Chapter Title: Questions about Questions

Book Title: Mostly Harmless Econometrics

Book Subtitle: An Empiricist's Companion

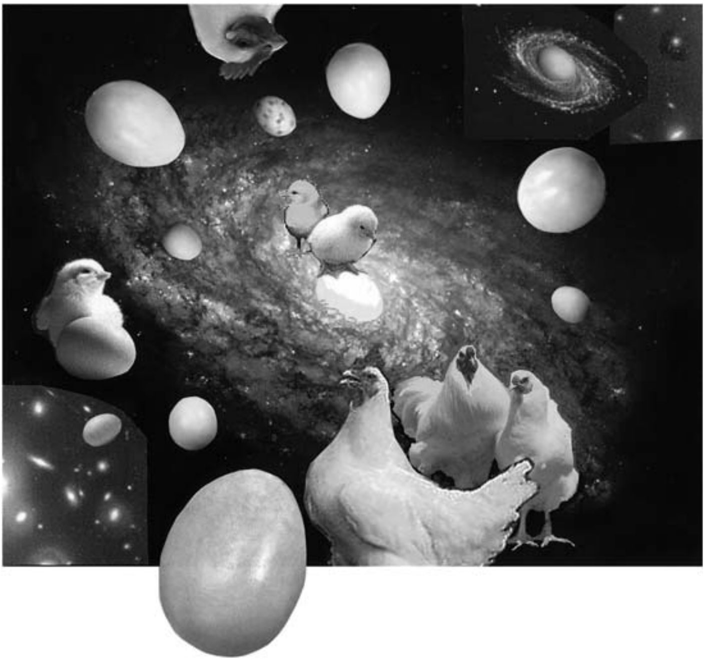
Book Author(s): Joshua D. Angrist and Jörn-Steffen Pischke

Published by: Princeton University Press

Stable URL: https://www.jstor.org/stable/j.ctvcm4j72.8

Part I

Preliminaries



Chapter 1

Questions about *Questions*

"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

This chapter briefly discusses the basis for a successful research project. Like the biblical story of Exodus, a research agenda can be organized around four questions. We call these frequently asked questions (FAQs), because they should be. The FAQs ask about the relationship of interest, the ideal experiment, the identification strategy, and the mode of inference.

In the beginning, we should ask, *What is the causal relationship of interest?* Although purely descriptive research has an important role to play, we believe that the most interesting research in social science is about questions of cause and effect, such as the effect of class size on children's test scores, discussed in chapters 2 and 6. A causal relationship is useful for making predictions about the consequences of changing circumstances or policies; it tells us what would happen in alternative (or "counterfactual") worlds. For example, as part of a research agenda investigating human productive capacity—what labor economists call human capital—we have both investigated the causal effect of schooling on wages (Card, 1999, surveys research in this area). The causal effect of schooling on wages is the increment to wages an individual would receive if he or she got more schooling. A range of studies suggest the causal effect of a college degree is about 40 percent higher wages on average, quite a payoff. The causal effect of schooling on wages is useful for predicting the earnings consequences of, say, changing the costs of attending college, or strengthening compulsory attendance laws. This relation is also of theoretical interest since it can be derived from an economic model.

As labor economists, we're most likely to study causal effects in samples of workers, but the unit of observation in causal research need not be an individual human being. Causal questions can be asked about firms or, for that matter, countries. Take, for example, Acemoglu, Johnson, and Robinson's (2001) research on the effect of colonial institutions on economic growth. This study is concerned with whether countries that inherited more democratic institutions from their colonial rulers later enjoyed higher economic growth as a consequence. The answer to this question has implications for our understanding of history and for the consequences of contemporary development policy. Today, we might wonder whether newly forming democratic institutions are important for economic development in Iraq and Afghanistan. The case for democracy is far from clear-cut; at the moment, China is enjoying robust economic growth without the benefit of complete political freedom, while much of Latin America has democratized without a big growth payoff.

The second research FAQ is concerned with *the experiment that could ideally be used to capture the causal effect of interest*. In the case of schooling and wages, for example, we can imagine offering potential dropouts a reward for finishing school, and then studying the consequences. In fact, Angrist and Lavy (2008) have run just such an experiment. Although their study looked at short-term effects such as college enrollment, a longer-term follow-up might well look at wages. In the case of political institutions, we might like to go back in time and randomly assign different government structures in former colonies on their independence day (an experiment that is more likely to be made into a movie than to get funded by the National Science Foundation).

Ideal experiments are most often hypothetical. Still, hypothetical experiments are worth contemplating because they help us pick fruitful research topics. We'll support this claim by asking you to picture yourself as a researcher with no budget constraint and no Human Subjects Committee policing your inquiry for social correctness: something like a well-funded Stanley Milgram, the psychologist who did pathbreaking work on the response to authority in the 1960s using highly controversial experimental designs that would likely cost him his job today.

Seeking to understand the response to authority, Milgram (1963) showed he could convince experimental subjects to administer painful electric shocks to pitifully protesting victims (the shocks were fake and the victims were actors). This turned out to be controversial as well as clever: some psychologists claimed that the subjects who administered shocks were psychologically harmed by the experiment. Still, Milgram's study illustrates the point that there are many experiments we can think about, even if some are better left on the drawing board.[1] If you can't devise an experiment that answers your question in a world where anything goes, then the odds of generating useful results with a modest budget and nonexperimental survey data seem pretty slim. The description of an ideal experiment also helps you formulate causal questions precisely. The mechanics of an ideal experiment highlight the forces you'd like to manipulate and the factors you'd like to hold constant.

Research questions that cannot be answered by any experiment are FUQs: fundamentally unidentified questions. What exactly does a FUQ look like? At first blush, questions about the causal effect of race or gender seem good candidates because these things are hard to manipulate in isolation ("imagine your chromosomes were switched at birth"). On the other hand, the issue economists care most about in the realm of race and sex, labor market discrimination, turns on whether someone treats you differently because they *believe* you to be black or white, male or female. The notion of a counterfactual world where men are perceived as women or vice versa has a long history and does not require Douglas Adams-style outlandishness to entertain (Rosalind disguised

[1] Milgram was later played by the actor William Shatner in a TV special, an honor that no economist has yet received, though Angrist is still hopeful.

as Ganymede fools everyone in Shakespeare's *As You Like It*). The idea of changing race is similarly near-fetched: in *The Human Stain*, Philip Roth imagines the world of Coleman Silk, a black literature professor who passes as white in professional life. Labor economists imagine this sort of thing all the time. Sometimes we even construct such scenarios for the advancement of science, as in audit studies involving fake job applicants and résumés.[2]

A little imagination goes a long way when it comes to research design, but imagination cannot solve every problem. Suppose that we are interested in whether children do better in school by virtue of having started school a little older. Maybe the 7-year-old brain is better prepared for learning than the 6-year-old brain. This question has a policy angle coming from the fact that, in an effort to boost test scores, some school districts are now imposing older start ages (Deming and Dynarski, 2008). To assess the effects of delayed school entry on learning, we could randomly select some kids to start first grade at age 7, while others start at age 6, as is still typical. We are interested in whether those held back learn more in school, as evidenced by their elementary school test scores. To be concrete, let's look at test scores in first grade.

The problem with this question—the effects of start age on first grade test scores—is that the group that started school at age 7 is . . . older. And older kids tend to do better on tests, a pure maturation effect. Now, it might seem we can fix this by holding age constant instead of grade. Suppose we wait to test those who started at age 6 until second grade and test those who started at age 7 in first grade, so that everybody is tested at age 7. But the first group has spent more time in school, a fact that raises achievement if school is worth anything. There is no way to disentangle the effect of start age on learning from maturation and time-in-school effects as long as kids are still in school. The problem here is that for students, start age

[2] A recent example is Bertrand and Mullainathan (2004), who compared employers' reponses to résumés with blacker-sounding and whiter-sounding first names, such as Lakisha and Emily (though Fryer and Levitt, 2004, note that names may carry information about socioeconomic status as well as race.)

equals current age minus time in school. This deterministic link disappears in a sample of adults, so we can investigate pure start-age effects on adult outcomes, such as earnings or highest grade completed (as in Black, Devereux, and Salvanes, 2008). But the effect of start age on elementary school test scores is impossible to interpret even in a randomized trial, and therefore, in a word, FUQed.

The third and fourth research FAQs are concerned with the nuts-and-bolts elements that produce a specific study. Question number 3 asks, *What is your identification strategy?* Angrist and Krueger (1999) used the term *identification strategy* to describe the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment. Returning to the schooling example, Angrist and Krueger (1991) used the interaction between compulsory attendance laws in American states and students' season of birth as a natural experiment to estimate the causal effects of finishing high school on wages (season of birth affects the degree to which high school students are constrained by laws allowing them to drop out after their 16th birthday). Chapters 3–6 are primarily concerned with conceptual frameworks for identification strategies.

Although a focus on credible identification strategies is emblematic of modern empirical work, the juxtaposition of ideal and natural experiments has a long history in econometrics. Here is our econometrics forefather, Trygve Haavelmo (1944, p. 14), appealing for more explicit discussion of both kinds of experimental designs:

> A design of experiments (a prescription of what the physicists call a "crucial experiment") is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although—unfortunately—most economists do not describe their design of experiments explicitly. If they did, they would see that the experiments they have in mind may be grouped into two different classes, namely, (1) experiments that *we should like to make* to see if certain real economic phenomena—when artificially isolated from "other influences"—would verify certain

> hypotheses, and (2) the stream of experiments that Nature is steadily turning out from her own enormous laboratory, and which we merely watch as passive observers. In both cases the aim of the theory is the same, to become master of the happenings of real life.

The fourth research FAQ borrows language from Rubin (1991): *What is your mode of statistical inference?* The answer to this question describes the population to be studied, the sample to be used, and the assumptions made when constructing standard errors. Sometimes inference is straightforward, as when you use census microdata samples to study the American population. Often inference is more complex, however, especially with data that are clustered or grouped. The last chapter covers practical problems that arise once you've answered question number 4. Although inference issues are rarely very exciting, and often quite technical, the ultimate success of even a well-conceived and conceptually exciting project turns on the details of statistical inference. This sometimes dispiriting fact inspired the following econometrics haiku, penned by Keisuke Hirano after completing his thesis:

T-stat looks too good
Try clustered standard errors—
Significance gone

As should be clear from the above discussion, the four research FAQs are part of a process of project development. The following chapters are concerned mostly with the econometric questions that come up after you've answered the research FAQs—in other words, issues that arise once your research agenda has been set. Before turning to the nuts and bolts of empirical work, however, we begin with a more detailed explanation of why randomized trials give us our benchmark.

This page intentionally left blank

Chapter Title: The Experimental Ideal

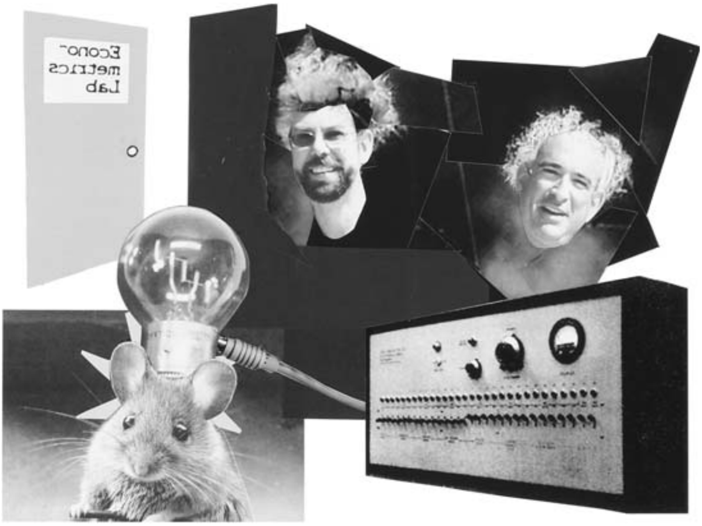
Book Title: Mostly Harmless Econometrics

Book Subtitle: An Empiricist's Companion

Book Author(s): Joshua D. Angrist and Jörn-Steffen Pischke

Published by: Princeton University Press

Stable URL: https://www.jstor.org/stable/j.ctvcm4j72.9


Econo-
metrics
Lab

Chapter 2

The Experimental Ideal

It is an important and popular fact that things are not always what they seem. For instance, on the planet Earth, man had always assumed that he was more intelligent than dolphins because he had achieved so much—the wheel, New York, wars and so on—while all the dolphins had ever done was muck about in the water having a good time. But conversely, the dolphins had always believed that they were far more intelligent than man—for precisely the same reasons. In fact there was only one species on the planet more intelligent than dolphins, and they spent a lot of their time in behavioral research laboratories running round inside wheels and conducting frighteningly elegant and subtle experiments on man. The fact that once again man completely misinterpreted this relationship was entirely according to these creatures' plans.

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

The most credible and influential research designs use random assignment. A case in point is the Perry preschool project, a 1962 randomized experiment designed to assess the effects of an early intervention program involving 123 black preschoolers in Ypsilanti, Michigan. The Perry treatment group was randomly assigned to an intensive intervention that included preschool education and home visits. It's hard to exaggerate the impact of the small but well-designed Perry experiment, which generated follow-up data through 1993 on the participants at age 27. Dozens of academic studies cite or use the Perry findings (see, e.g., Barnett, 1992). Most important, the Perry project provided the intellectual basis for the massive Head Start preschool program, begun in 1964, which ultimately served (and continues to serve) millions of American children.[1]

2.1 The Selection Problem

We take a brief time-out for a more formal discussion of the role experiments play in uncovering causal effects. Suppose you are interested in a causal if-then question. To be concrete, let us consider a simple example: Do hospitals make people healthier? For our purposes, this question is allegorical, but it is surprisingly close to the sort of causal question health economists care about. To make this question more realistic, let's imagine we're studying a poor elderly population that uses hospital emergency rooms for primary care. Some of these patients are admitted to the hospital. This sort of care is expensive, crowds hospital facilities, and is, perhaps, not very effective (see, e.g., Grumbach, Keane, and Bindman, 1993). In fact, exposure to other sick patients by those who are themselves vulnerable might have a net negative impact on their health.

Since those admitted to the hospital get many valuable services, the answer to the hospital effectiveness question still seems likely to be yes. But will the data back this up? The natural approach for an empirically minded person is to compare the health status of those who have been to the hospital with the health of those who have not. The National Health Interview Survey (NHIS) contains the information needed to make this comparison. Specifically, it includes a question, "During the past 12 months, was the respondent a patient in a hospital overnight?" which we can use to identify recent hospital visitors. The NHIS also asks, "Would you say your health in general is excellent, very good, good, fair, poor?"

[1] The Perry data continue to get attention, particularly as policy interest has returned to early education. A recent reanalysis by Michael Anderson (2008) confirmed many of the findings from the original Perry study, though Anderson also shows that the overall positive effects of the Perry project are driven entirely by the impact on girls. The Perry intervention seems to have done nothing for boys.

The following table displays the mean health status (assigning a 1 to poor health and a 5 to excellent health) among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean Health Status	Std. Error
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

The difference in means is 0.72, a large and highly significant contrast in favor of the nonhospitalized, with a *t*-statistic of 58.9.

Taken at face value, this result suggests that going to the hospital makes people sicker. It's not impossible this is the right answer since hospitals are full of other sick people who might infect us and dangerous machines and chemicals that might hurt us. Still, it's easy to see why this comparison should not be taken at face value: people who go to the hospital are probably less healthy to begin with. Moreover, even after hospitalization people who have sought medical care are not as healthy, on average, as those who were never hospitalized in the first place, though they may well be better off than they otherwise would have been.

To describe this problem more precisely, we can think about hospital treatment as described by a binary random variable, $\mathrm{D}_i = \{0, 1\}$. The outcome of interest, a measure of health status, is denoted by Y_i. The question is whether Y_i is *affected* by hospital care. To address this question, we assume we can imagine what might have happened to someone who went to the hospital if that person had not gone, and vice versa. Hence, for any individual there are two potential health variables:

$$\textit{Potential outcome} = \begin{cases} \mathrm{Y}_{1i} & \text{if } \mathrm{D}_i = 1 \\ \mathrm{Y}_{0i} & \text{if } \mathrm{D}_i = 0 \end{cases}.$$

In other words, Y_{0i} is the health status of an individual had he not gone to the hospital, irrespective of whether he actually went, while Y_{1i} is the individual's health status if he goes. We would like to know the difference between Y_{1i} and Y_{0i}, which can be said to be the causal effect of going to the hospital for individual i. This is what we would measure if we could go back in time and change a person's treatment status.[2]

The observed outcome, Y_i, can be written in terms of potential outcomes as

$$\mathrm{Y}_i = \begin{cases} \mathrm{Y}_{1i} & \text{if } \mathrm{D}_i = 1 \\ \mathrm{Y}_{0i} & \text{if } \mathrm{D}_i = 0 \end{cases}$$
$$= \mathrm{Y}_{0i} + (\mathrm{Y}_{1i} - \mathrm{Y}_{0i})\mathrm{D}_i. \qquad (2.1.1)$$

This notation is useful because $\mathrm{Y}_{1i} - \mathrm{Y}_{0i}$ is the causal effect of hospitalization for an individual. In general, there is likely to be a distribution of both Y_{1i} and Y_{0i} in the population, so the treatment effect can be different for different people. But because we never see both potential outcomes for any one person, we must learn about the effects of hospitalization by comparing the average health of those who were and were not hospitalized.

A naive comparison of averages by hospitalization status tells us something about potential outcomes, though not necessarily what we want to know. The comparison of average health conditional on hospitalization status is formally linked to the average causal effect by the equation:

$$\underbrace{E[\mathrm{Y}_i|\mathrm{D}_i = 1] - E[\mathrm{Y}_i|\mathrm{D}_i = 0]}_{\text{Observed difference in average health}} = \underbrace{E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1]}_{\text{Average treatment effect on the treated}}$$
$$+ \underbrace{E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 0]}_{\text{Selection bias}}.$$

The term

$$E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1]$$

is the *average causal effect of hospitalization on those who were hospitalized*. This term captures the averages difference between the health of the hospitalized, $E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1]$, and what would have happened to them had they not been hospitalized,

[2] The potential outcomes idea is a fundamental building block in modern research on causal effects. Important references developing this idea are Rubin (1974, 1977) and Holland (1986), who refers to a causal framework involving potential outcomes as the Rubin causal model.

