance weighting. Weighting is performed with respect to the marginal distribution of individuals across strata, but weighting is also performed with respect to the conditional variance of the treatment variable across strata as well. Thus, net of the weight given to stratum-specific effects solely as a function of $\Pr[S]$, the conditional-variance terms give more weight to stratum-specific causal effects in strata with propensity scores close to .5 and less weight to stratum-specific causal effects in strata with propensity scores close to either 0 or 1.

Why would the OLS estimator implicitly invoke conditional-variance weighting as a supplement to weighting simply by the marginal distribution of $S$? OLS is a minimum-variance-based estimator of the parameter of interest. As a result, it gives more weight to stratum-specific effects with the lowest expected variance, and the expected variance of each stratum-specific effect is an inverse function of the stratum-specific variance of the treatment variable $D$. Thus, if the two pieces of the weighting scheme are not aligned (i.e., the propensity score is close to 0 or 1 for strata that have high total probability mass but close to .5 for strata with low probability mass), then a regression estimator of this form, even under a fully flexible coding of $S$, can yield estimates that are far from the true average treatment effect even in an infinite sample.

To see the effects that supplemental weighting by the conditional variance of the treatment can have on a regression estimate, consider the alternative joint distributions for $S$ and $D$ presented in Table 5.6. For this example, suppose that the values of $E[Y^0|S, D]$, $E[Y^1|S, D]$, and $E[Y|S, D]$ in the final three panels of Table 5.1 again obtain, such that $S$ continues to offer a perfect stratification

Table 5.6: The Joint Probability Distribution for Two Variants of the Stratifying and Treatment Variables in Prior Regression Demonstration 1

| | Joint probability distribution of $S$ and $D$ | |
|---|---|---|
| | Control group: $D = 0$ | Treatment group: $D = 1$ |
| | Variant I | |
| $S = 1$ | $\Pr[S = 1, D = 0] = .40$ | $\Pr[S = 1, D = 1] = .04$ |
| $S = 2$ | $\Pr[S = 2, D = 0] = .20$ | $\Pr[S = 2, D = 1] = .04$ |
| $S = 3$ | $\Pr[S = 3, D = 0] = .16$ | $\Pr[S = 3, D = 1] = .16$ |
| | Variant II | |
| $S = 1$ | $\Pr[S = 1, D = 0] = .40$ | $\Pr[S = 1, D = 1] = .04$ |
| $S = 2$ | $\Pr[S = 2, D = 0] = .12$ | $\Pr[S = 2, D = 1] = .12$ |
| $S = 3$ | $\Pr[S = 3, D = 0] = .03$ | $\Pr[S = 3, D = 1] = .29$ |

of the data. Note that, for the two alternative joint distributions of $S$ and $D$ in Table 5.6, the marginal distribution of $S$ remains the same as in Regression Example 1: $\Pr[S = 1] = .44$, $\Pr[S = 2] = .24$, and $\Pr[S = 3] = .32$. As a result, the unconditional average treatment effect is the same for both variants of the joint distribution of $S$ and $D$ depicted in Table 5.6, and it matches the unconditional average treatment effect for the original example represented fully in Table 5.1. In particular, the same distribution of stratum-specific causal effects results in an unconditional average treatment effect of 2.64.

The difference represented by each variant of the joint distributions in Table 5.6 is in the propensity score for each stratum of $S$, which generates an alternative marginal distribution for $D$ and thus alternative true average treatment effects for the treated and for the untreated (and, as we will soon see, alternative regression estimates from the same specification).

For Variant I in Table 5.6, those with $S$ equal to 1 or 2 are much less likely to be in the treatment group, and those with $S$ equal to 3 are now only equally likely to be in the treatment group and the control group. As a result, the marginal distribution of $D$ is now different, with $\Pr[D = 0] = .76$ and $\Pr[D = 1] = .24$. The average treatment effect for the treated is now 3.33 whereas the average treatment effect among the untreated is 2.42. Both of these are larger than was the case for the example represented by Table 5.1 because (1) a greater proportion of those in the control group have $S = 3$ (i.e., $\frac{.16}{.76} > \frac{.12}{.6}$), (2) a greater proportion of those in the treatment group have $S = 3$ (i.e., $\frac{.16}{.24} > \frac{.2}{.4}$), and (3) those with $S = 3$ gain the most from the treatment.

For Variant II, those with $S$ equal to 1 are still very unlikely to be in the treatment group, but those with $S$ equal to 2 are again equally likely to be in the treatment group. But those with $S$ equal to 3 are now very likely to be in the treatment group. As a result, the marginal distribution of $D$ is now different again, with $\Pr[D=0]=.55$ and $\Pr[D=1]=.45$, and the average treatment effect for the treated is now 3.29 whereas the average treatment effect among the untreated is 2.11. Both of these are smaller than for Variant I because a smaller proportion of both the treatment group and the control group have $S=3$.

For these two variants of the joint distribution of $S$ and $D$, we have examples in which the unconditional average treatment effect is the same as it was for the example in Table 5.1, but the underlying average treatment effects for the treated and for the untreated differ considerably. Does the reestimation of Equation (5.12) for these variants of the example still generate an estimate for the coefficient on $D$ that is (a) relatively close to the true unconditional average treatment effect and (b) closer to the unconditional average treatment effect than either the average treatment effect for the treated or the average treatment effect for the untreated?

For Variant I, the regression model yields

$$\hat{Y} = 1.90 + 3.07(D) + 3.92(S2) + 8.56(S3) \tag{5.20}$$

for an infinite sample. In this case, the coefficient of 3.07 on $D$ is not particularly close to the unconditional average treatment effect of 2.64, and in fact it is closer to the average treatment effect for the treated of 3.33 (although still not particularly close). For Variant II, the regression model yields

$$\hat{Y} = 1.96 + 2.44(D) + 3.82(S2) + 9.45(S3). \tag{5.21}$$

In this case, the coefficient of 2.44 on $D$ is closer to the unconditional average treatment effect of 2.64, but not as close as was the case for the example in Regression Demonstration 1. It is now relatively closer to the average treatment effect for the untreated, which is 2.11 (although, again, still not particularly close).

For Variant I, the regression estimator is weighted more toward the stratum with $S=3$, for which the propensity score is .5. For this stratum, the causal effect is 4. For Variant II, in contrast, the regression estimator is weighted more toward the stratum with $S=2$, for which the propensity score is .5. And, for this stratum, the causal effect is 2.[16]

What is the implication of these alternative setups of the same basic demonstration? Given that the unconditional average treatment effect is the same for all three joint distributions of $S$ and $D$, it would be unwise for the incautious researcher to believe that this sort of a regression specification will provide a reliably close estimate to the unconditional average treatment effect, the average treatment effect for the treated, or the unconditional average treatment effect when there is reason to believe that these three average causal effects differ because of individual-level heterogeneity. The regression estimate will be weighted

[16] Recall that, because the marginal distribution of $S$ is the same for all three joint distributions of $S$ and $D$ by construction of the example, the $\Pr[S=s]$ pieces of the weights remain the same for all three alternatives. Thus, the differences between the regression estimates are produced entirely by differences in the $\text{Var}[D|S=s]$ pieces of the weights.

toward stratum-specific effects for which the propensity score is closest to .5, net of all else.