$E[\text{Y}_{0i}|\text{D}_i = 1]$. The observed difference in health status, however, adds to this causal effect a term called *selection bias*. This term is the difference in average Y_{0i} between those who were and those who were not hospitalized. Because the sick are more likely than the healthy to seek treatment, those who were hospitalized have worse values of Y_{0i}, making selection bias negative in this example. The selection bias may be so large (in absolute value) that it completely masks a positive treatment effect. The goal of most empirical economic research is to overcome selection bias, and therefore to say something about the causal effect of a variable like D_i.[3]

2.2 Random Assignment Solves the Selection Problem

Random assignment of D_i solves the selection problem because random assignment makes D_i independent of potential outcomes. To see this, note that

$$\begin{aligned} E[\text{Y}_i|\text{D}_i = 1] - E[\text{Y}_i|\text{D}_i = 0] &= E[\text{Y}_{1i}|\text{D}_i = 1] - E[\text{Y}_{0i}|\text{D}_i = 0] \\ &= E[\text{Y}_{1i}|\text{D}_i = 1] - E[\text{Y}_{0i}|\text{D}_i = 1], \end{aligned}$$

where the independence of Y_{0i} and D_i allows us to swap $E[\text{Y}_{0i}|\text{D}_i = 1]$ for $E[\text{Y}_{0i}|\text{D}_i = 0]$ in the second line. In fact, given random assignment, this simplifies further to

$$\begin{aligned} E[\text{Y}_{1i}|\text{D}_i = 1] - E[\text{Y}_{0i}|\text{D}_i = 1] &= E[\text{Y}_{1i} - \text{Y}_{0i}|\text{D}_i = 1] \\ &= E[\text{Y}_{1i} - \text{Y}_{0i}]. \end{aligned}$$

The effect of randomly assigned hospitalization on the hospitalized is the same as the effect of hospitalization on a randomly chosen patient. The main thing, however, is that random assignment of D_i eliminates selection bias. This does not mean that randomized trials are problem-free, but in principle they solve the most important problem that arises in empirical research.

[3] This section marks our first use of the conditional expectation operator (e.g., $E[\text{Y}_i|\text{D}_i = 1]$ and $E[\text{Y}_i|\text{D}_i = 0]$). We use this to denote the population (or infinitely large sample) average of one random variable with the value of another held fixed. A more formal and detailed definition appears in Chapter 3.

How relevant is our hospitalization allegory? Experiments often reveal things that are not what they seem on the basis of naive comparisons alone. A recent example from medicine is the evaluation of hormone replacement therapy (HRT). This is a medical intervention that was recommended for middle-aged women to reduce menopause symptoms. Evidence from the Nurses Health Study, a large and influential nonexperimental survey of nurses, showed better health among HRT users. In contrast, the results of a recently completed randomized trial showed few benefits of HRT. Worse, the randomized trial revealed serious side effects that were not apparent in the nonexperimental data (see, e.g., Women's Health Initiative [WHI], Hsia et al., 2006).

An iconic example from our own field of labor economics is the evaluation of government-subsidized training programs. These are programs that provide a combination of classroom instruction and on-the-job training for groups of disadvantaged workers such as the long-term unemployed, drug addicts, and ex-offenders. The idea is to increase employment and earnings. Paradoxically, studies based on nonexperimental comparisons of participants and nonparticipants often show that after training, the trainees earn less than plausible comparison groups (see, e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Lalonde 1995). Here, too, selection bias is a natural concern, since subsidized training programs are meant to serve men and women with low earnings potential. Not surprisingly, therefore, simple comparisons of program participants with nonparticipants often show lower earnings for the participants. In contrast, evidence from randomized evaluations of training programs generate mostly positive effects (see, e.g., Lalonde, 1986; Orr et al., 1996).

Randomized trials are not yet as common in social science as in medicine, but they are becoming more prevalent. One area where the importance of random assignment is growing rapidly is education research (Angrist, 2004). The 2002 Education Sciences Reform Act passed by the U.S. Congress mandates the use of rigorous experimental or quasi-experimental research designs for all federally funded education studies. We can therefore expect to see many more randomized trials in education research in the years to come. A pioneering randomized study from the field of education is the Tennessee STAR experiment, designed to estimate the effects of smaller classes in primary school.

Labor economists and others have a long tradition of trying to establish causal links between features of the classroom environment and children's learning, an area of investigation that we call "education production." This terminology reflects the fact that we think of features of the school environment as inputs that cost money, while the output that schools produce is student learning. A key question in research on education production is which inputs produce the most learning given their costs. One of the most expensive inputs is class size, since smaller classes can only be achieved by hiring more teachers. It is therefore important to know whether the expense of smaller classes has a payoff in terms of higher student achievement. The STAR experiment was meant to answer this question.

Many studies of education production using nonexperimental data suggest there is little or no link between class size and student learning. So perhaps school systems can save money by hiring fewer teachers, with no consequent reduction in achievement. The observed relation between class size and student achievement should not be taken at face value, however, since weaker students are often deliberately grouped into smaller classes. A randomized trial overcomes this problem by ensuring that we are comparing apples to apples, that is, that the students assigned to classes of different sizes are otherwise comparable. Results from the Tennessee STAR experiment point to a strong and lasting payoff to smaller classes (see Finn and Achilles, 1990, for the original study, and Krueger, 1999, for an econometric analysis of the STAR data).

The STAR experiment was unusually ambitious and influential, and therefore worth describing in some detail. It cost about $12 million and was implemented for a cohort of kindergartners in 1985–86. The study ran for four years, until the original cohort of kindergartners was in third grade, and involved about 11,600 children. The average class size in regular Tennessee classes in 1985–86 was about 22.3. The experiment assigned students to one of three treatments: small classes with 13–17 children, regular classes with 22–25 children and a part-time teacher's aide (the usual arrangement), or regular classes with a full-time teacher's aide. Schools with at least three classes in each grade could choose to participate in the experiment.

The first question to ask about a randomized experiment is whether the randomization successfully balanced subjects' characteristics across the different treatment groups. To assess this, it's common to compare pretreatment outcomes or other covariates across groups. Unfortunately, the STAR data fail to include any pretreatment test scores, though it is possible to look at characteristics of children such as race and age. Table 2.2.1, reproduced from Krueger (1999), compares the means of these variables. The student characteristics in the table are a free lunch variable, student race, and student age. Free lunch status is a good measure of family income, since only poor children qualify for a free school lunch. Differences in these characteristics across the three class types are small, and none is significantly different from zero, as indicated by the *p*-values in the last column. This suggests the random assignment worked as intended.

Table 2.2.1 also presents information on average class size, the attrition rate, and test scores, measured here on a percentile scale. The attrition rate (proportion of students lost to follow-up) was lower in small kindergarten classrooms. This is potentially a problem, at least in principle.[4] Class sizes are significantly lower in the assigned-to-be-small classrooms, which means that the experiment succeeded in creating the desired variation. If many of the parents of children assigned to regular classes had successfully lobbied teachers and principals to get their children assigned to small classes, the gap in class size across groups would be much smaller.

Because randomization eliminates selection bias, the difference in outcomes across treatment groups captures the average

[4] Krueger (1999) devotes considerable attention to the attrition problem. Differences in attrition rates across groups may result in a sample of students in higher grades that is not randomly distributed across class types. The kindergarten results, which were unaffected by attrition, are therefore the most reliable.

TABLE 2.2.1
Comparison of treatment and control characteristics in the Tennessee STAR experiment

	Class Size			
Variable	Small	Regular	Regular/Aide	*P*-value for equality across groups
Free lunch	.47	.48	.50	.09
White/Asian	.68	.67	.66	.26
Age in 1985	5.44	5.43	5.42	.32
Attrition rate	.49	.52	.53	.02
Class size in kindergarten	15.10	22.40	22.80	.00
Percentile score in kindergarten	54.70	48.90	50.00	.00

Notes: Adapted from Krueger (1999), table I. The table shows means of variables by treatment status for the sample of students who entered STAR in kindergarten. The *P*-value in the last column is for the *F*-test of equality of variable means across all three groups. The free lunch variable is the fraction receiving a free lunch. The percentile score is the average percentile score on three Stanford Achievement Tests. The attrition rate is the proportion lost to follow-up before completing third grade.

causal effect of class size (relative to regular classes with a part-time aide). In practice, the difference in means between treatment and control groups can be obtained from a regression of test scores on dummies for each treatment group, a point we expand on below. Regression estimates of treatment-control differences for kindergartners, reported in table 2.2.2 (derived from Krueger, 1999, table V), show a small-class effect of about five percentile points (other rows in the table show coefficients on control variables in the regressions). The effect size is about $.2\sigma$, where σ is the standard deviation of the percentile score in kindergarten. The small-class effect is significantly different from zero, while the regular/aide effect is small and insignificant.

The STAR study, an exemplary randomized trial in the annals of social science, also highlights the logistical difficulty, long duration, and potentially high cost of randomized trials.

TABLE 2.2.2
Experimental estimates of the effect of class size on test scores

Explanatory Variable	(1)	(2)	(3)	(4)
Small class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian	—	—	8.35 (1.35)	8.44 (1.36)
Girl	—	—	4.48 (.63)	4.39 (.63)
Free lunch	—	—	−13.15 (.77)	−13.07 (.77)
White teacher	—	—	—	−.57 (2.10)
Teacher experience	—	—	—	.26 (.10)
Teacher Master's degree	—	—	—	−0.51 (1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Notes: Adapted from Krueger (1999), table V. The dependent variable is the Stanford Achievement Test percentile score. Robust standard errors allowing for correlated residuals within classes are shown in parentheses. The sample size is 5,681.

In many cases, such trials are impractical.[5] In other cases, we would like an answer sooner rather than later. Much of

[5] Randomized trials are never perfect, and STAR is no exception. Pupils who repeated or skipped a grade left the experiment. Students who entered an experimental school one grade later were added to the experiment and randomly assigned to one of the classes. One unfortunate aspect of the experiment is that students in the regular and regular/aide classes were reassigned after the kindergarten year, possibly because of protests by the parents with children in the regular classrooms. There was also some switching of children after the kindergarten year. But Krueger's (1999) analysis suggests that none of these implementation problems affected the main conclusions of the study.

the research we do, therefore, attempts to exploit cheaper and more readily available sources of variation. We hope to find natural or quasi-experiments that mimic a randomized trial by changing the variable of interest while other factors are kept balanced. Can we always find a convincing natural experiment? Of course not. Nevertheless, we take the position that a notional randomized trial is our benchmark. Not all researchers share this view, but many do. We heard it first from our teacher and thesis advisor, Orley Ashenfelter, a pioneering proponent of experiments and quasi-experimental research designs in social science. Here is Ashenfelter (1991) assessing the credibility of the observational studies linking schooling and income:

> How convincing is the evidence linking education and income? Here is my answer: Pretty convincing. If I had to bet on what an ideal experiment would indicate, I bet that it would show that better educated workers earn more.

The quasi-experimental study of class size by Angrist and Lavy (1999) illustrates the manner in which nonexperimental data can be analyzed in an experimental spirit. The Angrist and Lavy study relied on the fact that in Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in a fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions, such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being “as good as randomly assigned.”

The Angrist-Lavy study compared students in grades with enrollments above and below bureaucratic class size cutoffs to construct well-controlled estimates of the effects of a sharp change in class size without the benefit of a real experiment. As in the Tennessee STAR study, the Angrist and Lavy (1999) results pointed to a strong link between class size and achievement. This was in marked contrast to naive analyses, also reported by Angrist and Lavy, based on simple comparisons between those enrolled in larger and smaller classes. These comparisons showed students in smaller classes doing worse on standardized tests. The hospital allegory of selection bias would therefore seem to apply to the class size question as well.[6]

2.3 Regression Analysis of Experiments

Regression is a useful tool for the study of causal questions, including the analysis of data from experiments. Suppose (for now) that the treatment effect is the same for everyone, say $Y_{1i} - Y_{0i} = \rho$, a constant. With constant treatment effects, we can rewrite (2.1.1) in the form

$$Y_i = \underset{\substack{\shortparallel \\ E(Y_{0i})}}{\alpha} + \underset{\substack{\shortparallel \\ (Y_{1i} - Y_{0i})}}{\rho} D_i + \underset{\substack{\shortparallel \\ Y_{0i} - E(Y_{0i}),}}{\eta_i,} \tag{2.3.1}$$

where η_i is the random part of Y_{0i}. Evaluating the conditional expectation of this equation with treatment status switched off and on gives

$$\begin{aligned} E[Y_i|D_i = 1] &= \alpha + \rho + E[\eta_i|D_i = 1] \\ E[Y_i|D_i = 0] &= \alpha + E[\eta_i|D_i = 0], \end{aligned}$$

so that

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \underbrace{\rho}_{\text{Treatment effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{Selection bias}}.$$

Thus, selection bias amounts to correlation between the regression error term, η_i, and the regressor, D_i. Since

$$E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0] = E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0],$$

this correlation reflects the difference in (no-treatment) potential outcomes between those who get treated and those who

[6] The Angrist-Lavy (1999) results turn up again in chapter 6, as an illustration of the quasi-experimental regression-discontinuity research design.

don't. In the hospital allegory, those who were treated had poorer health outcomes in the no-treatment state, while in the Angrist and Lavy (1999) study, students in smaller classes tended to have intrinsically lower test scores.

In the STAR experiment, where D_i is randomly assigned, the selection bias term disappears, and a regression of Y_i on D_i estimates the causal effect of interest, ρ. Table 2.2.2 shows different regression specifications, some of which include covariates other than the random assignment indicator, D_i. Covariates play two roles in regression analyses of experimental data. First, the STAR experimental design used conditional random assignment. In particular, assignment to classes of different sizes was random within schools but not across schools. Students attending schools of different types (say, urban versus rural) were a bit more or less likely to be assigned to a small class. The comparison in column 1 of table 2.2.2, which makes no adjustment for this, might therefore be contaminated by differences in achievement in schools of different types. To adjust for this, some of Krueger's regression models include school fixed effects, that is, a separate intercept for each school in the STAR data. In practice, the consequences of adjusting for school fixed effects is rather minor, but we wouldn't know this without taking a look. We have more to say about regression models with fixed effects in chapter 5.

The other controls in Krueger's table describe student characteristics such as race, age, and free lunch status. We saw before that these individual characteristics are balanced across class types, that is, they are not systematically related to the class size assignment of the student. If these controls, call them X_i, are uncorrelated with the treatment D_i, then they will not affect the estimate of ρ. In other words, estimates of ρ in the long regression,

$$\text{Y}_i = \alpha + \rho \text{D}_i + X_i'\gamma + \eta_i, \qquad (2.3.2)$$

will be close to estimates of ρ in the short regression, (2.3.1). This is a point we expand on in chapter 3.

Inclusion of the variables X_i, although not necessary in this case, may generate more precise estimates of the causal effect of interest. Notice that the standard error of the estimated treatment effects in column 3 is smaller than the corresponding standard error in column 2. Although the control variables, X_i, are uncorrelated with D_i, they have substantial explanatory power for Y_i. Including these control variables therefore reduces the residual variance, which in turn lowers the standard error of the regression estimates. Similarly, the standard errors of the estimates of ρ are reduced by the inclusion of school fixed effects because these too explain an important part of the variance in student performance. The last column adds teacher characteristics. Because teachers were randomly assigned to classes, and teacher characteristics have little to do with student achievement in these data, both the estimated effect of small classes and its standard error are unchanged by the addition of teacher variables.

Regression plays an exceptionally important role in empirical economic research. As we've seen in this chapter, regression is well-suited to the analysis of experimental data. In some cases, regression can also be used to approximate experiments in the absence of random assignment. But before we get into the important question of when a regression is likely to have a causal interpretation, it is useful to review a number of fundamental regression facts and properties. These facts and properties are reliably true for any regression, regardless of the motivation for running it.

Chapter Title: Making Regression Make Sense

Book Title: Mostly Harmless Econometrics

Book Subtitle: An Empiricist's Companion

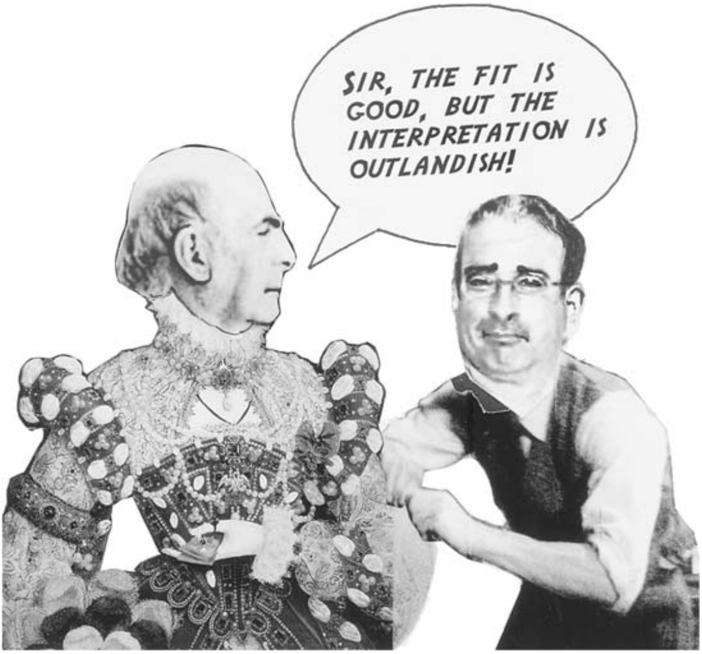
Book Author(s): Joshua D. Angrist and Jörn-Steffen Pischke

Published by: Princeton University Press

Stable URL: https://www.jstor.org/stable/j.ctvcm4j72.10

Part II

The Core


SIR, THE FIT IS GOOD, BUT THE INTERPRETATION IS OUTLANDISH!

Chapter 3

Making Regression Make Sense

"Let us think the unthinkable, let us do the undoable.
Let us prepare to grapple with the ineffable itself,
and see if we may not eff it after all."
Douglas Adams, *Dirk Gently's Holistic Detective Agency*

Angrist recounts:

I ran my first regression in the summer of 1979 between my freshman and sophomore years as a student at Oberlin College. I was working as a research assistant for Allan Meltzer and Scott Richard, faculty members at Carnegie-Mellon University, near my house in Pittsburgh. I was still mostly interested in a career in special education, and had planned to go back to work as an orderly in a state mental hospital, my previous summer job. But Econ 101 had got me thinking, and I could also see that at the same wage rate, a research assistant's hours and working conditions were better than those of a hospital orderly. My research assistant duties included data collection and regression analysis, though I did not understand regression or even statistics at the time.

The paper I was working on that summer (Meltzer and Richard, 1983) is an attempt to link the size of governments in democracies, measured as government expenditure over GDP, to income inequality. Most income distributions have a long right tail, which means that average income tends to be way above the median. When inequality grows, more voters find themselves with below-average incomes. Annoyed by this, those with incomes between the median and the average may join those with incomes below the median in voting for fiscal policies that take from the rich and give to the poor. The size of government consequently increases.

> I absorbed the basic theory behind the Meltzer and Richard project, though I didn't find it all that plausible, since voter turnout is low for the poor. I also remember arguing with my bosses over whether government expenditure on education should be classified as a public good (something that benefits everyone in society as well as those directly affected) or a private good publicly supplied, and therefore a form of redistribution like welfare. You might say this project marked the beginning of my interest in the social returns to education, a topic I went back to with more enthusiasm and understanding in Acemoglu and Angrist (2000).
>
> Today, I understand the Meltzer and Richard study as an attempt to use regression to uncover and quantify an interesting causal relation. At the time, however, I was purely a regression mechanic. Sometimes I found the RA work depressing. Days would go by when I didn't talk to anybody but my bosses and the occasional Carnegie-Mellon Ph.D. student, most of whom spoke little English anyway. The best part of the job was lunch with Allan Meltzer, a distinguished scholar and a patient and good-natured supervisor, who was happy to chat while we ate the contents of our brown bags (this did not take long, as Allan ate little and I ate fast). Once I asked Allan whether he found it satisfying to spend his days perusing regression output, which then came on reams of double-wide green-bar paper. Meltzer laughed and said there was nothing he would rather be doing.

Now we too spend our days happily perusing regression output, in the manner of our teachers and advisers in college and graduate school. This chapter explains why.

3.1 Regression Fundamentals

The end of the previous chapter introduced regression models as a computational device for the estimation of treatment-control differences in an experiment, with and without covariates. Because the regressor of interest in the class size study discussed in section 2.3 was randomly assigned, the resulting estimates have a causal interpretation. In most studies, however, regression is used with observational data. Without the benefit of random assignment, regression estimates may or may not have a causal interpretation. We return to the central question of what makes a regression causal later in this chapter.

Setting aside the relatively abstract causality problem for the moment, we start with the mechanical properties of regression estimates. These are universal features of the population regression vector and its sample analog that have nothing to do with a researcher's interpretation of his output. These properties include the intimate connection between the population regression function and the conditional expectation function and the sampling distribution of regression estimates.

3.1.1 Economic Relationships and the Conditional Expectation Function

Empirical economic research in our field of labor economics is typically concerned with the statistical analysis of individual economic circumstances, and especially differences between people that might account for differences in their economic fortunes. Differences in economic fortune are notoriously hard to explain; they are, in a word, random. As applied econometricians, however, we believe we can summarize and interpret randomness in a useful way. An example of "systematic randomness" mentioned in the introduction is the connection between education and earnings. On average, people with more schooling earn more than people with less schooling. The connection between schooling and earnings has considerable predictive power, in spite of the enormous variation in individual circumstances that sometimes clouds this fact. Of course, the fact that more educated people tend to earn more than less educated people does not mean that schooling *causes* earnings to increase. The question of whether the earnings-schooling relationship is causal is of enormous importance, and we come back to it many times. Even without resolving the difficult question of causality, however, it's clear that education predicts earnings in a narrow statistical sense. This predictive power is compellingly summarized by the conditional expectation function (CEF).

The CEF for a dependent variable Y_i, given a $\text{K} \times 1$ vector of covariates X_i (with elements x_{ki}), is the expectation, or population average, of Y_i, with X_i held fixed. The population average can be thought of as the mean in an infinitely large sample, or the average in a completely enumerated finite population. The CEF is written $E[\text{Y}_i|\text{X}_i]$ and is a function of X_i. Because X_i is random, the CEF is random, though sometimes we work with a particular value of the CEF, say $E[\text{Y}_i|\text{X}_i = 42]$, assuming 42 is a possible value for X_i. In chapter 2, we briefly considered the CEF $E[\text{Y}_i|\text{D}_i]$, where D_i is a zero-one variable. This CEF takes on two values, $E[\text{Y}_i|\text{D}_i = 1]$ and $E[\text{Y}_i|\text{D}_i = 0]$. Although this special case is important, we are most often interested in CEFs that are functions of many variables, conveniently subsumed in the vector X_i. For a specific value of X_i, say $\text{X}_i = x$, we write $E[\text{Y}_i|\text{X}_i = x]$. For continuous Y_i with conditional density $f_y(t|\text{X}_i = x)$ at $\text{Y}_i = t$, the CEF is

$$E[\text{Y}_i|\text{X}_i = x] = \int t f_y(t|\text{X}_i = x)dt.$$

If Y_i is discrete, $E[\text{Y}_i|\text{X}_i = x]$ equals the sum $\sum_t tP(\text{Y}_i = t|\text{X}_i = x)$, where $P(\text{Y}_i = t|\text{X}_i = x)$ is the conditional probability mass function for Y_i given $\text{X}_i = x$.

Expectation is a population concept. In practice, data usually come in the form of samples and rarely consist of an entire population. We therefore use samples to make inferences about the population. For example, the sample CEF is used to learn about the population CEF. This is necessary and important, but we postpone a discussion of the formal inference step taking us from sample to population until section 3.1.3. Our "population-first" approach to econometrics is motivated by the fact that we must define the objects of interest before we can use data to study them.[1]

[1] Examples of pedagogical writing using the "population-first" approach to econometrics include Chamberlain (1984), Goldberger (1991), and Manski (1991).

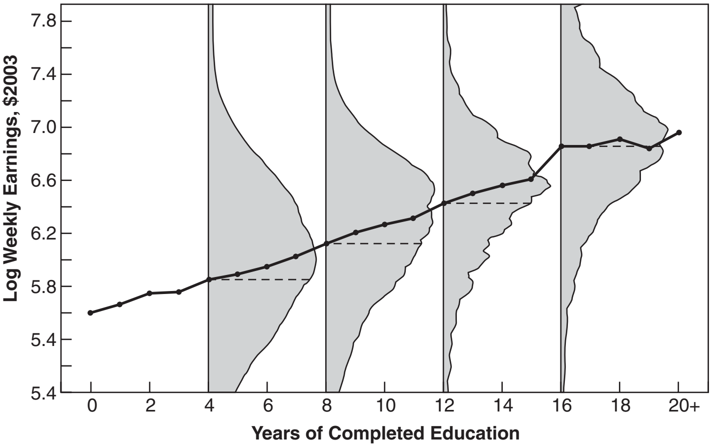


Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

Figure 3.1.1 plots the CEF of log weekly wages given schooling for a sample of middle-aged white men from the 1980 census. The distribution of earnings is also plotted for a few key values: 4, 8, 12, and 16 years of schooling. The CEF in the figure captures the fact that, notwithstanding the enormous variation individual circumstances, people with more schooling generally earn more. The average earnings gain associated with a year of schooling is typically about 10 percent.

An important complement to the CEF is the law of iterated expectations. This law says that an unconditional expectation can be written as the unconditional average of the CEF. In other words,

$$E[\text{Y}_i] = E\{E[\text{Y}_i|\text{X}_i]\}, \qquad (3.1.1)$$

where the outer expectation uses the distribution of X_i. Here is a proof of the law of iterated expectations for continuously distributed (X_i, Y_i) with joint density $f_{xy}(u, t)$, where $f_y(t|\text{X}_i = u)$ is the conditional distribution of Y_i given $\text{X}_i = u$ and $g_y(t)$

and $g_x(u)$ are the marginal densities:

$$\begin{aligned}
E\{E[\mathrm{Y}_i|\mathrm{X}_i]\} &= \int E[\mathrm{Y}_i|\mathrm{X}_i = u]g_x(u)du \\
&= \int \left[\int tf_y(t|\mathrm{X}_i = u)dt\right] g_x(u)du \\
&= \int \int tf_y(t|\mathrm{X}_i = u)g_x(u)dudt \\
&= \int t\left[\int f_y(t|\mathrm{X}_i = u)g_x(u)du\right] dt \\
&= \int t\left[\int f_{xy}(u,t)du\right] dt \\
&= \int tg_y(t)dt = E[\mathrm{Y}_i].
\end{aligned}$$

The integrals in this derivation run over the possible values of X_i and Y_i (indexed by u and t). We've laid out these steps because the CEF and its properties are central to the rest of this chapter.[2]

The power of the law of iterated expectations comes from the way it breaks a random variable into two pieces, the CEF and a residual with special properties.

Theorem 3.1.1 *The CEF Decomposition Property.*

$$\mathrm{Y}_i = E[\mathrm{Y}_i|\mathrm{X}_i] + \varepsilon_i,$$

where (i) ε_i is mean independent of X_i, that is, $E[\varepsilon_i|\mathrm{X}_i] = 0$, and therefore (ii) ε_i is uncorrelated with any function of X_i.

Proof. (i) $E[\varepsilon_i|\mathrm{X}_i] = E[\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]|\mathrm{X}_i] = E[\mathrm{Y}_i|\mathrm{X}_i] - E[\mathrm{Y}_i|\mathrm{X}_i] = 0$. (ii) Let $h(\mathrm{X}_i)$ be any function of X_i. By the law of iterated expectations, $E[h(\mathrm{X}_i)\varepsilon_i] = E\{h(\mathrm{X}_i)E[\varepsilon_i|\mathrm{X}_i]\}$, and by mean independence, $E[\varepsilon_i|\mathrm{X}_i] = 0$.

[2] A simple example illustrates how the law of iterated expectations works: Average earnings in a population of men and women is the average for men times the proportion male in the population plus the average for women times the proportion female in the population.

This theorem says that any random variable Y_i can be decomposed into a piece that is "explained by X_i"—that is, the CEF—and a piece left over that is orthogonal to (i.e., uncorrelated with) any function of X_i.

The CEF is a good summary of the relationship between Y_i and X_i, for a number of reasons. First, we are used to thinking of averages as providing a representative value for a random variable. More formally, the CEF is the best predictor of Y_i given X_i in the sense that it solves a minimum mean squared error (MMSE) prediction problem. This CEF prediction property is a consequence of the CEF decomposition property:

Theorem 3.1.2 *The CEF Prediction Property.*
Let $m(\mathrm{X}_i)$ be any function of X_i. The CEF solves

$$E[\mathrm{Y}_i|\mathrm{X}_i] = \underset{m(\mathrm{X}_i)}{\arg\min}\, E[(\mathrm{Y}_i - m(\mathrm{X}_i))^2],$$

so it is the MMSE predictor of Y_i given X_i.

Proof. Write

$$\begin{aligned}(\mathrm{Y}_i - m(\mathrm{X}_i))^2 &= ((\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i)))^2\\ &= (\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])^2 + 2(E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))\\ &\quad \times (\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))^2.\end{aligned}$$

The first term doesn't matter because it doesn't involve $m(\mathrm{X}_i)$. The second term can be written $h(\mathrm{X}_i)\varepsilon_i$, where $h(\mathrm{X}_i) \equiv 2(E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))$, and therefore has expectation zero by the CEF decomposition property. The last term is minimized at zero when $m(\mathrm{X}_i)$ is the CEF.

A final property of the CEF, closely related to both the decomposition and prediction properties, is the analysis of variance (ANOVA) theorem:

Theorem 3.1.3 *The ANOVA Theorem.*

$$V(\mathrm{Y}_i) = V(E[\mathrm{Y}_i|\mathrm{X}_i]) + E[V(\mathrm{Y}_i|\mathrm{X}_i)],$$

where $V(\cdot)$ denotes variance and $V(\mathrm{Y}_i|\mathrm{X}_i)$ is the conditional variance of Y_i given X_i.

Proof. The CEF decomposition property implies the variance of Y_i is the variance of the CEF plus the variance of the residual, $\varepsilon_i \equiv \text{Y}_i - E[\text{Y}_i|\text{X}_i]$, since ε_i and $E[\text{Y}_i|\text{X}_i]$ are uncorrelated. The variance of ε_i is

$$E[\varepsilon_i^2] = E[E[\varepsilon_i^2|\text{X}_i]] = E[V[\text{Y}_i|\text{X}_i]],$$

where $E[\varepsilon_i^2|\text{X}_i] = V[\text{Y}_i|\text{X}_i]$ because $\varepsilon_i \equiv \text{Y}_i - E[\text{Y}_i|\text{X}_i]$.

The two CEF properties and the ANOVA theorem may have a familiar ring. You might be used to seeing an ANOVA table in your regression output, for example. ANOVA is also important in research on inequality, where labor economists decompose changes in the income distribution into parts that can be accounted for by changes in worker characteristics and changes in what's left over after accounting for these factors (see, e.g., Autor, Katz, and Kearney, 2005). What may be unfamiliar is the fact that the CEF properties and ANOVA variance decomposition work in the population as well as in samples, and do not turn on the assumption of a linear CEF. In fact, the validity of linear regression as an empirical tool does not turn on linearity either.

3.1.2 Linear Regression and the CEF

So what's the regression you want to run? In our world, this question or one like it is heard almost every day. Regression estimates provide a valuable baseline for almost all empirical research because regression is tightly linked to the CEF, and the CEF provides a natural summary of empirical relationships. The link between regression functions—that is, the best-fitting line generated by minimizing expected squared errors—and the CEF can be explained in at least three ways. To lay out these explanations precisely, it helps to be precise about the regression function we have in mind. This section is concerned with the vector of population regression coefficients, defined as the solution to a population least squares problem. At this point we are not worried about causality. Rather, we let the $\textsc{k} \times 1$ regression coefficient vector β be defined by solving

$$\beta = \underset{b}{\arg\min}\; E[(\textsc{y}_i - \textsc{x}_i'b)^2]. \tag{3.1.2}$$

Using the first-order condition,

$$E[\textsc{x}_i(\textsc{y}_i - \textsc{x}_i'b)] = 0,$$

the solution can be written $\beta = E[\textsc{x}_i\textsc{x}_i']^{-1}E[\textsc{x}_i\textsc{y}_i]$. Note that by construction, $E[\textsc{x}_i(\textsc{y}_i - \textsc{x}_i'\beta)] = 0$. In other words, the population residual, which we define as $\textsc{y}_i - \textsc{x}_i'\beta = e_i$, is uncorrelated with the regressors, $\textsc{x}_i$. It bears emphasizing that this error term does not have a life of its own. It owes its existence and meaning to β. We return to this important point in the discussion of causal regression in section 3.2.

In the simple bivariate case where the regression vector includes only the single regressor, x_i, and a constant, the slope coefficient is $\beta_1 = \frac{Cov(\textsc{y}_i, x_i)}{V(x_i)}$, and the intercept is $\alpha = E[\textsc{y}_i] - \beta_1 E[\textsc{x}_i]$. In the multivariate case, with more than one nonconstant regressor, the slope coefficient for the kth regressor is given below:

REGRESSION ANATOMY

$$\beta_k = \frac{Cov(\textsc{y}_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}, \tag{3.1.3}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all the other covariates.

In other words, $E[\textsc{x}_i\textsc{x}_i']^{-1}E[\textsc{x}_i\textsc{y}_i]$ is the $\textsc{k} \times 1$ vector with kth element $\frac{Cov(\textsc{y}_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$. This important formula is said to describe the anatomy of a multivariate regression coefficient because it reveals much more than the matrix formula $\beta = E[\textsc{x}_i\textsc{x}_i']^{-1}E[\textsc{x}_i\textsc{y}_i]$. It shows us that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after partialing out all the other covariates.

To verify the regression anatomy formula, substitute

$$\text{Y}_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_{\text{K}} x_{\text{K}i} + e_i$$

in the numerator of (3.1.3). Since $\tilde{x}_{ki}$ is a linear combination of the regressors, it is uncorrelated with e_i. Also, since $\tilde{x}_{ki}$ is a residual from a regression on all the other covariates in the model, it must be uncorrelated with these covariates. Finally, for the same reason, the covariance of $\tilde{x}_{ki}$ with x_{ki} is just the variance of $\tilde{x}_{ki}$. We therefore have $Cov(\text{Y}_i, \tilde{x}_{ki}) = \beta_k V(\tilde{x}_{ki})$.[3]

The regression anatomy formula is probably familiar to you from a regression or statistics course, perhaps with one twist: the regression coefficients defined in this section are not estimators; rather, they are nonstochastic features of the joint distribution of dependent and independent variables. This joint distribution is what you would observe if you had a complete enumeration of the population of interest (or knew the stochastic process generating the data). You probably don't have such information. Still, it's good empirical practice to think about what population parameters mean before worrying about how to estimate them.

Below we discuss three reasons why the vector of population regression coefficients might be of interest. These reasons can be summarized by saying that you should be interested in regression parameters if you are interested in the CEF.

[3] The regression anatomy formula is usually attributed to Frisch and Waugh (1933). You can also do regression anatomy this way:

$$\beta_k = \frac{Cov(\tilde{\text{Y}}_{ki}, \tilde{x}_{ki})}{V(\tilde{x}_{ki})},$$

where $\tilde{\text{Y}}_{ki}$ is the residual from a regression of Y_i on every covariate except x_{ki}. This works because the fitted values removed from $\tilde{\text{Y}}_{ki}$ are uncorrelated with $\tilde{x}_{ki}$. Often it's useful to plot $\tilde{\text{Y}}_{ki}$ against $\tilde{x}_{ki}$; the slope of the least squares fit in this scatterplot is the multivariate β_k, even though the plot is two-dimensional. Note, however, that it's not enough to partial the other covariates out of Y_i only. That is,

$$\frac{Cov(\tilde{\text{Y}}_{ki}, x_{ki})}{V(x_{ki})} = \left[\frac{Cov(\tilde{\text{Y}}_{ki}, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}\right]\left[\frac{V(\tilde{x}_{ki})}{V(x_{ki})}\right] \neq \beta_k,$$

unless x_{ki} is uncorrelated with the other covariates.

Theorem 3.1.4 *The Linear CEF Theorem (Regression Justification I).*

Suppose the CEF is linear. Then the population regression function is it.

Proof. Suppose $E[\text{Y}_i|\text{X}_i] = \text{X}_i'\beta^*$ for a $\text{K} \times 1$ vector of coefficients, β^*. Recall that $E[\text{X}_i(\text{Y}_i - E[\text{Y}_i|\text{X}_i])] = 0$ by the CEF decomposition property. Substitute using $E[\text{Y}_i|\text{X}_i] = \text{X}_i'\beta^*$ to find that $\beta^* = E[\text{X}_i\text{X}_i']^{-1}E[\text{X}_i\text{Y}_i] = \beta$.

The linear CEF theorem raises the question of what makes a CEF linear. The classic scenario is joint normality, that is, the vector $(\text{Y}_i, \text{X}_i')'$ has a multivariate normal distribution. This is the scenario considered by Galton (1886), father of regression, who was interested in the intergenerational link between normally distributed traits such as height and intelligence. The normal case is clearly of limited empirical relevance since regressors and dependent variables are often discrete, while normal distributions are continuous. Another linearity scenario arises when regression models are saturated. As reviewed in section 3.1.4, a saturated regression model has a separate parameter for every possible combination of values that the set of regressors can take on. For example a saturated regression model with two dummy covariates includes both covariates (with coefficients known as the main effects) and their product (known as an interaction term). Such models are inherently linear, a point we also discuss in section 3.1.4.

The following two reasons for focusing on regression are relevant when the linear CEF theorem does not apply.

Theorem 3.1.5 *The Best Linear Predictor Theorem (Regression Justification II).*

The function $\text{X}_i'\beta$ *is the best linear predictor of* Y_i *given* X_i *in a MMSE sense.*

Proof. $\beta = E[\text{X}_i\text{X}_i']^{-1}E[\text{X}_i\text{Y}_i]$ solves the population least squares problem, (3.1.2).

In other words, just as the CEF, $E[\text{Y}_i|\text{X}_i]$, is the best (i.e., MMSE) predictor of Y_i given X_i in the class of *all* functions of

X_i, the population regression function is the best we can do in the class of *linear* functions.