In general, regression models do not offer consistent estimates of the average treatment effect when causal effect heterogeneity is present, even when a fully flexible coding is given to the only necessary adjustment variable(s). Regression estimators with fully flexible codings of the adjustment variables do provide consistent estimates of the average treatment effect if either (a) the true propensity score does not differ by strata or (b) the average stratum-specific causal effect does not vary by strata.[17] Condition (a) would almost never be true (because, if it were, one would not even think to adjust for $S$ because it is already independent of $D$). And condition (b) is probably not true in most applications, because rarely are investigators willing to assert that all consequential heterogeneity of a causal effect has been explicitly modeled.

Instead, for this type of a regression specification, in which all elements of a set of perfect stratifying variables $S$ are given fully flexible codings (i.e., a dummy variable coding for all but one of the possible combinations of the values for the variables in $S$), the OLS estimator $\hat{\delta}_{\text{OLS, multiple}}$ in Equation (5.9) is equal to

$$\frac{1}{c}\sum_{s} \mathrm{Var}_N[d_i|s_i = s]\ \Pr{}_N[s_i = s]\left\{E_N[y_i|d_i = 1, s_i = s] - E_N[y_i|d_i = 0, s_i = s]\right\} \tag{5.22}$$

in a sample of size $N$. Here, $c$ is a scaling constant equal to the sum (over all combinations of values $s$ of $S$) of the terms $\mathrm{Var}_N[d_i|s_i = s]\Pr_N[s_i = s]$.

There are two additional points to emphasize. First, the weighting scheme for stratified estimates in Equation (5.22) applies only when the fully flexible parameterization of $S$ is specified. Under a constrained specification of $S$ [e.g., in which some elements of $S$ are constrained to have linear effects, as in Equation (5.1)] the weighting scheme is more complex. The weights remain a function of the marginal distribution of $S$ and the stratum-specific conditional variance of $D$, but the specific form of each of these components becomes conditional on the specification of the regression model (see Section 2.3.1 of Angrist and Krueger 1999). The basic intuition here is that a linear constraint on a variable in $S$ in a regression model represents an implicit linearity assumption about true underlying propensity score that may not be linear in $S$.[18]

[17] As a by-product of either condition, the average treatment effect must be equal to the average treatment effect for the treated and the average treatment effect for the untreated. Thus, the regression estimator would be consistent for both of these as well.

[18] For a binary causal exposure variable $D$, a many-valued variable $S$ that is treated as an interval-scaled variable, and a regression equation

$$\hat{Y} = \hat{\alpha} + \hat{\delta}(D) + \hat{\beta}(S)\ ,$$

the OLS estimator $\hat{\delta}$ is equal to

$$\frac{1}{l}\sum_{s} \widetilde{\mathrm{Var}}_N[\hat{d}_i|s_i = s]\ \widetilde{\Pr}_N[s_i = s]\left\{E_N[y_i|d_i = 1, s_i = s] - E_N[y_i|d_i = 0, s_i = s]\right\}$$

Second, regression can make it all too easy to overlook the same sort of fundamental mismatch problems that were examined for Matching Demonstration 2 in Subsection 4.2.2. Regression will implicitly drop strata for which the propensity score is either 0 or 1 in the course of forming its weighted average by Equation (5.22). As a result, a researcher who interprets a regression result as a decent estimate of the average treatment effect, but with supplemental conditional-variance weighting, may be entirely wrong. No meaningful average causal effect may exist in the population. The second point is best explained by the following illustration.

## Regression Demonstration 3

Reconsider the hypothetical example presented as Matching Demonstration 2 from Chapter 4, beginning on page 95. The assumed relationships that generate the data are very similar to those for Regression Demonstrations 1 and 2, but, as shown in Table 5.7, no individual for whom $S$ is equal to 1 in the population is ever exposed to the treatment because $\Pr[S = 1, D = 1] = 0$ whereas $\Pr[S = 1, D = 0] = .4$. As a result, not even an infinite sample from the population would ever include an individual in the treatment group with $s_i = 1$.[19] Because

in a sample of size $N$, where $l$ is a scaling constant equal to the sum over all $s$ of $S$ the terms $\widetilde{\text{Var}}_N[\hat{d}_i|s_i = s]\ \widetilde{\Pr}_N[s_i = s]$.

The distinction between $\widetilde{\text{Var}}_N[\hat{d}_i|s_i = s]\ \widetilde{\Pr}_N[s_i = s]$ and $\text{Var}_N[d_i|s_i = s]\ \Pr_N[s_i = s]$ in the main text results from a constraint on the propensity score that is implicit in the regression equation. In specifying $S$ as an interval-scaled variable, least squares implicitly assumes that the true propensity score $\Pr[D|S]$ is linear in $S$. As a result, the first portion of the stratum-specific weight is

$$\widetilde{\text{Var}}_N[\hat{d}_i|s_i = s] \equiv \hat{p}_s(1 - \hat{p}_s),$$

where $\hat{p}_s$ is equal to the predicted stratum-specific propensity score from a linear regression of $d_i$ on $s_i$: $\hat{p}_s = \hat{\xi} + \hat{\phi}_s s$.

Perhaps somewhat less clear, the term $\widetilde{\Pr}_N[s_i = s]$ is also a function of the constraint on $S$. $\widetilde{\Pr}_N[s_i = s]$ is not simply the marginal distribution of $S$ in the sample, as $\Pr_N[s_i = s]$ is. Rather, one must use Bayes' rule to determine the implied marginal distribution of $S$, given the assumed linearity of the propensity score across levels of $S$. Rearranging

$$\Pr[d_i = 1|s_i] = \frac{\Pr[s_i|d_i = 1]\Pr[d_i = 1]}{\Pr[s_i]}$$

as

$$\Pr[s_i] = \frac{\Pr[s_i|d_i = 1]\Pr[d_i = 1]}{\Pr[d_i = 1|s_i]},$$

and then substituting $\hat{p}_s$ for $\Pr[d_i = 1|s_i]$, we then find that

$$\widetilde{\Pr}_N[s_i = s] = \frac{\Pr_N[s_i = s|d_i = 1]\Pr_N[d_i = 1]}{\hat{p}_s}.$$

The terms $\Pr_N[s_i = s|d_i = 1]$ and $\Pr_N[d_i = 1]$ are, however, unaffected by the linearity constraint on the propensity score. They are simply the true conditional probability of $S$ equal to $s$ given $D$ equal to $d$ as well as the marginal probability of $D$ equal to $d$ for a sample of size $N$.

Note that, if the true propensity score is linear in $S$, then the weighting scheme here is equivalent to the one in the main text.

[19] Again, recall that we assume no measurement error in general in this book. In the presence of measurement error, some individuals might be misclassified and therefore might show up in the data with $s_i = 1$ and $d_i = 1$.