Theorem 3.1.6 *The Regression CEF Theorem (Regression Justification III).*

The function $\mathrm{X}_i'\beta$ *provides the MMSE linear approximation to* $E[\mathrm{Y}_i|\mathrm{X}_i]$, *that is,*

$$\beta = \arg\min_b E\{(E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2\}. \tag{3.1.4}$$

Proof. Start by observing that β solves (3.1.2). Write

$$\begin{aligned}(\mathrm{Y}_i - \mathrm{X}_i'b)^2 &= \{(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)\}^2 \\ &= (\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])^2 + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2 \\ &\quad + 2(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])(E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b).\end{aligned}$$

The first term doesn't involve b and the last term has expectation zero by the CEF decomposition property (ii). The CEF approximation problem, (3.1.4), is therefore the same as the population least squares problem, (3.1.2).

These two theorems give us two more ways to view regression. Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable. On the other hand, if we prefer to think about approximating $E[\mathrm{Y}_i|\mathrm{X}_i]$, as opposed to predicting Y_i, the regression CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.

The regression CEF theorem is our favorite way to motivate regression. The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships without necessarily trying to pin them down exactly. The linear CEF theorem is for special cases only. The best linear predictor theorem is satisfyingly general, but seems to encourage an overly clinical view of empirical research. We're not really interested in predicting *individual* Y_i; it's the *distribution* of Y_i that we care about.

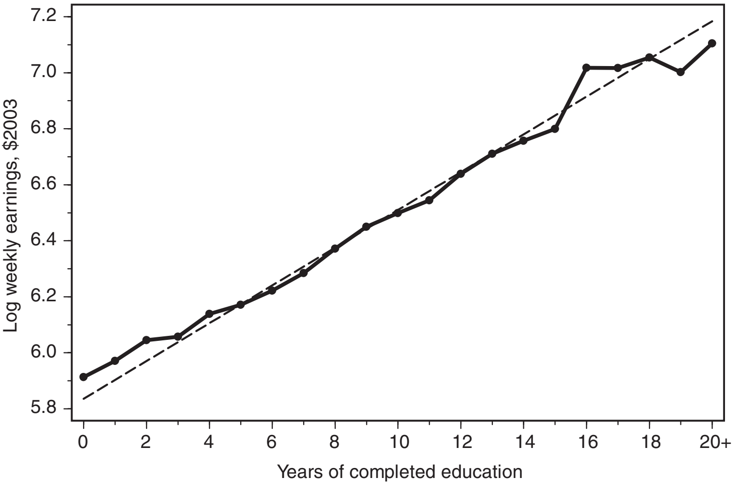


Figure 3.1.2 Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

Figure 3.1.2 illustrates the CEF approximation property for the same schooling CEF plotted in figure 3.1.1. The regression line fits the somewhat bumpy and nonlinear CEF as if we were estimating a model for $E[\text{Y}_i|\text{X}_i]$ instead of a model for Y_i. In fact, that is exactly what's going on. An implication of the regression CEF theorem is that regression coefficients can be obtained by using $E[\text{Y}_i|\text{X}_i]$ as a dependent variable instead of Y_i itself. To see this, suppose that X_i is a discrete random variable with probability mass function $g_x(u)$. Then

$$E\{(E[\text{Y}_i|\text{X}_i] - \text{X}_i'b)^2\} = \sum_u (E[\text{Y}_i|\text{X}_i = u] - u'b)^2 g_x(u).$$

This means that β can be constructed from the weighted least squares (WLS) regression of $E[\text{Y}_i|\text{X}_i = u]$ on u, where u runs over the values taken on by X_i. The weights are given by the distribution of X_i, that is, $g_x(u)$. An even simpler way to see this is to iterate expectations in the formula for β:

$$\beta = E[\text{X}_i\text{X}_i']^{-1}E[\text{X}_i\text{Y}_i] = E[\text{X}_i\text{X}_i']^{-1}E[\text{X}_iE(\text{Y}_i|\text{X}_i)]. \quad (3.1.5)$$

The CEF or grouped data version of the regression formula is of practical use when working on a project that precludes the analysis of microdata. For example, Angrist (1998) used grouped data to study the effect of voluntary military service on earnings later in life. One of the estimation strategies used in this project regresses civilian earnings on a dummy for veteran status, along with personal characteristics and the variables used by the military to screen soldiers. The earnings data come from the U.S. Social Security system, but Social Security earnings records cannot be released to the public. Instead of individual earnings, Angrist worked with average earnings conditional on race, sex, test scores, education, and veteran status.

To illustrate the grouped data approach to regression, we estimated the schooling coefficient in a wage equation using 21 conditional means, the sample CEF of earnings given schooling. As the Stata output reproduced in Figure 3.1.3 shows, a grouped data regression, weighted by the number of individuals at each schooling level in the sample, produces coefficients identical to those generated using the underlying microdata sample with hundreds of thousands of observations. Note, however, that the standard errors from the grouped regression do not measure the asymptotic sampling variance of the slope estimate in repeated micro-data samples; for that you need an estimate of the variance of $\text{Y}_i - \text{X}_i'\beta$. This variance depends on the microdata, in particular the second moments of $W_i \equiv [\text{Y}_i \; \text{X}_i']'$, a point we elaborate on in the next section.

3.1.3 Asymptotic OLS Inference

In practice, we don't usually know what the CEF or the population regression vector is. We therefore draw statistical inferences about these quantities using samples. Statistical inference is what much of traditional econometrics is about. Although this material is covered in any econometrics text, we don't want to skip the inference step completely. A review of basic asymptotic theory allows us to highlight the important fact that the process of statistical inference is distinct from the

A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS       Number of obs =  409435
-------------+------------------------------    F(  1,409433) =49118.25
       Model | 22631.4793      1  22631.4793    Prob > F      =  0.0000
    Residual |  188648.31 409433  .460755019    R-squared     =  0.1071
-------------+------------------------------    Adj R-squared =  0.1071
       Total | 211279.789 409434   .51602893    Root MSE      =  .67879
-------------+-----------------------------------------------------------
             |              Robust                  Old Fashioned
    earnings |      Coef.   Std. Err.      t           Std. Err.       t
-------------+-----------------------------------------------------------
      school |   .0674387   .0003447   195.63          .0003043   221.63
      const. |   5.835761   .0045507  1282.39          .0040043  1457.38
-------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS       Number of obs =      21
-------------+------------------------------    F(  1,    19) =  540.31
       Model |  1.16077332     1  1.16077332    Prob > F      =  0.0000
    Residual |  .040818796    19  .002148358    R-squared     =  0.9660
-------------+------------------------------    Adj R-squared =  0.9642
       Total |  1.20159212    20  .060079606    Root MSE      =  .04635
-------------+-----------------------------------------------------------
     average |              Robust                  Old Fashioned
   _earnings |      Coef.   Std. Err.      t           Std. Err.       t
-------------+-----------------------------------------------------------
      school |   .0674387   .0040352    16.71         .0029013     23.24
      const. |   5.835761   .0399452   146.09         .0381792    152.85
-------------------------------------------------------------------------
```

Figure 3.1.3 Microdata and grouped data estimates of the returns to schooling, from Stata regression output. *Source*: 1980 Census—IPUMS, 5 percent sample. The sample includes white men, age 40–49. Robust standard errors are heteroskedasticity consistent. Panel A uses individual-level microdata. Panel B uses earnings averaged by years of schooling.

question of how a particular set of regression estimates should be interpreted. Whatever a regression coefficient may mean, it has a sampling distribution that is easy to describe and use for statistical inference.[4]

[4] The discussion of asymptotic OLS inference in this section is largely a condensation of material in Chamberlain (1984). Important pitfalls and problems with asymptotic theory are covered in the last chapter.

We are interested in the distribution of the sample analog of

$$\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$$

in repeated samples. Suppose the vector $W_i \equiv [\mathrm{Y}_i\ \mathrm{X}_i']'$ is independently and identically distributed in a sample of size N. A natural estimator of the first population moment, $E[W_i]$, is the sum, $\frac{1}{N}\sum_{i=1}^{N} W_i$. By the law of large numbers, this vector of sample moments gets arbitrarily close to the corresponding vector of population moments as the sample size grows. We might similarly consider higher-order moments of the elements of W_i, for example the matrix of second moments, $E[W_i W_i']$, with sample analog $\frac{1}{N}\sum_{i=1}^{N} W_i W_i'$. Following this principle, the method of moments estimator of β replaces each expectation by a sum. This logic leads to the ordinary least squares (OLS) estimator

$$\hat{\beta} = \left[\sum_i \mathrm{X}_i\mathrm{X}_i'\right]^{-1}\sum_i \mathrm{X}_i\mathrm{Y}_i.$$

Although we derived $\hat{\beta}$ as a method of moments estimator, it is called the OLS estimator of β because it solves the sample analog of the least squares problem described at the beginning of section 3.1.2.[5]

The asymptotic sampling distribution of $\hat{\beta}$ depends solely on the definition of the estimand (i.e., the nature of the thing we're trying to estimate, β) and the assumption that the data constitute a random sample. Before deriving this distribution, it helps to summarize the general asymptotic distribution theory that covers our needs. This basic theory can be stated mostly in words. For the purposes of these statements, we assume the reader is familiar with the core terms and concepts of statistical theory—moments, mathematical expectation, probability

[5] Econometricians like to use matrices because the notation is so compact. Sometimes (not very often) we do too. Suppose X is the matrix whose rows are given by X_i' and y is the vector with elements Y_i, for $i = 1, \ldots, N$. The sample moment matrix $\frac{1}{N}\sum X_iX_i'$ is $X'X/N$ and the sample moment vector $\frac{1}{N}\sum X_iy_i$ is $X'y/N$. Then we can write $\hat{\beta} = (X'X)^{-1}X'y$, a widely used matrix formula.

limits, and asymptotic distributions. For definitions of these terms and a formal mathematical statement of the theoretical propositions given below, see Knight (2000).

THE LAW OF LARGE NUMBERS Sample moments converge in probability to the corresponding population moments. In other words, the probability that the sample mean is close to the population mean can be made as high as you like by taking a large enough sample.

THE CENTRAL LIMIT THEOREM Sample moments are asymptotically normally distributed (after subtracting the corresponding population moment and multiplying by the square root of the sample size). The asymptotic covariance matrix is given by the variance of the underlying random variable. In other words, in large enough samples, appropriately normalized sample moments are approximately normally distributed.

SLUTSKY'S THEOREM

1. Consider the sum of two random variables, one of which converges in distribution (in other words, has an asymptotic distribution) and the other converges in probability to a constant: the asymptotic distribution of this sum is unaffected by replacing the one that converges to a constant by this constant. Formally, let a_N be a statistic with an asymptotic distribution and let b_N be a statistic with probability limit b. Then $a_N + b_N$ and $a_N + b$ have the same asymptotic distribution.
2. Consider the product of two random variables, one of which converges in distribution and the other converges in probability to a constant: the asymptotic distribution of this product is unaffected by replacing the one that converges to a constant by this constant. Formally, let a_N be a statistic with an asymptotic distribution and let b_N be a statistic with probability limit b. Then $a_N b_N$ and $a_N b$ have the same asymptotic distribution.

THE CONTINUOUS MAPPING THEOREM Probability limits pass through continuous functions. For example, the probability limit of any continuous function of a sample moment is the function evaluated at the corresponding population moment. Formally, the probability limit of $h(b_N)$ is $h(b)$, where *plim* $b_N = b$ and $h(\cdot)$ is continuous at b.

THE DELTA METHOD Consider a vector-valued random variable that is asymptotically normally distributed. Continuously differentiable scalar functions of this random variable are also asymptotically normally distributed, with covariance matrix given by a quadratic form with the covariance matrix of the random variable on the inside and the gradient of the function evaluated at the probability limit of the random variable on the outside.[6] Formally, the asymptotic distribution of $h(b_N)$ is normal with covariance matrix $\nabla h(b)'\Omega\nabla h(b)$, where *plim* $b_N = b$, $h(\cdot)$ is continuously differentiable at b with gradient $\nabla h(b)$, and b_N has asymptotic covariance matrix Ω.[7]

We can use these results to derive the asymptotic distribution of $\hat{\beta}$ in two ways. A conceptually straightforward but somewhat inelegant approach is to use the delta method: $\hat{\beta}$ is a function of sample moments, and is therefore asymptotically normally distributed. It remains only to find the covariance matrix of the asymptotic distribution from the gradient of this function. (Note that consistency of $\hat{\beta}$ comes immediately from the continuous mapping theorem).[8] An easier and more instructive derivation uses the Slutsky and central limit theorems. Note first that we can write

$$\text{Y}_i = \text{X}_i'\beta + [\text{Y}_i - \text{X}_i'\beta] \equiv \text{X}_i'\beta + e_i, \qquad (3.1.6)$$

where the residual e_i is defined as the difference between the dependent variable and the population regression function, as

[6] A quadratic form is a matrix-weighted sum of squares. Suppose v is an $N \times 1$ vector and M is an $N \times N$ matrix. A quadratic form in v is $v'Mv$. If M is an $N \times N$ diagonal matrix with diagonal elements m_i, then $v'Mv = \sum_i m_i v_i^2$.

[7] For a derivation of the delta method formula using the Slutsky and continuous mapping theorems, see Knight (2000, pp. 120–121). We say "the asymptotic distribution of $h(b_N)$," but we really mean the asymptotic distribution of $\sqrt{N}(h(b_N) - h(b))$.

[8] An estimator is said to be *consistent* when it converges in probability to the target parameter.

before. In other words, $E[\mathrm{X}_i e_i] = 0$ is a consequence of $\beta = E[\mathrm{X}_i\ \mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$ and $e_i = \mathrm{Y}_i - \mathrm{X}_i'\beta$, and not an assumption about an underlying economic relation.[9]

Substituting the identity (3.1.6) for Y_i in the formula for $\hat{\beta}$, we have

$$\hat{\beta} = \beta + \left[\sum \mathrm{X}_i\mathrm{X}_i'\right]^{-1} \sum \mathrm{X}_i e_i.$$

The asymptotic distribution of $\hat{\beta}$ is the asymptotic distribution of $\sqrt{N}(\hat{\beta} - \beta) = N[\sum \mathrm{X}_i\mathrm{X}_i']^{-1}\frac{1}{\sqrt{N}}\sum \mathrm{X}_i e_i$. By the Slutsky theorem, this has the same asymptotic distribution as $E[\mathrm{X}_i\mathrm{X}_i']^{-1}\frac{1}{\sqrt{N}}\sum \mathrm{X}_i e_i$. Since $E[\mathrm{X}_i e_i] = 0$, $\frac{1}{\sqrt{N}}\sum \mathrm{X}_i e_i$ is a root-N normalized and centered sample moment. By the central limit theorem, this is asymptotically normally distributed with mean zero and covariance matrix $E[\mathrm{X}_i\mathrm{X}_i'e_i^2]$, since this matrix of fourth moments is the covariance matrix of $\mathrm{X}_i e_i$. Therefore, $\hat{\beta}$ has an asymptotic normal distribution with probability limit β and covariance matrix

$$E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i'e_i^2]E[\mathrm{X}_i\mathrm{X}_i']^{-1}. \tag{3.1.7}$$

The theoretical standard errors used to construct t-statistics are the square roots of the diagonal elements of (3.1.7). In practice these standard errors are estimated by substituting sums for expectations and using the estimated residuals, $\hat{e}_i = \mathrm{Y}_i - \mathrm{X}_i'\hat{\beta}$ to form the empirical fourth moment matrix, $\sum[\mathrm{X}_i\mathrm{X}_i\hat{e}_i^2]/N$.

Asymptotic standard errors computed in this way are known as heteroskedasticity-consistent standard errors, White (1980a) standard errors, or Eicker-White standard errors, in recognition of Eicker's (1967) derivation. They are also known as "robust" standard errors (e.g., in Stata). These standard errors are said to be robust because, in large enough samples, they provide accurate hypothesis tests and confidence intervals given minimal assumptions about the data and model. In particular, our derivation of the limiting distribution makes

[9] Residuals defined in this way are not necessarily mean independent of X_i; for mean independence, we need a linear CEF.

no assumptions other than those needed to ensure that basic statistical results like the central limit theorem go through. Robust standard errors are not, however, the standard errors that you get by default from packaged software. Default standard errors are derived under a homoskedasticity assumption, specifically, that $E[e_i^2|\mathrm{X}_i] = \sigma^2$, a constant. Given this assumption, we have

$$E[\mathrm{X}_i\mathrm{X}_i'e_i^2] = E(\mathrm{X}_i\mathrm{X}_i'E[e_i^2|\mathrm{X}_i]) = \sigma^2 E[\mathrm{X}_i\mathrm{X}_i'],$$

by iterating expectations. The asymptotic covariance matrix of $\hat{\beta}$ then simplifies to

$$\begin{aligned} &E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i'e_i^2]E[\mathrm{X}_i\mathrm{X}_i']^{-1} \\ &\quad = E[\mathrm{X}_i\mathrm{X}_i']^{-1}\sigma^2 E[\mathrm{X}_i\mathrm{X}_i']E[\mathrm{X}_i\mathrm{X}_i]^{-1} \\ &\quad = \sigma^2 E[\mathrm{X}_i\mathrm{X}_i']^{-1}. \end{aligned} \tag{3.1.8}$$

The diagonal elements of (3.1.8) are what SAS or Stata report unless you request otherwise.

Our view of regression as an approximation to the CEF makes heteroskedasticity seem natural. If the CEF is nonlinear and you use a linear model to approximate it, then the quality of fit between the regression line and the CEF will vary with X_i. Hence, the residuals will be larger, on average, at values of X_i where the fit is poorer. Even if you are prepared to assume that the conditional variance of $\textsc{y}_i$ given X_i is constant, the fact that the CEF is nonlinear means that $E[(\textsc{y}_i - \mathrm{X}_i'\beta)^2|\mathrm{X}_i]$ will vary with X_i. To see this, note that

$$\begin{aligned} &E[(\textsc{y}_i - \mathrm{X}_i'\beta)^2|\mathrm{X}_i] \\ &\quad = E\{[(\textsc{y}_i - E[\textsc{y}_i|\mathrm{X}_i]) + (E[\textsc{y}_i|\mathrm{X}_i] - \mathrm{X}_i'\beta)]^2|\mathrm{X}_i\} \\ &\quad = V[\textsc{y}_i|\mathrm{X}_i] + (E[\textsc{y}_i|\mathrm{X}_i] - \mathrm{X}_i'\beta)^2. \end{aligned} \tag{3.1.9}$$

Therefore, even if $V[\textsc{y}_i|\mathrm{X}_i]$ is constant, the residual variance increases with the square of the gap between the regression line and the CEF, a fact noted in White (1980b).[10]

[10] The cross-product term resulting from an expansion of the squared term in the middle of (3.1.9) is zero because $\textsc{y}_i - E[\textsc{y}_i|\mathrm{X}_i]$ is mean independent of X_i.

In the same spirit, it's also worth noting that while a linear CEF makes homoskedasticity possible, this is not a sufficient condition for homoskedasticity. Our favorite example in this context is the linear probability model (LPM). A linear probability model is any regression where the dependent variable is zero-one, that is, a dummy variable such as an indicator for labor force participation. Suppose the regression model is saturated, so the CEF given regressors is linear. Because the CEF is linear, the residual variance is also the conditional variance, $V[\mathrm{Y}_i|\mathrm{X}_i]$. But the dependent variable is a Bernoulli trial with conditional variance $P[\mathrm{Y}_i = 1|\mathrm{X}_i](1 - P[\mathrm{Y}_i = 1|\mathrm{X}_i])$. We conclude that LPM residuals are necessarily heteroskedastic unless the only regressor is a constant.

These points of principle notwithstanding, as an empirical matter, heteroskedasticity may matter little. In the microdata schooling regression depicted in figure 3.1.3, the robust standard error is .0003447, while the old-fashioned standard error is .0003043, not much smaller. The standard errors from the grouped data regression, which are necessarily heteroskedastic if group sizes differ, change somewhat more; compare the .004 robust standard to the .0029 conventional standard error. Based on our experience, these differences are typical. If heteroskedasticity matters a lot, say, more than a 30 percent increase or any marked decrease in standard errors, you should worry about possible programming errors or other problems. For example, robust standard errors below conventional may be a sign of finite-sample bias in the robust calculation.

Finally, a brief note on the textbook approach to inference that you might have seen elsewhere. Traditional econometric inference begins with stronger assumptions than those we have invoked in this section. The traditional set-up, sometimes called a classical normal regression model, postulates: fixed (non-stochastic) regressors, a linear CEF, normally distributed errors, and homoskedasticity (see, e.g., Goldberger, 1991). These stronger assumptions give us two things: (1) unbiasedness of the OLS estimator, (2) a formula for the sampling variance of the OLS estimator that is valid in small as well as large samples. Unbiasedness of the OLS estimators means that $E[\hat{\beta}] = \beta$, a property that holds in a sample of any size and is stronger than consistency, which means only that we can expect $\hat{\beta}$ to be close to β in large samples. It's easy to see when and why we get unbiasedness. In general,

$$E[\hat{\beta}] = \beta + E\left\{\left[\sum X_i X_i'\right]^{-1} \sum X_i e_i\right\}.$$

If the regressors are nonrandom (fixed in repeated samples) the expectation passes through and we have unbiasedness because $E[e_i] = 0$. Otherwise, with random regressors, we can iterate expectations and get unbiasedness if $E[e_i|X_i] = 0$. This is true when the CEF is linear, but not in our more general "agnostic regression" framework.

The variance formula obtained under classical assumptions is the same as the large-sample formula under homoskedasticity but—provided the strong classical assumptions are valid—this formula holds in a sample of any size. We've chosen to start with the asymptotic approach to inference because modern empirical work typically leans heavily on the large-sample theory that lies behind robust variance formulas. The payoff is valid inference under weak assumptions, in particular, a framework that makes sense for our less-than-literal approach to regression models. On the other hand, the large-sample approach is not without its dangers, a point we return to in the discussion of inference in chapter 8 and in the discussion of instrumental variables in chapter 4.

3.1.4 Saturated Models, Main Effects, and Other Regression Talk

We often discuss regression models using terms like *saturated* and *main effects*. These terms originate in an experimentalist tradition that uses regression to model the effects of discrete treatment-type variables. This language is now used more widely in many fields, however, including applied econometrics. For readers unfamiliar with these terms, this section provides a brief review.

Saturated regression models are regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the explanatory variables. For example, when working with a single explanatory variable indicating whether a worker is a college graduate, the model is saturated by including a single dummy for college graduates and a constant. We can also saturate when the regressor takes on many values. Suppose, for example, that $\text{S}_i = 0, 1, 2, \ldots, \tau$. A saturated regression model for S_i is

$$\text{Y}_i = \alpha + \beta_1 d_{1i} + \beta_2 d_{2i} + \cdots + \beta_\tau d_{\tau i} + \varepsilon_i,$$

where $d_{ji} = 1[\text{S}_i = j]$ is a dummy variable indicating schooling level j, and β_j is said to be the jth-level schooling *effect*.[11] Note that

$$\beta_j = E[\text{Y}_i|\text{S}_i = j] - E[\text{Y}_i|\text{S}_i = 0],$$

while $\alpha = E[\text{Y}_i|\text{S}_i = 0]$. In practice, you can pick any value of S_i for the reference group; a regression model is saturated as long as it has one parameter for every possible j in $E[\text{Y}_i|\text{S}_i = j]$. Saturated regression models fit the CEF perfectly because the CEF is a linear function of the dummy regressors used to saturate. This is an important special case of the linear CEF theorem.

If there are two explanatory variables—say, one dummy indicating college graduates and one dummy indicating sex—the model is saturated by including these two dummies, their product, and a constant. The coefficients on the dummies are known as main effects, while the product is called an *interaction term*. This is not the only saturated parameterization; any set of indicators (dummies) that can be used to identify each value taken on by all covariates produces a saturated model. For example, an alternative saturated model includes dummies for male college graduates, male nongraduates, female college graduates, and female nongraduates, but no intercept.

Here's some notation to make this more concrete. Let x_{1i} indicate college graduates and x_{2i} indicate women. The CEF

[11] We use the notation $1[\text{S}_i = j]$ to denote the indicator function, in this case a function that creates a dummy variable switched on when $\text{S}_i = j$.

given x_{1i} and x_{2i} takes on four values:

$$\begin{aligned}
&E[\text{Y}_i|x_{1i}=0,x_{2i}=0],\\
&E[\text{Y}_i|x_{1i}=1,x_{2i}=0],\\
&E[\text{Y}_i|x_{1i}=0,x_{2i}=1],\\
&E[\text{Y}_i|x_{1i}=1,x_{2i}=1].
\end{aligned}$$

We can label these using the following scheme:

$$\begin{aligned}
&E[\text{Y}_i|x_{1i}=0,x_{2i}=0]=\alpha\\
&E[\text{Y}_i|x_{1i}=1,x_{2i}=0]=\alpha+\beta_1\\
&E[\text{Y}_i|x_{1i}=0,x_{2i}=1]=\alpha+\gamma\\
&E[\text{Y}_i|x_{1i}=1,x_{2i}=1]=\alpha+\beta_1+\gamma+\delta_1.
\end{aligned}$$

Since there are four Greek letters and the CEF takes on four values, this parameterization does not restrict the CEF. It can be written in terms of Greek letters as

$$E[\text{Y}_i|x_{1i},x_{2i}]=\alpha+\beta_1 x_{1i}+\gamma x_{2i}+\delta_1(x_{1i}x_{2i}),$$

a parameterization with two main effects and one interaction term.[12] The saturated regression equation becomes

$$\text{Y}_i=\alpha+\beta_1 x_{1i}+\gamma x_{2i}+\delta_1(x_{1i}x_{2i})+\varepsilon_i.$$

We can combine the multivalued schooling variable with sex to produce a saturated model that has τ main effects for schooling, one main effect for sex, and τ sex-schooling interactions:

$$\text{Y}_i=\alpha+\sum_{j=1}^{\tau}\beta_j d_{ji}+\gamma x_{2i}+\sum_{j=1}^{\tau}\delta_j(d_{ji}x_{2i})+\varepsilon_i. \qquad (3.1.10)$$

The coefficients on the interaction terms, δ_j, tell us how each of the schooling effects differ by sex. The CEF in this case

[12] With a third dummy variable in the model, say x_{3i}, a saturated model includes three main effects, three second-order interaction terms $\{x_{1i}x_{2i}, x_{1i}x_{3i}, x_{2i}x_{3i}\}$, and one third-order term, $x_{1i}x_{2i}x_{3i}$.

takes on $2(\tau + 1)$ values, while the regression has this many parameters.

Note that there is a hierarchy of increasingly restrictive modeling strategies with saturated models at the top. It's natural to start with a saturated model because this fits the CEF. On the other hand, saturated models generate a lot of interaction terms, many of which may be uninteresting or estimated imprecisely. You might therefore sensibly choose to omit some or all of these terms. Equation (3.1.10) without interaction terms approximates the CEF using a purely additive model for schooling and sex. This is a good approximation if the returns to college are similar for men and women. In any case, schooling coefficients in the additive specification give a (weighted) average return across both sexes, as discussed in section 3.3.1. On the other hand, it would be strange to estimate a model that included interaction terms but omitted the corresponding main effects. In the case of schooling, this is something like

$$\mathrm{Y}_i = \alpha + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji} x_{2i}) + \varepsilon_i. \tag{3.1.11}$$

This model allows schooling to shift wages only for women, something very far from the truth. Consequently, the results of estimating (3.1.11) are likely to be hard to interpret.

Finally, it's important to recognize that a saturated model fits the CEF perfectly regardless of the distribution of Y_i. For example, this is true for linear probability models and other limited dependent variable models (e.g., non-negative Y_i), a point we return to at the end of this chapter.

3.2 Regression and Causality

Section 3.1.2 shows how regression gives the best (MMSE) linear approximation to the CEF. This understanding, however, does not help us with the deeper question of when regression has a causal interpretation. When can we think of a regression coefficient as approximating the causal effect that might be revealed in an experiment?

3.2.1 The Conditional Independence Assumption

A regression is causal when the CEF it approximates is causal. This doesn't answer the question, of course. It just passes the buck up one level, since, as we've seen, a regression inherits its legitimacy from a CEF. Causality means different things to different people, but researchers working in many disciplines have found it useful to think of causal relationships in terms of the potential outcomes notation used in chapter 2 to describe what would happen to a given individual in a hypothetical comparison of alternative hospitalization scenarios. Differences in these potential outcomes were said to be the causal effect of hospitalization. The CEF is causal when it describes differences in average potential outcomes for a fixed reference population.