Table 5.7: The Joint Probability Distribution and Conditional Population Expectations for Regression Demonstration 3

| | Joint probability distribution of $S$ and $D$ | |
|---|---|---|
| | Control group: $D = 0$ | Treatment group: $D = 1$ |
| $S = 1$ | $\Pr[S = 1, D = 0] = .4$ | $\Pr[S = 1, D = 1] = 0$ |
| $S = 2$ | $\Pr[S = 2, D = 0] = .1$ | $\Pr[S = 2, D = 1] = .13$ |
| $S = 3$ | $\Pr[S = 3, D = 0] = .1$ | $\Pr[S = 3, D = 1] = .27$ |
| | Potential outcomes under the control state | |
| $S = 1$ | $E[Y^0 \vert S = 1, D = 0] = 2$ | |
| $S = 2$ | $E[Y^0 \vert S = 2, D = 0] = 6$ | $E[Y^0 \vert S = 2, D = 1] = 6$ |
| $S = 3$ | $E[Y^0 \vert S = 3, D = 0] = 10$ | $E[Y^0 \vert S = 3, D = 1] = 10$ |
| | Potential outcomes under the treatment state | |
| $S = 1$ | $E[Y^1 \vert S = 1, D = 0] =4$ | |
| $S = 2$ | $E[Y^1 \vert S = 2, D = 0] = 8$ | $E[Y^1 \vert S = 2, D = 1] = 8$ |
| $S = 3$ | $E[Y^1 \vert S = 3, D = 0] = 14$ | $E[Y^1 \vert S = 3, D = 1] = 14$ |
| | Observed outcomes | |
| $S = 1$ | $E[Y \vert S = 1, D = 0] = 2$ | |
| $S = 2$ | $E[Y \vert S = 2, D = 0] = 6$ | $E[Y \vert S = 2, D = 1] = 8$ |
| $S = 3$ | $E[Y \vert S = 3, D = 0] = 10$ | $E[Y \vert S = 3, D = 1] = 14$ |

of this structural zero in the joint distribution of $S$ and $D$ , the three conditional expectations, $E[Y^0|S = 1, D = 0]$, $E[Y^1|S = 1, D = 0]$, and $E[Y|S = 1, D = 0]$, are properly regarded as undefined and hence are omitted from the last three panels of Table 5.7.

As shown earlier in Subsection 4.2.2, the naive estimator can still be calculated for this example and will be equal to 8.05 in an infinite sample. Moreover, the average treatment effect for the treated can be estimated consistently as 3.35 by considering only the values for those with $S$ equal to 2 and 3. But there is no way to consistently estimate the treatment effect for the untreated, and hence no way to consistently estimate the unconditional average treatment effect.

Consider now the estimated values that would be obtained with data arising from this joint distribution for a regression model specified equivalently as in Equations (5.12), (5.20), and (5.21):

$$\hat{Y} = 2.00 + 3.13(D) + 3.36(S2) + 8.64(S3). \tag{5.23}$$

In this case, the OLS estimator is still equivalent to Equation (5.22), which in an infinite sample would then be equal to Equation (5.13). But, with reference to Equation (5.13), note that the weight for the first term,

$$\frac{\text{Var}[D|S=1]\,\text{Pr}[S=1]}{\sum\limits_{S}\text{Var}[D|S=s]\,\text{Pr}[S=s]},$$

is equal to zero because $\text{Var}[D|S=1]$ is equal to 0 in the population by construction. Accordingly, the numerator of the stratum-specific weight is zero, and it enters into the summation of the denominator of the other two stratum-specific weights as zero. As a result, the regression estimator yields a coefficient on $D$ that is 3.13, which is biased downward as an estimate of the average treatment effect for the treated and has no relationship with the undefined average treatment effect. If interpreted as an estimate of the average treatment effect for the treated, but with supplemental conditional-variance weighting, then the coefficient of 3.13 is interpretable. But it cannot be interpreted as a meaningful estimate of the average treatment effect in the population once one commits to the potential outcome framework because the average treatment effect does not exist.

The importance of this demonstration is only partly revealed in this way of presenting the results. Imagine that a researcher simply observes $\{y_i, d_i, s_i\}_{i=1}^{N}$ and then estimates the model in Equation (5.23) without first considering the joint distribution of $S$ and $D$ as presented in Table 5.7. It would be entirely unclear to such a researcher that there are no individuals in the sample (or in the population) whose values for both $D$ and $S$ are 1. Such a researcher might therefore be led to believe that the coefficient estimate for $D$ is a meaningful estimate of the causal effect of $D$ for all members of the population.

All too often, regression modeling, at least as practiced in the social sciences, makes it too easy for an analyst to overlook fundamental mismatches between treatment and control cases. And, thus, one can obtain average treatment effect estimates with regression techniques even when no meaningful average treatment effect exists. Even though this is the case, we do not want to push this argument too far. Therefore, in the next section we make the (perhaps obvious) point that regression can be used as a technique to execute a perfect stratification of the data under the same assumptions that justified matching as a stratification estimator in Chapter 4.

### 5.3.2 Regression as an Implementation of a Perfect Stratification

Matching and regression can both be used to carry out a perfect stratification of the data. Consider how the matching estimates presented in Matching Demonstration 1 (see page 92) could have been generated by standard regression routines. For that hypothetical example, an analyst could specify $S$ as two dummy variables and $D$ as one dummy variable. If all two-way interactions between $S$ and $D$ are then included in a regression model predicting the observed outcome $Y$, then the analyst has enacted the same perfect stratification of the data by fitting a model that is saturated in both $S$ and $D$ to all of the cells of the first panel of Table 4.2 (or see instead the reproduction in Table 5.1 on page 125):

$$\hat{Y} = 2 + 2(D) + 4(S2) + 8(S3) + 0(D \times S2) \; + 2(D \times S3). \tag{5.24}$$

The values of each of the six cells of the panel are unique functions of the six estimated coefficients from the regression model. Accordingly, by use of the marginal distribution of $S$ and the joint distribution of $S$ given $D$, coefficient contrasts can be averaged across the relevant distributions of $S$ in order to obtain consistent estimates of the average treatment effect, the treatment effect among the treated, and the treatment effect among the untreated.

Nevertheless, for many applications, such a saturated model may not be possible, and in some cases this impossibility may be misinterpreted. For Regression Demonstration 3 (see page 149), if one were to fit the seemingly saturated model with the same six parameters as in Equation (5.24), the coefficient on $D$ would be dropped by standard software routines. One might then attribute this to the size of the dataset and then instead use a more constrained parameterization [i.e., either enter $S$ as a simple linear term interacted with $D$ or instead specify the model in Equation (5.23)]. These models must then be properly interpreted, and in no case could they be interpreted as yielding unbiased and consistent estimates of the average treatment effect. In a sense, this problem is simply a matter of model misspecification. But, at a deeper level, it may be that regression as a method tends to encourage the analyst to oversimplify these important model specification issues.[20]

What if a zero cell in the joint distribution of $S$ and $D$ occurred by chance in any single dataset? In other words, what if there is no fundamental overlap problem in the distribution of $S$ across $D$, but instead the only problem is a finite dataset? In this case, regression can be reformulated as a weighting estimator, as we describe in the next subsection, in order to solve the sparseness problem.

### 5.3.3 Matching as Weighted Regression

In this subsection, we explore further the connections between matching and regression estimators, demonstrating how weighted regression can be used to estimate causal effects. Imbens (2004) reviews alternative ways to calculate the same sorts of weighted averages that we present here, and he fully accounts for the connections between inverse-probability-weighting procedures and nonparametric regression.

As was shown for the hypothetical example in Matching Demonstration 3 (see page 100), matching can be considered a method to weight the data in order

[20] Rubin (1977) provides simple and elegant examples of all such complications, highlighting the importance of assumptions about the relationships between covariates and outcomes (see also Holland and Rubin 1983 and Rosenbaum 1984a, 1984b).

to balance predictors of treatment selection and thereby calculate contrasts that can be given causal interpretations. In this section, we show that the three propensity-score-weighting estimators in Equations (4.12) – (4.14) can be specified as three weighted OLS regression estimators. In fact, if one defines a weighting variable appropriately, then any standard software package that estimates weighted regression can be used.

To see how to do this, note first that the naive estimator in Equation (2.7) can be written as an OLS estimator, $(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}$, where (1) $\mathbf{Q}$ is an $n \times 2$ matrix that contains a vector of 1s in its first column and a vector of the values of $d_i$ for each individual in its second column and (2) $\mathbf{y}$ is an $n \times 1$ column vector containing values of $y_i$ for each individual. To estimate each of the propensity-score-weighting estimators in Equations (4.12)–(4.14), simply estimate a weighted OLS estimator:

$$\hat{\delta}_{\text{OLS, weighted}} \equiv (\mathbf{Q}'\mathbf{P}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{P}\mathbf{y}, \tag{5.25}$$

where $\mathbf{P}$ is an appropriately chosen weight matrix, depending on the average treatment effect of interest.

For the treatment effect for the treated, specify $\mathbf{P}$ as an $n \times n$ diagonal matrix with 1 in the $i \times i$th place for members of the treatment group and $\hat{p}_i/(1-\hat{p}_i)$ in the $i \times i$th place for members of the control group (where, as defined earlier for the hypothetical example in Matching Demonstration 3, $\hat{p}_i$ is the estimated propensity score; see Subsection 4.3.2). For the treatment effect for the untreated, specify $\mathbf{P}$ as an $n \times n$ diagonal matrix with $(1-\hat{p}_i)/\hat{p}_i$ in the $i \times i$th place for members of the treatment group and 1$s$ in the $i \times i$th place for members of the control group. Finally, for the unconditional average treatment effect, specify $\mathbf{P}$ as an $n \times n$ diagonal matrix with $1/\hat{p}_i$ in the $i \times i$th place for members of the treatment group and $1/(1-\hat{p}_i)$ in the $i \times i$th place for members of the control group. Consider the following demonstration, which builds directly on the prior presentation of matching in Section 4.3.

### Regression Demonstration 4

Consider first how the matching estimates in the hypothetical example in Matching Demonstration 3 (beginning on page 100) could have been generated by a standard regression routine. As shown there for Equations (4.8) and (4.9), the potential outcomes were specified as functions of individual values for $A$ and $B$:

$$y_i^1 = 102 + 6a_i + 4b_i + v_i^1, \tag{5.26}$$

$$y_i^0 = 100 + 3a_i + 2b_i + v_i^0, \tag{5.27}$$

where $A$ and $B$ are distributed as independent uniform random variables with a minimum of .1 and a maximum of 1, and where $v_i^1$ and $v_i^0$ are independent random draws from a normal distribution with expectation 0 and a standard deviation of 5. For each individual, $y_i$ is then equal to $y_i^1(d_i) + (1-d_i)y_i^0$, where the value of $d_i$ is determined by a Bernoulli distribution with the probability of 1 rather than 0 as the nonlinear function in $A$ and $B$ that is presented in Figure 4.1 in Subsection 4.3.2. The first panel of Table 5.8 reproduces the true average treatment effects for this example from the prior Table 4.5; the unconditional average treatment effect is 4.53 whereas the average treatment effects for the treated and the untreated are 4.89 and 4.40, respectively.

The second panel of Table 5.8 introduces a second variant on this basic setup. For this variant, Equations (5.26) and (5.27) are replaced with

$$\begin{aligned} y_i^1 &= 102 + 3a_i + 2b_i + 6(a_i \times b_i) + v_i^1, && (5.28) \\ y_i^0 &= 100 + 2a_i + 1b_i - 2(a_i \times b_i) + v_i^0, && (5.29) \end{aligned}$$

but everything else remains the same. These alternative potential outcome definitions result in a slightly more dramatic pattern for the average treatment effects. As shown in the second panel of Table 5.8, the unconditional average treatment effect is 5.05 whereas the average treatment effects for the treated and the untreated are now 5.77 and 4.79, respectively. Although this difference is notable, the nonlinearity of the individual-level treatment effects is of most consequence here. The opposite-signed parameters specified for the cross-product interaction of $A$ and $B$ ensures that those with high levels of $A$ and $B$ together have much larger individual-level treatment effects than others. For Variant I in the first panel of Table 5.8, the differential sizes of the treatment effects were separable into simple linear pieces that could be independently attributed to $A$ and $B$.

For each variant of this example within its corresponding panel, three coefficients on $D$ (for an infinite sample so that sampling error is zero) are presented for three OLS regression specifications: (1) $Y$ regressed on $D$, (2) $Y$ regressed on $D$, $A$, and $B$, and (3) $Y$ regressed on $D$, $A$, $A$-squared, $B$, and $B$-squared. All three of these OLS estimates are placed in the column for the average treatment effect because such estimates are commonly interpreted as average treatment effect estimates. But, of course, this is somewhat misleading, because regression estimates such as these are often presented without any reference to the average treatment effect (i.e., often as an estimate of an implicit constant structural effect of $D$ on $Y$). None of these OLS estimates is particularly close to its respective true average treatment effect. And, although the quadratic specifications of $A$ and $B$ help to some degree, the estimates are still far from their targets.

For the last row of each panel, we then implemented the weighted regression models specified earlier in Equation (5.25). For these estimates, we take the estimated propensity scores (i.e., the $\hat{p}_i$ used to construct each of the three $\mathbf{P}$ matrices) from the row labeled "Perfectly specified propensity score estimates" in Table 4.5. Each of the estimates lands exactly on the targeted parameter, as will always be the case in a sufficiently large dataset if the propensity score is estimated flawlessly.