It's easiest to expand on the somewhat murky notion of a causal CEF in the context of a particular question, so let's stick with the schooling example. The causal connection between schooling and earnings can be defined as the functional relationship that describes what a given individual would earn if he or she obtained different levels of education. In particular, we might think of schooling decisions as being made in a series of episodes where the decision maker can realistically go one way or another, even if certain choices are more likely than others. For example, in the middle of junior year, restless and unhappy, Angrist glumly considered his options: dropping out of high school and hopefully getting a job, staying in school but taking easy classes that would lead to a quick and dirty high school diploma, or plowing on in an academic track that would lead to college. Although the consequences of such choices are usually unknown in advance, the idea of alternative paths leading to alternative outcomes for a given individual seems uncontroversial. Philosophers have argued over whether this personal notion of potential outcomes is precise enough to be scientifically useful, but individual decision makers seem to have no trouble thinking about their lives and choices in this manner (as in Robert Frost's celebrated "The Road Not Taken": the traveler-narrator sees himself looking back on a moment of choice. He believes that the decision to follow the road less traveled "has made all the difference," though he also recognizes that counterfactual outcomes are unknowable).

In empirical work, the causal relationship between schooling and earnings tells us what people would earn, on average, if we could either change their schooling in a perfectly controlled environment or change their schooling randomly so that those with different levels of schooling would be otherwise comparable. As we discussed in chapter 2, experiments ensure that the causal variable of interest is independent of potential outcomes so that the groups being compared are truly comparable. Here, we would like to generalize this notion to causal variables that take on more than two values, and to more complicated situations where we must hold a variety of control variables fixed for causal inferences to be valid. This leads to the *conditional independence assumption* (CIA), a core assumption that provides the (sometimes implicit) justification for the causal interpretation of regression estimates. This assumption is also called selection on observables because the covariates to be held fixed are assumed to be known and observed (e.g., in Goldberger, 1972; Barnow, Cain, and Goldberger, 1981). The big question, therefore, is what these control variables are, or should be. We'll say more about that shortly. For now, we just do the econometric thing and call the covariates X_i. As far as the schooling problem goes, it seems natural to imagine that X_i is a vector that includes measures of ability and family background.

For starters, think of schooling as a binary decision, such as whether Angrist goes to college. Denote this by a dummy variable, C_i. The causal relationship between college attendance and a future outcome such as earnings can be described using the same potential outcomes notation we used to describe experiments in chapter 2. To address this question, we imagine two potential earnings variables:

$$\textit{Potential outcome} = \begin{cases} Y_{1i} & \text{if } C_i = 1 \\ Y_{0i} & \text{if } C_i = 0 \end{cases}.$$

In this case, Y_{0i} is *i*'s earnings without college, while Y_{1i} is *i*'s earnings if he goes. We would like to know the difference between Y_{1i} and Y_{0i}, which is the causal effect of college attendance on individual i. This is what we would measure if we could go back in time and nudge i onto the road not taken. The observed outcome, Y_i, can be written in terms of potential outcomes as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})C_i.$$

We get to see one of Y_{1i} or Y_{0i}, but never both. We therefore hope to measure the average of $Y_{1i} - Y_{0i}$, or the average for some group, such as those who went to college. This is $E[Y_{1i} - Y_{0i}|C_i = 1]$.

In general, comparisons of those who do and don't go to college are likely to be a poor measure of the causal effect of college attendance. Following the logic in chapter 2, we have

$$\underbrace{E[Y_i|C_i = 1] - E[Y_i|C_i = 0]}_{\text{Observed difference in earnings}} = \underbrace{E[Y_{1i} - Y_{0i}|C_i = 1]}_{\text{Average treatment effect on the treated}} + \underbrace{E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]}_{\text{Selection bias}}. \tag{3.2.1}$$

It seems likely that those who go to college would have earned more anyway. If so, selection bias is positive and the naive comparison, $E[Y_i|C_i = 1] - E[Y_i|C_i = 0]$, exaggerates the benefits of college attendance.

The CIA asserts that conditional on observed characteristics, X_i, selection bias disappears. Formally, this means

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp C_i|X_i, \tag{3.2.2}$$

where the symbol "$\perp\!\!\!\perp$" denotes the independence relation and random variables to the right of the vertical bar are the conditioning set. Given the CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation. In other words,

$$E[Y_i|X_i, C_i = 1] - E[Y_i|X_i, C_i = 0] = E[Y_{1i} - Y_{0i}|X_i].$$

Now, we'd like to expand the conditional independence assumption to causal relations that involve variables that can take on more than two values, such as years of schooling, S_i.

The causal relationship between schooling and earnings is likely to be different for each person. We therefore use the individual-specific functional notation,

$$\textsc{y}_{si} \equiv f_i(s),$$

to denote the potential earnings that person i would receive after obtaining s years of education. If s takes on only two values, 12 and 16, then we are back to the college/no college example:

$$\textsc{y}_{0i} = f_i(12); \textsc{y}_{1i} = f_i(16).$$

More generally, the function $f_i(s)$ tells us what i would earn for *any* value of schooling, s. In other words, $f_i(s)$ answers causal "what if" questions. In the context of theoretical models of the relationship between human capital and earnings, the form of $f_i(s)$ may be determined by aspects of individual behavior, by market forces, or both.

The CIA in this more general setup becomes

$$\textsc{y}_{si} \perp\!\!\!\perp \textsc{s}_i | \text{X}_i, \text{ for all } s. \tag{CIA}$$

In many randomized experiments, the CIA crops up because $\textsc{s}_i$ *is* randomly assigned conditional on X_i (in the Tennessee STAR experiment, for example, small classes were randomly assigned within schools). In an observational study, the CIA means that $\textsc{s}_i$ can be said to be "as good as randomly assigned," conditional on X_i.

Conditional on X_i, the *average causal effect* of a one-year increase in schooling is $E[f_i(s) - f_i(s-1)|\text{X}_i]$, while the average causal effect of a four-year increase in schooling is $E[f_i(s) - E[f_i(s-4)]|\text{X}_i]$. The data reveal only $\textsc{y}_i = f_i(\textsc{s}_i)$, that is, $f_i(s)$ for $s = \textsc{s}_i$. But given the CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation. In other words,

$$\begin{aligned} &E[\textsc{y}_i|\text{X}_i, \textsc{s}_i = s] - E[\textsc{y}_i|\text{X}_i, \textsc{s}_i = s-1] \\ &\quad = E[f_i(s) - f_i(s-1)|\text{X}_i] \end{aligned}$$

for any value of s. For example, we can compare the earnings of those with 12 and 11 years of schooling to learn about the average causal effect of high school graduation:

$$
\begin{aligned}
&E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 12] - E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 11] \\
&\quad = E[f_i(12)|\mathrm{X}_i, \mathrm{s}_i = 12] - E[f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 11].
\end{aligned}
$$

This comparison has a causal interpretation because, given the CIA,

$$
\begin{aligned}
&E[f_i(12)|\mathrm{X}_i, \mathrm{s}_i = 12] - E[f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 11] \\
&\quad = E[f_i(12) - f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 12].
\end{aligned}
$$

Here, selection bias comes from differences in the potential dropout earnings of high school graduates and nongraduates. Given the CIA, however, high school graduation is independent of potential earnings conditional on X_i, so the selection bias vanishes. Note also that in this case, the causal effect of graduating from high school on high school graduates is equal to the average high school graduation effect at X_i:

$$
E[f_i(12) - f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 12] = E[f_i(12) - f_i(11)|\mathrm{X}_i].
$$

This is important, but less important than the elimination of selection bias.

So far, we have constructed separate causal effects for each value taken on by the conditioning variables. This leads to as many causal effects as there are values of X_i, an embarrassment of riches. Empiricists almost always find it useful to boil a set of estimates down to a single summary measure, such as the unconditional or overall average causal effect. By the law of iterated expectations, the unconditional average causal effect of high school graduation is

$$
\begin{aligned}
&E\{E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 12] - E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 11]\} && (3.2.3) \\
&\quad = E\{E[f_i(12) - f_i(11)|\mathrm{X}_i]\} \\
&\quad = E[f_i(12) - f_i(11)]. && (3.2.4)
\end{aligned}
$$

In the same spirit, we might be interested in the average causal effect of high school graduation on high school graduates:

$$
\begin{aligned}
&E\{E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 12] - E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 11]|\mathrm{s}_i = 12\} && (3.2.5) \\
&\quad = E\{E[f_i(12) - f_i(11)|\mathrm{X}_i]|\mathrm{s}_i = 12\} \\
&\quad = E[f_i(12) - f_i(11)|\mathrm{s}_i = 12]. && (3.2.6)
\end{aligned}
$$

This parameter tells us how much high school graduates gained by virtue of having graduated. Likewise, for the effects of college graduation there is a distinction between $E[f_i(16) - f_i(12)|\mathrm{s}_i = 16]$, the average causal effect on college graduates, and $E[f_i(16) - f_i(12)]$, the unconditional average effect.

The unconditional average effect, (3.2.3), can be computed by averaging all the X-specific effects weighted by the marginal distribution of X_i, while the average effect on high school or college graduates averages the X-specific effects weighted by the distribution of X_i in these groups. In both cases, the empirical counterpart is a matching estimator: we make comparisons across schooling groups for individuals with the same covariate values, compute the difference in their average earnings, and then average these differences in some way.

In practice, there are many details to worry about when implementing a matching strategy. We fill in some of the technical details on the mechanics of matching in section 3.3.1. Here we note that a drawback of the matching approach is that it is not automatic; rather, it requires two steps, matching and averaging. Estimating the standard errors of the resulting estimates may not be straightforward, either. A third consideration is that the two-way contrast at the heart of this subsection (high school or college completers versus dropouts) does not do full justice to the problem at hand. Since s_i takes on many values, there are separate average causal effects for each possible increment in s_i, which also must be summarized in some way.[13] These considerations lead us back to regression.

Regression provides an easy-to-use empirical strategy that automatically turns the CIA into causal effects. Two routes can be traced from the CIA to regression. One assumes that $f_i(s)$ is both linear in s and the same for everyone except for an additive error term, in which case linear regression is a

[13] For example, we might construct the average effect over s using the distribution of s_i. In other words, we estimate $E[f_i(s) - f_i(s-1)]$ for each s by matching, and then compute the average difference

$$\sum E[f_i(s) - f_i(s-1)]P(s),$$

where $P(s)$ is the probability mass function for s_i. This is a discrete approximation to the average derivative, $E[f_i'(\mathrm{s}_i)]$.

natural tool to estimate the features of $f_i(s)$. A more general but somewhat longer route recognizes that $f_i(s)$ almost certainly differs for different people, and moreover need not be linear in s. Even so, allowing for random variation in $f_i(s)$ across people and for nonlinearity for a given person, regression can be thought of as a strategy for the estimation of a weighted average of the individual-specific difference, $f_i(s) - f_i(s-1)$. In fact, regression can be seen as a particular sort of matching estimator, capturing an average causal effect, much as (3.2.3) or (3.2.5) does.

At this point, we want to focus on the conditions required for regression to have a causal interpretation and not on the details of the regression-matching analog. We therefore start with the first route, a linear constant effects causal model. Suppose that

$$f_i(s) = \alpha + \rho s + \eta_i. \tag{3.2.7}$$

In addition to being linear, this equation says that the functional relationship of interest is the same for everyone. Again, s is written without an i subscript, because equation (3.2.7) tells us what person i would earn for any value of s, and not just the realized value, s_i. In this case, however, the only individual-specific and random part of $f_i(s)$ is a mean-zero error component, η_i, which captures unobserved factors that determine potential earnings.

Substituting the observed value s_i for s in equation (3.2.7), we have

$$\mathrm{Y}_i = \alpha + \rho \mathrm{s}_i + \eta_i. \tag{3.2.8}$$

Equation (3.2.8) looks like a bivariate regression model, except that equation (3.2.7) explicitly associates the coefficients in (3.2.8) with a causal relationship. Importantly, because equation (3.2.7) is a causal model, s_i may be correlated with potential outcomes, $f_i(s)$, or, in this case, the residual term in (3.2.8), η_i.

Suppose now that the CIA holds given a vector of observed covariates, X_i. In addition to the functional form assumption for potential outcomes embodied in (3.2.8), we decompose the random part of potential earnings, η_i, into a linear function of observable characteristics, X_i, and an error term, v_i:

$$\eta_i = \mathrm{X}_i'\gamma + v_i,$$

where γ is a vector of population regression coefficients that is assumed to satisfy $E[\eta_i|\mathrm{X}_i] = \mathrm{X}_i'\gamma$. Since γ is defined by the regression of η_i on X_i, the residual v_i and X_i are uncorrelated by construction. Moreover, by virtue of the CIA, we have

$$\begin{aligned} E[f_i(s)|\mathrm{X}_i, \mathrm{s}_i] = E[f_i(s)|\mathrm{X}_i] &= \alpha + \rho s + E[\eta_i|\mathrm{X}] \\ &= \alpha + \rho s + \mathrm{X}_i'\gamma. \end{aligned}$$

The residual in the linear causal model

$$\mathrm{Y}_i = \alpha + \rho \mathrm{s}_i + \mathrm{X}_i'\gamma + v_i \tag{3.2.9}$$

is therefore uncorrelated with the regressors, s_i and X_i, and the regression coefficient ρ is the causal effect of interest.

It bears emphasizing once again that the key assumption here is that the observable characteristics, X_i, are the only reason why η_i and s_i (equivalently, $f_i(s)$ and s_i) are correlated. This is the selection-on-observables assumption for regression models discussed over a quarter century ago by Barnow, Cain, and Goldberger (1981). It remains the basis of most empirical work in economics.

3.2.2 The Omitted Variables Bias Formula

In addition to the variable of interest, s_i, we have now introduced a set of control variables, X_i, into our regression. The omitted variables bias (OVB) formula describes the relationship between regression estimates in models with different sets of control variables. This important formula is often motivated by the notion that a longer regression—one with controls, such as (3.2.9)—has a causal interpretation, while a shorter regression does not. The coefficients on the variables included in the shorter regression are therefore said to be biased. In fact, the OVB formula is a mechanical link between coefficient vectors that applies to short and long regressions whether or not the longer regression is causal. Nevertheless, we follow convention and refer to the difference between the included coefficients in a long regression and a short regression as being determined by the OVB formula.

To make this discussion concrete, suppose the relevant set of control variables in the schooling regression can be boiled down to a combination of family background, intelligence, and motivation. Let these specific factors be denoted by a vector, A_i, which we refer to by the shorthand term “ability.” The regression of wages on schooling, s_i, controlling for ability can be written as

$$\text{Y}_i = \alpha + \rho \text{s}_i + A_i'\gamma + e_i, \tag{3.2.10}$$

where α, ρ, and γ are population regression coefficients and e_i is a regression residual that is uncorrelated with all regressors by definition. If the CIA applies given A_i, then ρ can be equated with the coefficient in the linear causal model, (3.2.7), while the residual e_i is the random part of potential earnings that is left over after controlling for A_i.

In practice, ability is hard to measure. For example, the American Current Population Survey (CPS), a large data set widely used in applied microeconomics (and the source of U.S. government data on unemployment rates), tells us nothing about adult respondents’ family background, intelligence, or motivation. What are the consequences of leaving ability out of regression (3.2.10)? The resulting “short regression” coefficient is related to the “long regression” coefficient in equation (3.2.10) as follows:

OMITTED VARIABLES BIAS FORMULA

$$\frac{Cov(\text{Y}_i, \text{s}_i)}{V(\text{s}_i)} = \rho + \gamma'\delta_{As}, \tag{3.2.11}$$

where δ_{As} is the vector of coefficients from regressions of the elements of A_i on s_i. To paraphrase, the OVB formula says:

> Short equals long plus the effect of omitted times the regression of omitted on included.

This formula is easy to derive: plug the long regression into the short regression formula, $\frac{Cov(\text{Y}_i, \text{s}_i)}{V(\text{s}_i)}$. Not surprisingly, the OVB formula is closely related to the regression anatomy formula, (3.1.3), from section 3.1.2. Both the OVB formula and the regression anatomy formula tell us that short and long regression coefficients are the same whenever the omitted and included variables are uncorrelated.[14]

We can use the OVB formula to get a sense of the likely consequences of omitting ability for schooling coefficients. Ability variables have positive effects on wages, and these variables are also likely to be positively correlated with schooling. The short regression coefficient may therefore be "too big" relative to what we want. On the other hand, as a matter of economic theory, the direction of the correlation between schooling and ability is not entirely clear. Some omitted variables may be negatively correlated with schooling, in which case the short regression coefficient may be too small.[15]

Table 3.2.1 illustrates these points using data from the NLSY. The first three entries in the table show that the schooling coefficient decreases from .132 to .114 when family background variables—in this case, parents' education—as well as a few basic demographic characteristics (age, race, census region of residence) are included as controls. Further control for individual ability, as proxied by the Armed Forces Qualification Test (AFQT) score, reduces the schooling coefficient to .087 (the AFQT is used by the military to select soldiers). The OVB formula tells us that these reductions are a result of the fact that the additional controls are positively correlated with both wages and schooling.[16]

[14] Here is the multivariate generalization of OVB: Let β_1^s denote the coefficient vector on a $\kappa_1 \times 1$ vector of variables, X_{1i} in a (short) regression that has no other variables, and let β_1^l denote the coefficient vector on these variables in a (long) regression that includes a $\kappa_2 \times 1$ vector of additional variables, X_{2i}, with coefficient vector β_2^l. Then $\beta_1^s = \beta_1^l + E[X_{1i}X_{1i}']^{-1}E[X_{1i}X_{2i}']\beta_2^l$.

[15] As highly educated people, we like to think that ability and schooling are positively correlated. This is not a foregone conclusion, however: Mick Jagger dropped out of the London School of Economics and Bill Gates dropped out of Harvard, perhaps because the opportunity cost of schooling for these high-ability guys was high (of course, they may also be a couple of very lucky college dropouts).

[16] A large empirical literature investigates the consequences of omitting ability variables from schooling equations. Key early references include Griliches and Mason (1972), Taubman (1976), Griliches (1977), and Chamberlain (1978).

TABLE 3.2.1
Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age Dummies	Col. (2) and Additional Controls*	Col. (3) and AFQT Score	Col. (4), with Occupation Dummies
	.132 (.007)	.131 (.007)	.114 (.007)	.087 (.009)	.066 (.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2,434.

*Additional controls are mother's and father's years of schooling, and dummy variables for race and census region.

Although simple, the OVB formula is one of the most important things to know about regression. The importance of the OVB formula stems from the fact that if you claim an absence of omitted variables bias, then typically you're also saying that the regression you've got is the one you want. And the regression you want usually has a causal interpretation. In other words, you're prepared to lean on the CIA for a causal interpretation of the long regression estimates.

At this point, it's worth considering when the CIA is most likely to give a plausible basis for empirical work. The best-case scenario is random assignment of s_i, conditional on X_i, in some sort of (possibly natural) experiment. An example is the study of a mandatory retraining program for unemployed workers by Black et al. (2003). The authors of this study were interested in whether the retraining program succeeded in raising earnings later on. They exploited the fact that eligibility for the training program they studied was determined on the basis of personal characteristics and past unemployment and job histories. Workers were divided into groups on the basis of these characteristics. While some of these groups of workers were ineligible for training, workers in other groups were required to take training if they did not take a job. When some of the mandatory training groups contained more workers than training slots, training opportunities were distributed by lottery. Hence, training requirements were randomly assigned conditional on the covariates used to assign workers to groups. A regression on a dummy for training plus the personal characteristics, past unemployment variables, and job history variables used to classify workers seems very likely to provide reliable estimates of the causal effect of training.[17]

In the schooling context, there is usually no lottery that directly determines whether someone will go to college or finish high school.[18] Still, we might imagine subjecting individuals of similar ability and from similar family backgrounds to an experiment that encourages school attendance. The Education Maintenance Allowance, which pays British high school students in certain areas to attend school, is one such policy experiment (Dearden et al. 2003).

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

The CIA in Angrist (1998) amounts to the claim that after conditioning on all these observed characteristics, veterans and nonveterans are comparable. This assumption seems worth entertaining since, conditional on X_i, variation in veteran status in the Angrist (1998) study comes solely from the fact

[17] This program appears to raise earnings, primarily because workers offered training went back to work more quickly.

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

that some qualified applicants fail to enlist at the last minute. Of course, the considerations that lead a qualified applicant to "drop out" of the enlistment process could be related to earnings potential, so the CIA is clearly not guaranteed even in this case.

3.2.3 Bad Control

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

The essence of the bad control problem is a version of selection bias, albeit somewhat more subtle than the selection bias discussed in chapter 2 and section 3.2.1. To illustrate, suppose we are interested in the effects of a college degree on earnings and that people can work in one of two occupations, white collar and blue collar. A college degree clearly opens the door to higher-paying white collar jobs. Should occupation therefore be seen as an omitted variable in a regression of wages on schooling? After all, occupation is highly correlated with both education and pay. Perhaps it's best to look at the effect of college on wages for those within an occupation, say white collar only. The problem with this argument is that once we acknowledge the fact that college affects occupation, comparisons of wages by college degree status within an occupation are no longer apples-to-apples comparisons, *even* if college degree completion is randomly assigned.