As this demonstration shows, the relationships between matching and regression are now well established in the literature. In this section, we have offered a demonstration to show some of these connections. But, more generally,

Table 5.8: OLS and Weighted OLS Estimates of Treatment Effects for Regression Demonstration 4

| | Average treatment effects | | |
|---|---|---|---|
| | $E[\delta]$ | $E[\delta \mid D=1]$ | $E[\delta \mid D=0]$ |
| Variant I: $Y^1$and $Y^0$ linear in $A$ and $B$ | | | |
| True treatment effects | 4.53 | 4.89 | 4.40 |
| OLS regression estimates: | | | |
| $Y$regressed on $D$ | 5.39 | | |
| $Y$regressed on $D$ and linear $A$ and $B$ | 4.75 | | |
| $Y$regressed on $D$ and quadratic $A$ and $B$ | 4.74 | | |
| Weighted OLS regression of $Y$on $D$ | 4.53 | 4.89 | 4.40 |
| Variant II: $Y^1$and $Y^0$nonlinear in $A$ and $B$ | | | |
| True treatment effects | 5.05 | 5.77 | 4.79 |
| OLS regression estimates: | | | |
| $Y$regressed on $D$ | 5.88 | | |
| $Y$regressed on $D$ and linear $A$ and $B$ | 5.47 | | |
| $Y$regressed on $D$ and quadratic $A$ and $B$ | 5.44 | | |
| Weighted OLS regression of $Y$on $D$ | 5.05 | 5.77 | 4.79 |

it is now known that most matching estimators can be rewritten as forms of nonparametric regression (see Abadie and Imbens 2006; Hahn 1998; Heckman, Ichimura, and Todd 1998; Hirano et al. 2003; Imbens 2004; Lunceford and Davidian 2004). And, as shown earlier in this chapter, OLS regression, under certain specifications, can be seen as a form of matching with supplemental conditional-variance weighting (which may or may not be useful, depending on the application). We have further shown with this demonstration one particular advantage of a matching estimator, whether carried out as specified earlier in Chapter 4, or as a weighted regression estimator as shown in this subsection. Matching may significantly outperform regression models when the true functional form of a regression is nonlinear but a simple linear specification is used mistakenly. However, as shown earlier, the superior performance requires that the propensity score be estimated effectively, and perhaps flawlessly, in the case of weighting estimators.

### 5.3.4 Regression as Supplemental Adjustment When Matching

For Equation (5.25), we defined the **Q** matrix as containing a column of 1*s* and a column of individual-level values $d_i$. In fact, there is a literature that argues that all variables that predict treatment selection should be included as additional columns in **Q** as well. The idea is to offer what James Robins refers to as a "doubly robust" or "doubly protected" estimator (see Bang and Robins 2005; Robins and Rotnitzky 2001). Robins and Rotnitzky reflect on the fallibility of both standard regression methods and propensity-score-based weighting estimators:

> There has been considerable debate as to which approach to confounder control is to be preferred, as the first is biased if the outcome regression model is misspecified while the second approach is biased if the treatment regression, i.e., propensity, model is misspecified. This controversy could be resolved if an estimator were available that was guaranteed to be consistent for $\theta$ whenever at least one of the two models was correct under an asymptotic sequence in which the outcome and treatment regression models remain fixed as the sample size $n$ increases to infinity. We refer to such combined methods as doubly-robust or doubly-protected as they can protect against misspecification of either the outcome or treatment model, although not against simultaneous misspecification of both. (Robins and Rotnitzky 2001:922)

The basic motivation of this practice is to give the analyst two chances to "get it right," in hopes that misspecifications of the propensity-score-estimating equation and the final regression equation will neutralize each other. And, although Robins is credited with developing the recent asymptotic justification for a variety of specific procedures (see Robins and Ritov 1997; Robins, Rotnitzky, and Zhao 1994; Scharfstein, Rotnitzky, and Robins 1999; van der Laan and Robins 2003), the idea of using matching and regression together is quite general and has a long legacy in applied work (see Cochran and Rubin 1973; Gelman and King 1990; Heckman, Ichimura, and Todd 1998; Hirano and Imbens 2001; Rubin and Thomas 1996, 2000; Smith and Todd 2005). Consider the following demonstration that shows some of these possibilities.

#### Regression Demonstration 5

Recall Matching Demonstration 4, beginning on page 110, and consider now how regression can be used to supplement a matching algorithm.[21] Recall that for Matching Demonstration 4, we presented matching estimates of the Catholic school effect on achievement for simulated data. As we discuss there, the treatment effect for the treated is 6.96 in the simulated data, whereas the treatment

[21] The results for this demonstration are drawn from Morgan and Harding (2006), and we thereby thank David Harding for his contribution to this section. More details on these results are presented in the article as well.

Table 5.9: Combined Matching and Regresison Estimates for the Simulated Effect of Catholic Schooling on Achievement, as Specified in Matching Demonstration 4

| | Poorly specified propensity-score-estimating equation | | Well-specified propensity-score-estimating equation | |
|---|---|---|---|---|
| | TT Estimate | Bias | TT Estimate | Bias |
| OLS Regression: | | | | |
| Not restricted to region of common support | 7.79 | 0.83 | 6.81 | -0.15 |
| Restricted to region of common support | 7.88 | 0.92 | 6.80 | -0.16 |
| Matching with regression adjustment: | | | | |
| Interval with variable blocks (B&I) | 7.95 | 0.99 | 6.70 | -0.26 |
| One Nearest Neighbor with caliper = 0.001 (L&S) | 8.05 | 1.09 | 7.15 | 0.19 |
| One Nearest Neighbor without caliper (Abadie) | 7.78 | 0.82 | 6.88 | -0.08 |
| Five Nearest Neighbors with caliper = 0.001 (L&S) | 7.92 | 0.96 | 7.17 | 0.21 |
| Five Nearest Neighbors without caliper (Abadie) | 7.82 | 0.86 | 7.20 | 0.24 |

*Notes:* B&I denotes the software of Becker and Ichino; L&S denotes the software of Leuven and Sianesi; Abadie denotes the software of Abadie et al.

effect for the untreated is 5.9. In combination, the average treatment effect is then 6.0.

In this demonstration, we present in the first two lines of Table 5.9 least squares regression estimates of the treatment effect under two specifications, including the same variables for the propensity-score-estimating equation directly in the regression equation (and in the two different specifications used for the earlier Table 4.6). We present regression estimates in two variants: (1) without regard to the distributions of the variables and (2) based on a subsample restricted to the region of common support, as defined by the propensity score estimated from the covariates utilized for the respective scenario. Comparing these estimates with the values from Table 4.6, linear regression does about as well as the matching algorithms as an estimator of the treatment effect for the treated. In some cases, these estimates outperform some of the matching estimates. In fairness to the matching estimates, however, it should be pointed out that the data analyzed for this example are well suited to regression because the assumed functional form of each potential outcome variable is linear and hence relatively simple. Although we believe that this is reasonable for the simulated application, there are surely scenarios in which matching can be shown to clearly outperform regression because of nonlinearities that are not parameterized by the relevant regression model.

In the second panel of Table 5.9, we provide five examples of matching combined with regression adjustment. Interval matching with regression adjustment calculates the treatment effect within blocks after adjusting for the same covariates included in the propensity-score-estimating equation for the particular scenario, averaging over blocks to produce an overall treatment effect estimate. With nearest-neighbor matching, one accomplishes regression adjustment by regressing the outcome on the treatment and covariates using the matched sample, with appropriate weights for duplicated observations in the matched control group and for multiple neighbor matching. When the results of Tables 4.6 and 5.9 are compared, supplemental regression adjustment reduces the bias for only the nearest-neighbor matching with one match, whereas it offers no improvement for the other three matching estimators.

As this demonstration shows, supplemental regression adjustment may provide a slight improvement over an analogous matching estimator implemented without regression adjustment. But this is not true in all cases, especially when the matching estimator chosen already uses many cases to match to each target case.

Another advantage of combining matching and regression has emerged recently. Rather than consider regression as a supplement to a possibly faulty matching routine, one can consider matching a remedy to artifactual regression results that have been produced by incautious data mining. Ho et al. (2005) suggest that the general procedure one should carry out in any multivariate analysis is to first balance one's data with a matching routine and then to estimate a regression model on the balanced data. From this perspective, matching is a data preprocessor, which can be used to prepare the data for subsequent analysis with something such as a regression routine.[22]

## 5.4 Extensions and Other Perspectives

In this chapter, we have focused almost exclusively on the estimation of the effect of a binary cause on an interval-scaled outcome, and we have considered only least squares adjustments. Before carrying on to discuss least squares estimation of the effects of many-valued causes, we of course must concede what the reader is surely aware of: We have considered only a tiny portion of what falls under the general topic of regression modeling. We have not considered categorical outcome variables, time series analysis, nested data structures, variance-component models, and so on. One can gain a full perspective of the variants of regression modeling in just sociology and economics by consulting Agresti (2002), Arminger, Clogg, and Sobel (1995), Berk (2004), Hamilton (1994), Hayashi (2000), Hendry (1995), Long (1997), Powers and Xie (2000), Raudenbush and Bryk (2002), Ruud (2000), and Wooldridge (2002).

In the next subsection, we consider only one modest extension: least squares regression models for many-valued causes. This presentation then leads naturally to a discussion that follows in the next subsection of what might be labelled the "all-cause correct specification" tradition of regression analysis. Informed by the demonstrations offered in this chapter, we discuss the attractiveness of the promise of this alternative perspective but also the implausibility of the

[22] As we note in the conclusion to this chapter, our perspective is that matching and regression should be used together, and to the extent that the distinction between them fades away. Brand and Halaby (2006) is an example of this approach.

perspective as a general guide for either causal analysis or regression practice in the social sciences.

### 5.4.1 Regression Estimators for Many-Valued Causes

We suspect that the vast majority of published regression estimates of causal effects in the social sciences are for causes with more than two values. Accordingly, as in Subsection 4.6.3 on matching estimators for many-valued causes, we must discuss the additional complexities of analogous regression estimators. We will again, however, restrict attention to an interval-scaled outcome.

First, again recall the basic setup for many-valued causes from Chapter 2, Appendix B, in which we have a set of $J$ treatment states, a corresponding set of $J$ causal exposure dummy variables, $\{Dj\}_{j=1}^{J}$, and a corresponding set of $J$ potential outcome random variables, $\{Y^{Dj}\}_{j=1}^{J}$. The treatment received by each individual is $Dj^*$.

How would one estimate the causal effect of such a $J$-valued cause with regression methods? The first answer should be clear from our presentation in the last section: Because regression can be seen as a form of matching, one can use the same basic strategies outlined for matching estimators of many-valued causes in Subsection 4.6.3. One could form a series of two-way comparisons between the values of the cause, estimate a separate propensity score for each contrast, and then use a weighted regression model to estimate each pairwise causal effect.

If the number of causal states is relatively large, then this general strategy is infeasible. Some smoothing across pairwise comparisons would be necessary, either by collapsing some of the $J$ causal states or by imposing an ordering on the distribution of the causal effect across the $J$ causal states. The most common parametric restriction would be to assume that the causal effect is linear in $j$ for each individual $i$. For example, for a set of causal states (such as years of schooling) enumerated by values from 0 to 1, 2, 3, to $J$, the linearity assumption is the assumption that $y_i^{Dj} = y_i^{D0} + \beta_i(j)$ for all $j > 0$, which requires that the difference $y_i^{Dj} - y_i^{Dj-1}$ for each individual $i$ be equal to a constant $\beta_i$ for all $j > 0$. In this case, the individual-level causal effect is then a slope $\beta_i$ [rather than the simple difference in potential outcomes, $\delta_i$, specified earlier in Equation (2.1)]. This setup is analogous to the dose-response models for matching estimators discussed in Subsection 4.6.3, but it explicitly leaves open the possibility that the dose-response relationship varies across individuals even though it remains linear.

Angrist and Krueger (1999) show in a very clear example how both a linearity assumption on the individual-specific, dose-response relationship and a fully flexible coding of adjustment variables results in an OLS weighting scheme for the average value of $\beta_i$ in a sample that is even more complex than what we discussed earlier for simple binary causes (see Regression Demonstration 2 in Subsection 5.3.1). A form of conditional-variance weighting is present again, but now the weighting is in multiple dimensions because least squares must calculate average derivatives across the linearly ordered causal variable (see Angrist and

Krueger 1999, Equation 34). Because one cannot intuitively grasp how these weights balance out across all the dimensions of the implicit weighting scheme (at least we cannot do so), Angrist and Krueger help by offering a familiar example: an OLS estimate of the average causal effect of an additional year of schooling on labor market earnings, assuming linearity in years of schooling and using a fully flexible coding of adjustment variables for age, race, and residence location. They show that, for this example, OLS implicitly gives more weight to the causal effect of shifting from 13 to 14 years of schooling and from 14 to 15 years of schooling than for much more common differences such as the shift from 11 to 12 years of schooling (primarily because the net conditional unexplained variance of schooling is greatest for the contrasts between 13 and 14 years and between 14 and 15 years). They also show that, for this example, the piecewise increases in average earnings happen to be largest for the years of schooling that OLS systematically weights downward. The result is a least squares estimate under the linearity constraint of .094, which is smaller than the weighted average estimate of .144 that one can calculate by dropping the linearity constraint and then averaging year-specific estimates over the marginal distribution of years of schooling.

For other examples, the weighting schemes may not generate sufficiently different estimates, as the overall weighting is a complex function of the relationship between the unaccounted for variance of the causal variable within strata of the adjustment variables and the level of nonlinearity of the conditional expectation function. But the general point is clear and should be sobering: Linearity constraints across causal states may lead OLS models to generate nonintuitive (and sometimes misleading) averages of otherwise easily interpretable stratum-specific causal effects.

### 5.4.2 Data Mining and the Challenge of Regression Specification

In this subsection, we discuss the considerable appeal of what can be called the all-cause, complete-specification tradition of regression analysis. We argue that this orientation is impractical for most of the social sciences, for which theory is too weak and the disciplines too contentious to furnish perfect specifications that can be agreed on. At the same time, we argue that inductive approaches to discovering flawless regression models that represent all causes are mostly a form of self-deception, even though some computer programs now exist that can prevent the worst forms of abuse.