Here is a formal illustration of the bad control problem in the college/occupation example.[19] Let W_i be a dummy variable that denotes white collar workers and let Y_i denote earnings.

[19] The same problem arises in conditional-on-positive comparisons, discussed in detail in section 3.4.2.

The realization of these variables is determined by college graduation status and potential outcomes that are indexed against C_i. We have

$$\text{Y}_i = \text{C}_i\text{Y}_{1i} + (1 - \text{C}_i)\text{Y}_{0i}$$
$$\text{W}_i = \text{C}_i\text{W}_{1i} + (1 - \text{C}_i)\text{W}_{0i},$$

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

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

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

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

$$\begin{aligned} &E[\text{Y}_i|\text{W}_i = 1, \text{C}_i = 1] - E[\text{Y}_i|\text{W}_i = 1, \text{C}_i = 0] \\ &\quad = E[\text{Y}_{1i}|\text{W}_{1i} = 1, \text{C}_i = 1] - E[\text{Y}_{0i}|\text{W}_{0i} = 1, \text{C}_i = 0]. \end{aligned} \tag{3.2.12}$$

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

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

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

$$\begin{aligned} &E[\text{Y}_{1i}|\text{W}_{1i} = 1] - E[\text{Y}_{0i}|\text{W}_{0i} = 1] \\ &\quad = \underbrace{E[\text{Y}_{1i} - \text{Y}_{0i}|\text{W}_{1i} = 1]}_{\text{Causal effect}} + \underbrace{\{E[\text{Y}_{0i}|\text{W}_{1i} = 1] - E[\text{Y}_{0i}|\text{W}_{0i} = 1]\}.}_{\text{Selection bias}} \end{aligned}$$

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

The selection bias in this context can be positive or negative, depending on the relation between occupational choice, college attendance, and potential earnings. The main point is that even if $\text{Y}_{1i} = \text{Y}_{0i}$, so that there is no causal effect of college on wages, the conditional comparison in (3.2.12) will not tell us this (the regression of Y_i on W_i and C_i has exactly the same problem). It is also incorrect to say that the conditional comparison captures the part of the effect of college that is "not explained by occupation." In fact, the conditional comparison does not tell us much that is useful without a more elaborate model of the links between college, occupation, and earnings.[20]

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

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

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

similar to (3.2.10),

$$\text{Y}_i = \alpha + \rho \text{s}_i + \gamma a_i + e_i, \qquad (3.2.13)$$

where for the purposes of this discussion we've replaced the vector of controls A_i with a scalar ability measure a_i. Think of this as an IQ score that measures innate ability in eighth grade, before any relevant schooling choices are made (assuming everyone completes eighth grade). The error term in this equation satisfies $E[\text{s}_i e_i] = E[a_i e_i] = 0$ by definition. Since a_i is measured before s_i is determined, it is a good control.

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

$$a_{li} = \pi_0 + \pi_1 \text{s}_i + \pi_2 a_i. \qquad (3.2.14)$$

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

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

$$\text{Y}_i = \left(\alpha - \gamma \frac{\pi_0}{\pi_2}\right) + \left(\rho - \gamma \frac{\pi_1}{\pi_2}\right) \text{s}_i + \frac{\gamma}{\pi_2} a_{li} + e_i. \qquad (3.2.15)$$

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

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

One moral of both the bad control and the proxy control stories is that when thinking about controls, timing matters. Variables measured before the variable of interest was determined are generally good controls. In particular, because these variables were determined before the variable of interest, they cannot themselves be outcomes in the causal nexus. Often, however, the timing is uncertain or unknown. In such cases, clear reasoning about causal channels requires explicit assumptions about what happened first, or the assertion that none of the control variables are themselves caused by the regressor of interest.[21]

3.3 Heterogeneity and Nonlinearity

As we saw in the previous section, a linear causal model in combination with the CIA leads to a linear CEF with a causal interpretation. Assuming the CEF is linear, the population

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

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

3.3.1 Regression Meets Matching

The past decade or two has seen increasing interest in matching as an empirical tool. Matching as a strategy to control for covariates is typically motivated by the CIA, as with causal regression in the previous section. For example, Angrist (1998) used matching to estimate the effects of voluntary military service on the later earnings of soldiers. These matching estimates have a causal interpretation assuming that, conditional on the individual characteristics the military uses to select soldiers (age, schooling, test scores), veteran status is independent of potential earnings. Matching estimators are appealingly simple: at bottom, matching amounts to covariate-specific treatment-control comparisons, weighted together to produce a single overall average treatment effect.

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

To flesh out this idea, it helps to look more deeply into the mathematical structure of the matching and regressions *estimands*, that is, the population quantities that these methods attempt to estimate. For regression, of course, the estimand is a vector of population regression coefficients. The matching estimand is typically a weighted average of contrasts or comparisons across cells defined by covariates. This is easiest to see in the case of discrete covariates, as in the military service example, and for a discrete regressor such as veteran status, which we denote here by the dummy $\textsc{d}_i$. Since treatment takes on only two values, we can use the notation $\textsc{y}_{1i}$ and $\textsc{y}_{0i}$ to denote potential outcomes. A parameter of primary interest in this context is the average effect of treatment on the treated, $E[\textsc{y}_{1i} - \textsc{y}_{0i}|\textsc{d}_i = 1]$. This tells us the difference between the average earnings of soldiers, $E[\textsc{y}_{1i}|\textsc{d}_i = 1]$, an observable quantity, and the counterfactual average earnings they would have obtained if they had not served, $E[\textsc{y}_{0i}|\textsc{d}_i = 1]$. Simple comparisons of earnings by veteran status give a biased measure of the effect of treatment on the treated unless $\textsc{d}_i$ is independent of $\textsc{y}_{0i}$. Specifically,

$$\begin{aligned}&E[\textsc{y}_i|\textsc{d}_i = 1] - E[\textsc{y}_i|\textsc{d}_i = 0]\\&\quad = E[\textsc{y}_{1i} - \textsc{y}_{0i}|\textsc{d}_i = 1] + \{E[\textsc{y}_{0i}|\textsc{d}_i = 1] - E[\textsc{y}_{0i}|\textsc{d}_i = 0]\}.\end{aligned}$$

In other words, the observed earnings difference by veteran status equals the average effect of treatment on the treated plus selection bias. This parallels the discussion of selection bias in chapter 2.

The CIA in this context says that

$$\{\textsc{y}_{0i}, \textsc{y}_{1i}\} \perp\!\!\!\perp \textsc{d}_i|X_i.$$

Given the CIA, selection bias disappears after conditioning on X_i, so the effect of treatment on the treated can be constructed by iterating expectations over X_i:

$$\begin{aligned}\delta_{TOT} &\equiv E[\textsc{y}_{1i} - \textsc{y}_{0i}|\textsc{d}_i = 1]\\&= E\{E[\textsc{y}_{1i} - \textsc{y}_{0i}|X_i, \textsc{d}_i = 1]|\textsc{d}_i = 1\}\\&= E\{E[\textsc{y}_{1i}|X_i, \textsc{d}_i = 1] - E[\textsc{y}_{0i}|X_i, \textsc{d}_i = 1]|\textsc{d}_i = 1\}.\end{aligned}$$

Of course, $E[\text{Y}_{0i}|\text{X}_i, \text{D}_i = 1]$ is counterfactual. By virtue of the CIA, however,

$$E[\text{Y}_{0i}|\text{X}_i, \text{D}_i = 0] = E[\text{Y}_{0i}|\text{X}_i, \text{D}_i = 1].$$

Therefore,

$$\begin{aligned}\delta_{TOT} &= E\{E[\text{Y}_{1i}|\text{X}_i, \text{D}_i = 1] - E[\text{Y}_{0i}|\text{X}_i, \text{D}_i = 0]|\text{D}_i = 1\} \\ &= E[\delta_X|\text{D}_i = 1], \qquad (3.3.1)\end{aligned}$$

where

$$\delta_X \equiv E[\text{Y}_i|\text{X}_i, \text{D}_i = 1] - E[\text{Y}_i|\text{X}_i, \text{D}_i = 0],$$

is the difference in mean earnings by veteran status at each value of X_i. At a particular value, say $\text{X}_i = x$, we write δ_x.

The matching estimator in Angrist (1998) uses the fact that X_i is discrete to construct the sample analog of the right-hand side of (3.3.1). In the discrete case, the matching estimand can be written

$$E[\text{Y}_{1i} - \text{Y}_{0i}|\text{D}_i = 1] = \sum_x \delta_x P(\text{X}_i = x|\text{D}_i = 1), \qquad (3.3.2)$$

where $P(\text{X}_i = x|\text{D}_i = 1)$ is the probability mass function for X_i given $\text{D}_i = 1$.[22] In this case, X_i takes on values determined by all possible combinations of year of birth, test score group, year of application to the military, and educational attainment at the time of application. The test score in this case is from the AFQT, used by the military to categorize the mental abilities of applicants (we included this as a control in the schooling regression discussed in section 3.2.2). The Angrist (1998) matching estimator replaces δ_X by the sample veteran-nonveteran earnings difference for each combination of covariates and then combines these in a weighted average using the empirical distribution of covariates among veterans.

[22] This matching estimator is discussed by Rubin (1977) and used by Card and Sullivan (1988) to estimate the effect of subsidized training on employment.

Note also that we can just as easily construct the unconditional average treatment effect,

$$\begin{aligned}\delta_{ATE} &= E\{E[\mathrm{Y}_{1i}|\mathrm{X}_i, \mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{X}_i, \mathrm{D}_i = 0]\}\\ &= \sum_x \delta_x P(\mathrm{X}_i = x) = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}]. \qquad (3.3.3)\end{aligned}$$

This is the expectation of δ_X using the marginal distribution of X_i instead of the distribution among the treated. δ_{TOT} tells us how much the typical *soldier* gained or lost as a consequence of military service, while δ_{ATE} tells us how much the typical *applicant* to the military gained or lost (since the Angrist, 1998, population consists of applicants.)

The U.S. military tends to be fairly picky about its soldiers, especially after downsizing at the end of the cold war. For the most part, the military now takes only high school graduates with test scores in the upper half of the test score distribution. Applicant screening therefore generates positive selection bias in naive comparisons of veteran and nonveteran earnings. This can be seen in table 3.3.1, which reports differences-in-means, matching, and regression estimates of the effect of voluntary military service on the 1988–91 Social Security–taxable earnings of men who applied to join the military between 1979 and 1982. The matching estimates were constructed from the sample analog of (3.3.2). Although white veterans earned \$1,233 more than white nonveterans, this estimated veteran effect becomes negative once differences in covariates are matched away. Similarly, while nonwhite veterans earned \$2,449 more than nonwhite nonveterans, controlling for covariates reduces this difference to \$840.

Table 3.3.1 also shows regression estimates of the effect of voluntary military service, controlling for the same set of covariates that were used to construct the matching estimates. These are estimates of δ_R in the equation

$$\mathrm{Y}_i = \sum_x d_{ix}\alpha_x + \delta_R \mathrm{D}_i + e_i, \qquad (3.3.4)$$

where $d_{ix} = 1[\mathrm{X}_i = x]$ is a dummy variable that indicates $\mathrm{X}_i = x$, α_x is a regression effect for $\mathrm{X}_i = x$ and δ_R is the

TABLE 3.3.1
Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings

Race	Average Earnings in 1988–1991 (1)	Differences in Means by Veteran Status (2)	Matching Estimates (3)	Regression Estimates (4)	Regression Minus Matching (5)
Whites	14,537	1,233.4 (60.3)	−197.2 (70.5)	−88.8 (62.5)	108.4 (28.5)
Non-whites	11,664	2,449.1 (47.4)	839.7 (62.7)	1,074.4 (50.7)	234.7 (32.5)

Notes: Adapted from Angrist (1998, tables II and V). Standard errors are reported in parentheses. The table shows estimates of the effect of voluntary military service on the 1988–91 Social Security–taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 nonwhites in the sample.

regression estimand. Note that this regression model allows a separate parameter for every value taken on by the covariates. This model can therefore be said to be saturated-in-X_i, since it includes a parameter for every value of X_i. It is not fully saturated, however, because there is a single additive effect for D_i with no $D_i \cdot X_i$ interactions.

Despite the fact that the matching and regression estimates control for the same variables, the regression estimates in table 3.3.1 are somewhat larger for nonwhites and less negative for whites. In fact, the differences between the matching and regression results are statistically significant. At the same time, the two estimation strategies present a broadly similar picture of the effects of military service. The reason the regression and matching estimates are similar is that regression, too, can be seen as a sort of matching estimator: the regression estimand differs from the matching estimands only in the weights used to combine the covariate-specific effects, δ_X, into a single average effect. In particular, while matching uses the distribution of covariates among the treated to weight covariate-specific estimates into an estimate of the effect of treatment on the treated, regression produces a variance-weighted average of these effects.

To see this, start by using the regression anatomy formula to write the coefficient on D_i in the regression of Y_i on X_i and D_i as

$$\begin{aligned}\delta_R &= \frac{Cov(\text{Y}_i, \tilde{\text{D}}_i)}{V(\tilde{\text{D}}_i)} && (3.3.5)\\ &= \frac{E[(\text{D}_i - E[\text{D}_i|X_i])\text{Y}_i]}{E[(\text{D}_i - E[\text{D}_i|X_i])^2]}\\ &= \frac{E\{(\text{D}_i - E[\text{D}_i|X_i])E[\text{Y}_i|\text{D}_i, X_i]\}}{E[(\text{D}_i - E[\text{D}_i|X_i])^2]}. && (3.3.6)\end{aligned}$$

The second equality in this set of expressions uses the fact that saturating the model in X_i means $E[\text{D}_i|X_i]$ is linear. Hence, $\tilde{\text{D}}_i$, which is defined as the residual from a regression of D_i on X_i, is the difference between D_i and $E[\text{D}_i|X_i]$. The third equality uses the fact that the regression of Y_i on D_i and X_i is the same as the regression of Y_i on $E[\text{Y}_i|\text{D}_i,X_i]$ (this we know from the regression CEF theorem, 3.1.6).

To simplify further, we expand the CEF, $E[\text{Y}_i|\text{D}_i,X_i]$, to get

$$E[\text{Y}_i|\text{D}_i, X_i] = E[\text{Y}_i|\text{D}_i = 0, X_i] + \delta_X \text{D}_i,$$

and then substitute for $E[\text{Y}_i|\text{D}_i,X_i]$ in the numerator of (3.3.6). This gives

$$\begin{aligned}&E\{(\text{D}_i - E[\text{D}_i|X_i])E[\text{Y}_i|\text{D}_i, X_i]\}\\ &\quad = E\{(\text{D}_i - E[\text{D}_i|X_i])E[\text{Y}_i|\text{D}_i = 0, X_i]\}\\ &\qquad + E\{(\text{D}_i - E[\text{D}_i|X_i])\text{D}_i\delta_X\}.\end{aligned}$$

The first term on the right-hand side is zero because $E[\text{Y}_i|\text{D}_i = 0,X_i]$ is a function of X_i only and is therefore uncorrelated with $(\text{D}_i - E[\text{D}_i|X_i])$. Similarly, the second term simplifies to

$$E\{(\text{D}_i - E[\text{D}_i|X_i])\text{D}_i\delta_X\} = E\{(\text{D}_i - E[\text{D}_i|X_i])^2\delta_X\}.$$

At this point, we've shown

$$\delta_R = \frac{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2\delta_X]}{E[(\mathrm{D}_i - E[\mathrm{D}\ {}_i|\mathrm{X}_i])^2]}$$

$$= \frac{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\delta_X\}}{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\}} = \frac{E[\sigma_D^2(\mathrm{X}_i)\delta_X]}{E[\sigma_D^2(\mathrm{X}_i)]}, \quad (3.3.7)$$

where

$$\sigma_D^2(\mathrm{X}_i) \equiv E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]$$

is the conditional variance of D_i given X_i. This establishes that the regression model, (3.3.4), produces a treatment-variance weighted average of δ_X.

Because the regressor of interest, D_i, is a dummy variable, one last step can be taken. In this case, $\sigma_D^2(\mathrm{X}_i) = P(\mathrm{D}_i = 1|\mathrm{X}_i)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i))$, so

$$\delta_R = \frac{\sum_x \delta_x[P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))]P(\mathrm{X}_i = x)}{\sum_x [P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))]P(\mathrm{X}_i = x)}.$$

This shows that the regression estimand weights each covariate-specific treatment effect by $[P(\mathrm{X}_i = x|\mathrm{D}_i = 1)(1 - P(\mathrm{X}_i = x|\mathrm{D}_i = 1))]P(\mathrm{X}_i = x)$. In contrast, the matching estimand for the effect of treatment on the treated can be written

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1] = \sum_x \delta_x P(\mathrm{X}_i = x|\mathrm{D}_i = 1)$$

$$= \frac{\sum_x \delta_x P(\mathrm{D}_i = 1|\mathrm{X}_i = x)P(\mathrm{X}_i = x)}{\sum_x P(\mathrm{D}_i = 1|\mathrm{X}_i = x)P(\mathrm{X}_i = x)},$$

using the fact that

$$P(\mathrm{X}_i = x|\mathrm{D}_i = 1) = \frac{P(\mathrm{D}_i = 1|\mathrm{X}_i = x) \cdot P(\mathrm{X}_i = x)}{P(\mathrm{D}_i = 1)}.$$

So the weights used to construct $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1]$ are proportional to the probability of treatment at each value of the covariates. The regression and matching weighting schemes therefore differ unless treatment is independent of covariates.

An important point coming out of this derivation is that the treatment-on-the-treated estimand puts the most weight on covariate cells containing those who are most likely to be treated. In contrast, regression puts the most weight on covariate cells where the conditional variance of treatment status is largest. As a rule, treatment variance is maximized when $P(\mathrm{D}_i = 1|\mathrm{X}_i = x) = \frac{1}{2}$, in other words, for cells where there are equal numbers of treated and control observations. The difference in weighting schemes is of little importance if δ_x does not vary across cells (though weighting still affects the statistical efficiency of estimators). In this example, however, men who were most likely to serve in the military appear to benefit least from their service. This is probably because those most likely to serve were most qualified and therefore also had the highest civilian earnings potential. This fact leads matching estimates of the effect of military service to be smaller than regression estimates based on the same vector of control variables.[23]

Also important is the fact that neither the regression nor the covariate-matching estimands give any weight to covariate cells that do not contain both treated and control observations. Consider a value of X_i, say x^*, where either no one is treated or everyone is treated. Then, δ_{x^*} is undefined, while the regression weights, $[P(\mathrm{D}_i = 1|\mathrm{X}_i = x^*)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x^*))]$, are zero. In the language of the econometric literature on matching, with saturated control for covariates both the regression and matching estimands impose *common support*, that is, they are limited to covariate values where both treated and control observations are found.[24]

[23] It's no surprise that regression gives the most weight to cells where $P(\mathrm{D}_i = 1|\mathrm{X}_i = x) = \frac{1}{2}$ since regression is efficient for a homoskedastic constant effects linear model. We should expect an efficient estimator to give the most weight to cells where the common treatment effect is estimated most precisely. With homoskedastic residuals, the most precise treatment effects come from cells where the probability of treatment equals $\frac{1}{2}$.

[24] The *support* of a random variable is the set of realizations that occur with positive probability. See Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2001) for a discussion of common support in matching.

The step from estimand to estimator is a little more complicated. In practice, both regression and matching estimators are implemented using modeling assumptions that implicitly involve a certain amount of extrapolation across cells. For example, matching estimators often combine covariate cells with few observations. This violates common support if the cells being combined do not all have both treated and nontreated observations. Regression models that are not saturated in X_i may also violate common support, since covariate cells without both treated and control observations can end up contributing to the estimates by extrapolation. Here, too, however, we see a symmetry between the matching and regression strategies: they are in the same class, in principle, and require the same sort of compromises in practice.[25]

Even More on Regression and Matching: Ordered and Continuous Treatments★

Does the quasi-matching interpretation of regression outlined above for a binary treatment variable apply to models with ordered and continuous treatments? The long answer is fairly technical and may be more than you want to know. The short answer is, to one degree or another, yes.

As we've already discussed, the population OLS slope vector always provides the MMSE linear approximation to the CEF. This, of course, works for ordered and continuous regressors as well as for binary. A related property is the fact that regression coefficients have an "average derivative" interpretation. In multivariate regression models, this interpretation is unfortunately complicated by the fact that the OLS slope vector is a matrix-weighted average of the gradient of the

[25] Matching problems involving finely distributed X-variables are often solved by aggregating values to make coarser groupings or by pairing observations that have similar, though not necessarily identical, values. See Cochran (1965), Rubin (1973), or Rosenbaum (1995, chapter 3) for discussions of this approach. With continuously distributed covariates, matching estimators are biased because matches are imperfect. Abadie and Imbens (2008) have recently shown that a regression-based bias correction can eliminate the (asymptotic) bias from imperfect matches.

CEF. Matrix-weighted averages are difficult to interpret except in special cases (see Chamberlain and Leamer, 1976). An important special case when the average derivative property is relatively straightforward is in regression models for an ordered or continuous treatment with a saturated model for covariates. To avoid lengthy derivations, we simply explain the formulas. A derivation is sketched in the appendix to this chapter. For additional details, see the appendix to Angrist and Krueger (1999).

For the purposes of this discussion, the treatment intensity, s_i, is assumed to be a continuously distributed random variable, not necessarily non-negative. Suppose that the CEF of interest can be written $h(t) \equiv E[Y_i|s_i = t]$ with derivative $h'(t)$. Then

$$\frac{E[Y_i(s_i - E[s_i])]}{E[s_i(s_i - E[s_i])]} = \frac{\int h'(t)\mu_t dt}{\int \mu_t dt}, \tag{3.3.8}$$

where

$$\mu_t \equiv \{E[s_i|s_i \geq t] - E[s_i|s_i < t]\}\{P(s_i \geq t)[1 - P(s_i \geq t)\}, \tag{3.3.9}$$

and the integrals in (3.3.8) run over the possible values of s_i. This formula (derived by Yitzhaki, 1996) weights each possible value of s_i in proportion to the difference in the conditional mean of s_i above and below that value. More weight is also given to points close to the median of s_i, since $P(s_i \geq t) \cdot [1 - P(s_i \geq t)]$ is maximized there.

With covariates, X_i, the weights in (3.3.8) become X-specific. A covariate-averaged version of the same formula applies to the multivariate regression coefficient of Y_i on s_i, after partialing out X_i. In particular,

$$\frac{E[Y_i(s_i - E[s_i|X_i])]}{E[s_i(s_i - E[s_i|X_i])]} = \frac{E[\int h'_X(t)\mu_{tX} dt]}{E[\int \mu_{tX} dt]}, \tag{3.3.10}$$

where $h'_X(t) \equiv \frac{\partial E[Y_i|X_i, s_i = t]}{\partial t}$ and

$$\mu_{tX} \equiv \{E[s_i|X_i, s_i \geq t] - E[s_i|X_i, s_i < t]\} \\ \times \{P(s_i \geq t|X_i)[1 - P(s_i \geq t|X_i)]\}.$$