Consider first a scenario in which one has a theoretical model that one believes is true. It suggests all of the inputs that determine the outcome of interest, as a set of observable variables, and it is in the form of a specific function that relates all inputs to the outcome. In this case, one can claim to have the correct specification for a regression of the outcome on some function of the variables suggested by the theoretical model. The only remaining challenges are then measurement, sampling, and observation.

The weakness of this approach is that critics can claim that the model is not true and hence that the entailed regression specification is wrong. Fighting off any such critics with empirical results can then be difficult, given that the regression specification used to generate the empirical results has been called into question.

In general, if members of a community of competing researchers assert their own true models and then offer up purportedly flawless regression models, the result may be a war of attrition in which no scientific progress is possible. It is therefore natural to ask: Can the *data* generate an all-cause, complete-specification regression model that all competing researchers can jointly adopt?

The first step in answering this question is to determine what an all-cause, complete specification would be, which is sometimes simply labeled a "correct specification."[23] In his 1978 book, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Edward Leamer lays out the following components of what he labels The Axiom of Correct Specification:

> (a) The set of explanatory variables that are thought to determine (linearly) the dependent variable must be (1) unique, (2) complete, (3) small in number, and (4) observable. (b) Other determinants of the dependent variable must have a probability distribution with at most a few unknown parameters. (3) All unknown parameters must be constant. (Leamer 1978:4)

But Leamer then immediately undermines the axiom as it applies to observational data analysis in the social sciences:

> If this axiom were, in fact, accepted, we would find one equation estimated for every phenomenon, and we would have books that compiled these estimates published with the same scientific fanfare that accompanies estimates of the speed of light or the gravitational constant. Quite the contrary, we are literally deluged with regression equations, all offering to "explain" the same event, and instead of a book of findings we have volumes of competing estimates. (Leamer 1978:4)

One can quibble with Leamer's axiom [e.g., that component (3) is not essential and so on], but the literature seems to provide abundant support for his conclusion. Few examples of flawless regression models suggested by true theoretical models can be found in the social science literature. One might hope for such success in the future, but the past 40 years of research do not give much reason for optimism.

Leamer instead argues that most regression models are produced by what he labels a data-instigated specification search, which he characterizes as a Sherlock Holmes form of inference wherein one refrains from developing a model

[23] The literature has never clearly settled on a definition that has achieved consensus, but Leamer's is as good as any. Part of the confusion arises from the recognition that a descriptively motivated regression model is always correct, no matter what its specification happens to be.

or any firm hypotheses before first considering extensively all the facts of a case. Leamer argues that this approach to variable selection and specification is fraught with potential danger and invalidates traditional notions of inference.

Consider the example of the Catholic school effect on learning, and in particular the research of James Coleman and his colleagues. In seeking to estimate the effect of Catholic schooling on achievement, Coleman did not draw a complete specification for his regression models from a specific theoretical model of human learning. This decision was not because no such models existed, nor because Coleman had no appreciation for the need for such models. He was, in contrast, well aware of classic behaviorist models of learning (see Bush and Mosteller 1955, 1959) that specified complex alternative mechanisms for sequences of responses to learning trials. Although he appreciated these models, he recognized (see Coleman 1964:38) that they could not be deployed effectively in the complex environments of secondary schooling in the United States, the context of which he had already studied extensively (see Coleman 1961).

As a result, Coleman did not specify a learning model that justified the regression models that he and his colleagues presented (see Sørensen 1998; Sørensen and Morgan 2000).[24] Their basic specification strategy was instead to attempt to adjust for a sufficient subset of other causes of learning so that, net of these effects, it could be claimed that Catholic and public school students were sufficiently equivalent. The specific variables that Coleman and his colleagues chose to include in their models were based in part on Coleman's deep knowledge of what predicts learning in high school (and one could argue that Coleman was the most knowledgeable social scientist on the topic in the world at the time). But he and his colleagues also adopted an empirical approach, as indicated parenthetically at the end of the following account of their selection of adjustment variables:

> In order to minimize the effects of differences in initial selection masquerading as effects of differences in the sectors themselves, achievement subtests were regressed, by sector and grade, on a larger number of background variables that measure both objective and subjective differences in the home. Some of these subjective differences may not be prior to the student's achievement, but may in part be consequences of it, so that there may be an overcompensation for

[24] When the 1982 data on seniors became available to supplement the 1980 data on sophomores, Coleman and his colleagues did move toward a stronger foundation for their specifications, providing an underlying model for the lagged achievement gain regression model that was an outgrowth of Coleman's early work on Markov chains and his proposals for longitudinal data analysis (Coleman 1964, 1981). In Hoffer et al. (1985:89–91), he and his colleagues showed that (subject to restrictions on individual heterogeneity) the lagged test score model is a linearized reduced-form model of two underlying rates (learning and forgetting) for the movement between two states (know and don't know) for each item on the cognitive test. Although the model is plausible, it is clearly constrained so that it can be estimated with simple regression techniques (see Coleman 1981:8–9 for an explanation of his *modus operandi* in such situations), and this is of course not the sort of constraint that one must adopt if one is truly interested in laying out the correct theoretical model of learning.

> background differences. It was felt desirable to do this so as to compensate for possible unmeasured differences in family background; but of course the results may be to artificially depress the resulting levels of background-controlled achievement in Catholic and other private schools. (A few additional background variables were initially included; those that showed no effects beyond the ones listed in the following paragraph were eliminated from the analysis.) (Coleman et al. 1982:147)

Coleman and his colleagues then reported that the final list of variables included 10 they considered "clearly prior" to school sector – including family income, parents' education, number of siblings, and number of rooms in the home – as well as 7 other variables that they considered "not clearly prior" to school sector – including more than 50 books in the home, owning a pocket calculator, and having a mother who thinks the student should go to college after high school.

As so often occurs in causal controversies of public importance, critics found this resulting list inadequate. From the perspective of their critics, Coleman and his colleagues had not provided a clear enough accounting of why some students were observed in Catholic schools whereas others were observed in public schools and why levels of learning should be considered a linear function of background and the specific school characteristics selected. After arguing that more would be known when follow-up data were collected and test score gains from sophomore to senior year could be analyzed, Alexander and Pallas (1983) argued that Coleman and his colleagues should have searched harder for additional adjustment variables:

> Failing this [estimating models with pretest and posttest data], another possibility would be to scout about for additional controls that might serve as proxies for student input differences that remain after socioeconomic adjustments. One candidate is the student's curriculum placement in high school. (Alexander and Pallas 1983:171)

Alexander and Pallas then laid out a rationale for this proxy approach, and they offered models that showed that the differences between public and private schools are smaller after conditioning on type of curriculum.

As this example shows, it is often simply unclear how one should go about selecting a sufficient set of conditioning variables to include in a regression equation when adopting the "adjustment for all other causes" approach to causal inference. Coleman and colleagues clearly included some variables that they believed that perhaps they should not, and they presumably tossed out some variables that they thought they should perhaps include but that proved to be insufficiently powerful predictors of test scores. Even so, Alexander and Pallas criticized Coleman and his colleagues for too little scouting.[25]

[25] Contrary to the forecasts of Coleman and his critics, after the 1982 data were released, the specification debate did not end. It simply moved on to new concerns, primarily how

Leamer, as mentioned earlier, would characterize such scouting as a Sherlock-Holmes-style, data-driven specification search. Leamer argues that this search strategy turns classical inference on its head:

> ... if theories are constructed after having studied the data, it is difficult to establish by how much, if at all, the data favor the data-instigated hypothesis. For example, suppose I think that a certain coefficient ought to be positive, and my reaction to the anomalous result of a negative estimate is to find another variable to include in the equation so that the estimate is positive. Have I found evidence that the coefficient is positive? (Leamer 1983:40)

Taken to its extreme, the Sherlock Holmes regression approach may discover relationships between candidate independent variables and the outcome variable that are due to sampling variability and nothing else. David Freedman showed this possibility in a simple simulation exercise, in which he sought to demonstrate that "... in a world with a large number of unrelated variables and no clear a priori specifications, uncritical use of standard [regression] methods will lead to models that appear to have a lot of explanatory power" (Freedman 1983:152). To show the plausibility of this conclusion, Freedman constructed an artificial dataset with 100 individuals, one outcome variable $Y$, and 50 other variables $X_1$ through $X_{50}$. The 100 values for each of these 51 variables were then independent random draws from the standard normal distribution. Thus, the data represent complete noise with only chance dependencies between the variables that mimic what any real-world sampling procedure would produce. The data were then subjected to regression analysis, with $Y$ regressed on $X_1$ through $X_{50}$. For these 50 variables, 1 variable yielded a coefficient with a $p$ value of less than .05 and another 14 had $p$ values of less than .25. Freedman then ran a second regression of $Y$ on the 15 variables that had $p$ values of less than .25, and in this second pass, 14 of them again turned up with $p$ values of less than .25. Most troubling, 6 of them now had $p$ values of less than .05, and the model as a whole had an $R^2$ of .36. From pure noise and simulated sampling variability, Freedman produced a regression model that looks similar to any number of those published in social science articles. It had six coefficients that passed conventional standards of statistical significance, and it explained a bit more than one third of the variance of the outcome variable.[26]

The danger of data-driven specification searches is important to recognize, but not all procedures are similarly in danger, especially given developments since Leamer first presented his critique in the 1970s and 1980s. There is a new literature on data mining and statistical learning that has devised techniques to avoid the problems highlighted by Freedman's simulation (see Hastie,

to adjust for sophomore test scores (with or without a family background adjustment, with or without curriculum differences, with only a subset of sophomore test scores, and with or without adjustment for attenuation that is due to measurement error).

[26] Raftery (1995) repeated Freedman's simulation experiment and obtained even more dramatic results.

Tibshirani, and Friedman 2001). For a very clear overview of these methods, see Berk (2006, 2007).[27]

Even so, data-instigated specifications of regression equations remain a problem in practice, because few applied social scientists use the fair and disciplined algorithms in the statistical learning literature. The Catholic school example is surely a case in which scouting led to the inclusion of variables that may not have been selected by a statistical learning algorithm. But, nonetheless, none of the scholars in the debate dared to reason backwards from their regression models in order to declare that they had inductively constructed a true model of learning. And, in general, it is hard to find examples of complete inductive model building in the published literature; scholars are usually driven by some theoretical predilections, and the results of mistaken induction are often fragile enough to be uncovered in the peer review process.[28] Milder forms of misspecification are surely pervasive.

## 5.5 Conclusions

Regression models, in their many forms, remain one of the most popular techniques for the evaluation of alternative explanations in the social sciences. In this chapter, we have restricted most of our attention to OLS regression of an interval-scaled variable on a binary causal variable. And, although we have considered how regression modeling can be used as a descriptive data reduction tool, we have focused mostly on regression as a parametric adjustment technique for estimating causal effects, while also presenting the deep connections between regression and matching as complementary forms of a more general conditioning estimation strategy.

We conclude this chapter by discussing the strengths and weaknesses of regression as a method for causal inference from observational data. The main strengths of regression analysis are clearly its computational simplicity, its myriad forms, its familiarity to a wide range of social scientists, and the ease with which one can induce computer software to generate standard errors. These are all distinct advantages over the matching techniques that we summarized in Chapter 4.

[27] And, as we noted earlier in Chapter 4, there are cases in which a data-driven specification search is both permissive and potentially quite useful. Consider again the causal graph in Figure 3.10 and suppose that one has a large number of variables that may be associated with both $D$ and $Y$ in one's dataset and that one presumes may be members of either $S$ or $X$. Accordingly, one has the choice of conditioning on two different types of variables that lie along the back-door path from $D$ to $Y$: the variables in $S$ that predict $D$ or the variables in $X$ that predict $Y$. Engaging in a data-driven specification search for variables that predict $Y$ will fall prey to inferential difficulties about the causal effect of $D$ on $Y$ for exactly the reasons just discussed. But a data-driven specification search for variables that predict $D$ will not fall prey to the same troubles, because in this search one does not use any direct information about the outcome $Y$.

[28] That being said, predictions about the behavior of financial markets can come close. See Krueger and Kennedy (1990) for discussion and interpretation of the apparent effect of Super Bowl victories on the U.S. stock market.

But, as we have shown in this chapter, regression models have some serious weaknesses. Their ease of estimation tends to suppress attention to features of the data that matching techniques force researchers to consider, such as the potential heterogeneity of the causal effect and the alternative distributions of covariates across those exposed to different levels of the cause. Moreover, the traditional exogeneity assumption of regression (e.g., in the case of least squares regression that the independent variables must be uncorrelated with the regression error term) often befuddles applied researchers who can otherwise easily grasp the stratification and conditioning perspective that undergirds matching. As a result, regression practitioners can too easily accept their hope that the specification of plausible control variables generates an as-if randomized experiment.

Focusing more narrowly on least squares models, we have shown through several demonstrations that they generate causal effect estimates that are both nonintuitive and inappropriate when consequential heterogeneity has not been fully parameterized. In this sense, the apparent simplicity of least squares regression belies the complexity of how the data are reduced to a minimum mean-squared-error linear prediction. For more complex regression models, the ways in which such heterogeneity is implicitly averaged are presently unknown. But no one seems to suspect that the complications of unparameterized heterogeneity are less consequential for fancier maximum-likelihood-based regression models in the general linear modeling tradition.

Our overall conclusion is thus virtually the same as for matching: Regression is a statistical method for analyzing available data, and for the estimation of causal effects it may have some advantages in some situations. The joint implication of these conclusions for causal analysis is that matching and regression are probably best used together, or at least used in ways such that the distinction between them fades away.