Equation (3.3.10) reflects two types of averaging: an integral that averages *along* the length of a nonlinear CEF at fixed covariate values, and an expectation that averages *across* covariate cells. An important point in this context is that population regression coefficients contain no information about the effect of s_i on the CEF for values of X_i where $P(\mathrm{s}_i \geq t|\mathrm{X}_i)$ equals zero or one. This includes values of X_i where s_i is fixed. It's also worth noting that if s_i is a dummy variable, we can extract equation (3.3.7) from the more general formula, (3.3.10).

Angrist and Krueger (1999) constructed the average weighting function for a schooling regression with state of birth and year of birth covariates. Although equations (3.3.8) and (3.3.10) may seem arcane or at least nonobvious, in this example the average weights, $E[\mu_{tX}]$, turn out to be a reasonably smooth symmetric function of t, centered at the mode of s_i.

The implications of (3.3.8) or (3.3.10) can be explored further given a model for the distribution of regressors. Suppose, for example, that s_i is normally distributed. Let $z_i = \frac{\mathrm{s}_i - E(\mathrm{s}_i)}{\sigma_s}$, where σ_s is the standard deviation of s_i, so that z_i is standard normal. Then

$$\begin{aligned} E[\mathrm{s}_i|\mathrm{s}_i \geq t] &= E(\mathrm{s}_i) + \sigma_s E\left[z_i|z_i \geq \frac{t - E(\mathrm{s}_i)}{\sigma_s}\right] \\ &= E(\mathrm{s}_i) + \sigma_s E[z_i|z_i \geq t^*]. \end{aligned}$$

From truncated normal formulas (see, e.g., Johnson and Kotz, 1970), we know that

$$E[z_i|z_i > t^*] = \frac{\phi(t^*)}{[1 - \Phi(t^*)]} \quad \text{and} \quad E[z_i|z_i < t^*] = \frac{-\phi(t^*)}{\Phi(t^*)}.$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions. Substituting in the formula for μ_t, (3.3.9), we have

$$\mu_t = \sigma_s \left\{ \frac{\phi(t^*)}{[1 - \Phi(t^*)]} - \frac{-\phi(t^*)}{\Phi(t^*)} \right\} [1 - \Phi(t^*)]\Phi(t^*) = \sigma_s \phi(t^*).$$

We have therefore shown that

$$\frac{Cov(\mathrm{Y}_i, \mathrm{s}_i)}{V(\mathrm{s}_i)} = E[h'(\mathrm{s}_i)].$$

In other words, when s_i is normal, the regression of Y_i on s_i is the unconditional average derivative, $E[h'(s_i)]$. Of course, this result is a special case of a special case.[26] Still, it seems reasonable to imagine that normality might not matter very much. And in our empirical experience, the average derivatives (also called "marginal effects") constructed from parametric nonlinear models (e.g., probit or Tobit) are usually indistinguishable from the corresponding regression coefficients, regardless of the distribution of regressors. We expand on this point in section 3.4.2.

3.3.2 Control for Covariates Using the Propensity Score

The most important result in regression theory is the OVB formula, which tells us that coefficients on included variables are unaffected by the omission of variables when the variables omitted are uncorrelated with the variables included. The propensity score theorem, due to Rosenbaum and Rubin (1983), extends this idea to estimation strategies that rely on matching instead of regression, where the causal variable of interest is a treatment dummy.[27]

The propensity score theorem says that if potential outcomes are independent of treatment status conditional on a multivariate covariate vector X_i, then potential outcomes are independent of treatment status conditional on a scalar function of covariates, the propensity score, defined as $p(X_i) \equiv E[D_i|X_i] = P[D_i = 1|X_i]$. Formally, we have the following theorem:

Theorem 3.3.1 *The Propensity Score Theorem.*
Suppose the CIA holds such that $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i|X_i$. *Then* $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i|p(X_i)$.

[26] Other specialized results in this spirit appear in Yitzhaki (1996) and Ruud (1986), who considers distribution-free estimation of limited-dependent-variable models.

[27] Propensity score methods can be adapted to multivalued treatments, though this has yet to catch on. See Imbens (2000) for an effort in this direction.

Proof. It's enough to show that $P[\text{D}_i = 1|\text{Y}_{ji}, p(\text{X}_i)]$ does not depend on Y_{ji} for $j = 0, 1$:

$$\begin{aligned} P[\text{D}_i = 1|\text{Y}_{ji}, p(\text{X}_i)] &= E[\text{D}_i|\text{Y}_{ji}, p(\text{X}_i)] \\ &= E\{E[\text{D}_i|\text{Y}_{ji}, p(\text{X}_i), \text{X}_i]|\text{Y}_{ji}, p(\text{X}_i)\} \\ &= E\{E[\text{D}_i|\text{Y}_{ji}, \text{X}_i]|\text{Y}_{ji}, p(\text{X}_i)\} \\ &= E\{E[\text{D}_i|\text{X}_i]|\text{Y}_{ji}, p(\text{X}_i)\}, \text{ by the CIA.} \end{aligned}$$

But $E\{E[\text{D}_i|\text{X}_i]|\text{Y}_{ji}, p(\text{X}_i)\} = E\{p(\text{X}_i)|\text{Y}_{ji}, p(\text{X}_i)\}$, which is clearly just $p(\text{X}_i)$.

Like the OVB formula for regression, the propensity score theorem says that you need only control for covariates that affect the probability of treatment. But it also says something more: the only covariate you really need to control for is the probability of treatment itself. In practice, the propensity score theorem is usually used for estimation in two steps: first, $p(\text{X}_i)$ is estimated using some kind of parametric model, say, logit or probit. Then estimates of the effect of treatment are computed either by matching on the estimated score from this first step or using a weighting scheme described below (see Imbens, 2004, for an overview).

Direct propensity score matching works in the same way as covariate matching except that we match on the score instead of the covariates directly. By the propensity score theorem and the CIA,

$$\begin{aligned} &E[\text{Y}_{1i} - \text{Y}_{0i}|\text{D}_i = 1] \\ &\quad = E\{E[\text{Y}_i|p(\text{X}_i), \text{D}_i = 1] - E[\text{Y}_i|p(\text{X}_i), \text{D}_i = 0]|\text{D}_i = 1\}. \end{aligned}$$

Estimates of the effect of treatment on the treated can therefore be obtained by stratifying on an estimate of $p(\text{X}_i)$ and substituting conditional sample averages for expectations or by matching each treated observation to controls with similar values of the propensity score (both of these approaches were used by Dehejia and Wahba, 1999). Alternatively, a model-based or nonparametric estimate of $E[\text{Y}_i|p(\text{X}_i),\text{D}_i]$ can be substituted for these conditional mean functions and the outer expectation replaced with a sum (as in Heckman, Ichimura, and Todd, 1998).

The somewhat niftier weighting approach to propensity score estimation skips the cumbersome matching step by exploiting the fact that the CIA implies $E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)}\right] = E[\mathrm{Y}_{1i}]$ and $E\left[\frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{(1-p(\mathrm{X}_i))}\right] = E[\mathrm{Y}_{0i}]$.[28] Therefore, given a scheme for estimating $p(\mathrm{X}_i)$, we can construct estimates of the average treatment effect from the sample analog of

$$\begin{aligned} E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] &= E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)} - \frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{1-p(\mathrm{X}_i)}\right] \\ &= E\left[\frac{(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i}{p(\mathrm{X}_i)(1-p(\mathrm{X}_i))}\right]. \qquad (3.3.11) \end{aligned}$$

This last expression is an estimand of the form suggested by Newey (1990) and Robins, Mark, and Newey (1992). We can similarly calculate the effect of treatment on the treated from the sample analog of:

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{D}_i = 1] = E\left[\frac{(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i}{(1-p(\mathrm{X}_i))P(\mathrm{D}_i = 1)}\right]. \qquad (3.3.12)$$

The idea that you can correct for nonrandom sampling via weighting by the reciprocal of the probability of selection dates back to Horvitz and Thompson (1952). Of course, to make this approach feasible, and for the resulting estimates to be consistent, we need a consistent estimator of $p(\mathrm{X}_i)$.

The Horvitz-Thompson version of the propensity score approach is appealing, since the estimator is essentially automated, with no cumbersome matching required. The Horvitz-Thompson approach also highlights the close link between propensity score matching and regression, much as discussed for covariate matching in section 3.3.1. Consider again the regression estimand, δ_R, for the population regression of Y_i on D_i, controlling for a saturated model for covariates. This estimand can be written

$$\delta_R = \frac{E[(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i]}{E[p(\mathrm{X}_i)(1-p(\mathrm{X}_i))]}. \qquad (3.3.13)$$

[28] To see this, iterate over X_i: $E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)}\right] = E\left\{E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)}|\mathrm{X}_i\right]\right\}$; $E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)}|\mathrm{X}_i\right] = \frac{E[\mathrm{Y}_i|\mathrm{D}_i=1,\mathrm{X}_i]p(\mathrm{X}_i)}{p(\mathrm{X}_i)} = E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1, \mathrm{X}_i] = E[\mathrm{Y}_{1i}|\mathrm{X}_i]$.

The two Horvitz-Thompson matching estimands, (3.3.11) and (3.3.12), and the regression estimand are all in the class of weighted average estimands considered by Hirano, Imbens, and Ridder (2003):

$$E\left\{g(\mathrm{X}_i)\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)} - \frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{(1-p(\mathrm{X}_i))}\right]\right\}, \qquad (3.3.14)$$

where $g(\mathrm{X}_i)$ is a known weighting function. (To go from estimand to estimator, replace $p(\mathrm{X}_i)$ with a consistent estimator and replace expectations with sums.) For the average treatment effect, set $g(\mathrm{X}_i) = 1$; for the effect on the treated, set $g(\mathrm{X}_i) = \frac{p(\mathrm{X}_i)}{P(\mathrm{D}_i=1)}$; and for regression, set

$$g(\mathrm{X}_i) = \frac{p(\mathrm{X}_i)(1-p(\mathrm{X}_i))}{E[p(\mathrm{X}_i)(1-p(\mathrm{X}_i))]}.$$

This similarity highlights once again the fact that regression and matching—including propensity score matching—are not really different animals, at least not until we specify a model for the propensity score.

A big question here is how best to model and estimate $p(\mathrm{X}_i)$, or how much smoothing or stratification to use when estimating $E[\mathrm{Y}_i|p(\mathrm{X}_i), \mathrm{D}_i]$, especially if the covariates are continuous. The regression analog of this question is how to parameterize the control variables (e.g., polynomials or main effects and interaction terms if the covariates are coded as discrete). The answer to this is inherently application-specific. A growing empirical literature suggests that a logit model for the propensity score with a few polynomial terms in continuous covariates works well in practice, though this cannot be a theorem, and inevitably, some experimentation will be required (see, e.g., Dehejia and Wahba, 1999).[29]

A developing theoretical literature has produced some thought-provoking theorems on the efficient use of the propensity score. First, from the point of view of asymptotic efficiency, there is usually a cost to matching on the propensity

[29] Andrea Ichino and Sascha Becker have posted Stata programs that implement various matching estimators; see Becker and Ichino (2002).

score instead of full covariate matching. We can get lower asymptotic standard errors by matching on any covariate that explains outcomes, whether or not it turns up in the propensity score. This we know from Hahn's (1998) investigation of the maximal precision of estimates of treatment effects under the CIA, with and without knowledge of the propensity score. For example, in Angrist (1998), there is an efficiency gain from matching on year of birth, even if the probability of serving in the military is unrelated to birth year, because earnings are related to birth year. A regression analog for this point is the result that even in a scenario with no OVB, the long regression generates more precise estimates of the coefficients on the variables included in a short regression whenever the omitted variables have some predictive power for outcomes (see section 3.1.3).

Hahn's (1998) results raise the question of why we should ever bother with estimators that use the propensity score. A philosophical argument is that the propensity score rightly focuses researcher attention on models for treatment assignment, something about which we may have reasonably good information, instead of the typically more complex and mysterious process determining outcomes. This view seems especially compelling when treatment assignment is the product of human institutions or government regulations, while the process determining outcomes is more anonymous (e.g., a market). For example, in a time series evaluation of the causal effects of monetary policy, Angrist and Kuersteiner (2004) argue that we know more about how the Federal Reserve sets interest rates than about the process determining GDP. In the same spirit, it may also be easier to validate a model for treatment assignment than to validate a model for outcomes (see Rosenbaum and Rubin, 1985, for a version of this argument).

A more precise though purely statistical argument for using the propensity score is laid out in Angrist and Hahn (2004). This paper shows that even though there is no *asymptotic* efficiency gain from the use of estimators based on the propensity score, there will often be a gain in precision in finite samples. Since all real data sets are finite, this result is empirically relevant. Intuitively, if the covariates omitted from the propensity score explain little of the variation in outcomes (in a purely statistical sense), it may be better to ignore them than to bear the statistical burden imposed by the need to estimate their effects. This is easy to see in studies using data sets such as the NLSY, where there are hundreds of covariates that might predict outcomes. In practice, we focus on a small subset of all possible covariates. This subset is usually chosen with an eye to what predicts treatment.

Finally, Hirano, Imbens, and Ridder (2003) provide an alternative asymptotic resolution of the "propensity score paradox" generated by Hahn's (1998) theorems. They show that even though estimates of treatment effects based on a known propensity score are inefficient, for models with continuous covariates, a Horvitz-Thompson-type weighting estimator is efficient when the weighting scheme uses a *nonparametric* estimate of the score. The facts that the propensity score is estimated and that it is estimated nonparametrically are both key for the Hirano, Imbens, and Ridder conclusions.

Do the Hirano, Imbens, and Ridder (2003) results resolve the propensity score paradox? For the moment, we prefer the finite-sample resolution given by Angrist and Hahn (2004). The latter result highlights the fact that it is the researchers' willingness to impose restrictions on the score that gives propensity score-based inference its conceptual and statistical power. In Angrist (1998), for example, an application with high-dimensional though discrete covariates, the unrestricted nonparametric estimator of the score is just the empirical probability of treatment in each covariate cell. With this nonparametric estimator plugged in for $p(X_i)$, it is straightforward to show that the sample analogs of (3.3.11) and (3.3.12) are algebraically equivalent to the corresponding full-covariate matching estimators. Hence, it's no surprise that score-based estimation comes out efficient, since full-covariate matching is the asymptotically efficient benchmark. An essential element of propensity score methods is the use of prior knowledge for dimension reduction. The statistical payoff is an improvement in finite-sample behavior. If you're not prepared to smooth, restrict, or otherwise reduce the dimensionality of the matching problem in a manner that has real empirical consequences, then you might as well go for full covariate matching or saturated regression control.

3.3.3 Propensity Score Methods versus Regression

Propensity score methods shift attention from the estimation of $E[\text{Y}_i|\text{X}_i,\text{D}_i]$ to the estimation of the propensity score, $p(\text{X}_i) \equiv E[\text{D}_i|\text{X}_i]$. This is attractive in applications where the latter is easier to model or motivate. For example, Ashenfelter (1978) showed that participants in government-funded training programs often have suffered a marked preprogram dip in earnings, a pattern found in many later studies. If this dip is the only thing that makes trainees special, then we can estimate the causal effect of training on earnings by controlling for past earnings dynamics. In practice, however, it's hard to match on earnings dynamics since earnings histories are both continuous and multidimensional. Dehejia and Wahba (1999) argue in this context that the causal effects of training programs are better estimated by conditioning on the propensity score than by conditioning on the earnings histories themselves.

The propensity score estimates reported by Dehejia and Wahba are remarkably close to the estimates from the randomized trial that constitute their benchmark. Nevertheless, we believe regression should be the starting point for most empirical projects. This is not a theorem; undoubtedly, there are circumstances in which propensity score matching provides more reliable estimates of average causal effects. The first reason we don't find ourselves on the propensity score bandwagon is practical: there are many details to be filled in when implementing propensity score matching, such as how to model the score and how to do inference; these details are not yet standardized. Different researchers might therefore reach very different conclusions, even when using the same data and covariates. Moreover, as we've seen with the Horvitz-Thompson estimands, there isn't very much theoretical daylight between regression and propensity score weighting. If the regression model for covariates is fairly flexible, say, close to saturated, regression can be seen as a type of propensity score weighting, so the difference is mostly in the implementation. In practice you may be far from saturation, but with the right covariates this shouldn't matter.

The face-off between regression and propensity score matching is illustrated here using the same National Supported Work (NSW) sample featured in Dehejia and Wahba (1999).[30] The NSW is a mid-1970s program that provided work experience to recipients with weak labor force attachment. Somewhat unusually for its time, the NSW was evaluated in a randomized trial. Lalonde's (1986) pathbreaking analysis compared the results from the NSW randomized study to econometric results using nonexperimental control groups drawn from the PSID and the CPS. He came away pessimistic because plausible nonexperimental methods generated a wide range of results, many of which were far from the experimental estimates. Moreover, Lalonde argued, an objective investigator, not knowing the results of the randomized trial, would be unlikely to pick the best econometric specifications and observational control groups.

In a striking second take on the Lalonde (1986) findings, Dehejia and Wahba (1999) found they could come close to the NSW experimental results by matching the NSW treatment group to observational control groups selected using the propensity score. They demonstrated this using various comparison groups. Following Dehejia and Wahba (1999), we look again at two of the CPS comparison groups, first, a largely unselected sample (CPS-1), and then a narrower comparison group selected from the recently unemployed (CPS-3).

Table 3.3.2 (columns 1–4 of which are a replication of table 1 in Dehejia and Wahba, 1999) reports descriptive statistics for the NSW treatment group, the randomly selected NSW control group, and our two observational control groups. The NSW treatment group and the randomly selected NSW control groups are younger, less educated, more likely to be nonwhite, and have much lower earnings than the general population represented by the CPS-1 sample. The CPS-3 sample matches the NSW treatment group more closely but still shows

[30] A more extended propensity-score face-off appears in the exchange between Smith and Todd (2005) and Dehejia (2005).

Table 3.3.2
Covariate means in the NSW and observational control samples

	NSW		Full Comparison Samples		P-Score Screened Comparison Samples	
Variable	Treated (1)	Control (2)	CPS-1 (3)	CPS-3 (4)	CPS-1 (5)	CPS-3 (6)
Age	25.82	25.05	33.23	28.03	25.63	25.97
Years of schooling	10.35	10.09	12.03	10.24	10.49	10.42
Black	.84	.83	.07	.20	.96	.52
Hispanic	.06	.11	.07	.14	.03	.20
Dropout	.71	.83	.30	.60	.60	.63
Married	.19	.15	.71	.51	.26	.29
1974 earnings	2,096	2,107	14,017	5,619	2,821	2,969
1975 earnings	1,532	1,267	13,651	2,466	1,950	1,859
Number of obs.	185	260	15,992	429	352	157

Notes: Adapted from Dehejia and Wahba (1999), table 1. The samples in the first four columns are as described in Dehejia and Wahba (1999). The samples in the last two columns are limited to comparison group observations with a propensity score between .1 and .9. Propensity score estimates use all the covariates listed in the table.

some differences, particularly in terms of race and preprogram earnings.

Table 3.3.3 reports estimates of the NSW treatment effect. The dependent variable is annual earnings in 1978, a year or two after treatment. Rows of the table show results with alternative sets of controls: none; all the demographic variables in table 3.3.2; lagged (1975) earnings; demographics plus lagged earnings; demographics and two lags of earnings. All estimates are from regressions of 1978 earnings on a treatment dummy plus controls (the raw treatment-control difference appears in the first row).

Estimates using the experimental control group, reported in column 1, are on the order of $1,600–1,800. Not surprisingly, these estimates vary little across specifications. In contrast, the raw earnings gap between NSW participants and the CPS-1 sample, reported in column 2, is roughly −$8,500, suggesting this comparison is heavily contaminated by selection bias.

TABLE 3.3.3
Regression estimates of NSW training effects using alternative controls

	Full Comparison Samples			P-Score Screened Comparison Samples	
	NSW	CPS-1	CPS-3	CPS-1	CPS-3
Specification	(1)	(2)	(3)	(4)	(5)
Raw difference	1,794 (633)	−8,498 (712)	−635 (657)		
Demographic controls	1,670 (639)	−3,437 (710)	771 (837)	−3,361 (811) [139/497]	890 (884) [154/154]
1975 earnings	1,750 (632)	−78 (537)	−91 (641)	No obs. [0/0]	166 (644) [183/427]
Demographics, 1975 earnings	1,636 (638)	623 (558)	1,010 (822)	1,201 (722) [149/357]	1,050 (861) [157/162]
Demographics, 1974 and 1975 earnings	1,676 (639)	794 (548)	1,369 (809)	1,362 (708) [151/352]	649 (853) [147/157]

Notes: The table reports regression estimates of training effects using the Dehejia-Wahba (1999) data with alternative sets of controls. The demographic controls are age, years of schooling, and dummies for black, Hispanic, high school dropout, and married. Standard errors are reported in parentheses. Observation counts are reported in brackets [treated/control]. There are no observations with an estimated propersity score in the interval [.1, .9] using only 1975 earnings as a covariate with CPS-1 data.

The addition of demographic controls and lagged earnings narrows the gap considerably; the estimated treatment effect reaches (positive) $800 in the last row. The results are even better in column 3, which uses the narrower CPS-3 comparison group. The characteristics of this group are much closer to those of NSW participants; consistent with this, the raw earnings difference is only −$635. The fully controlled estimate, reported in the last row, is close to $1,400, not far from the experimental treatment effect.

A drawback of the process taking us from CPS-1 to CPS-3 is the ad hoc nature of the rules used to construct the smaller and more carefully selected CPS-3 comparison group. The CPS-3 selection criteria can be motivated by the NSW program rules, which favor individuals with low earnings and weak labor-force attachment, but in practice, there are many ways to implement this. We'd therefore like a more systematic approach to prescreening. In a recent paper, Crump, Hotz, Imbens, and Mitnik (2009) suggest that the propensity score be used for systematic sample selection as a precursor to regression estimation. This contrasts with our earlier discussion of the propensity score as the basis for an estimator.

We implemented the Crump et al. (2009) suggestion by first estimating the propensity score on a pooled NSW-treatment and observational-comparison sample, and then picking only those observations with $0.1 < p(X_i) < 0.9$. In other words, the estimation sample is limited to observations with a predicted probability of treatment equal to at least 10 percent but no more than 90 percent. This ensures that regressions are estimated in a sample including only covariate cells where there are at least a few treated and control observations. Estimation using screened samples therefore requires no extrapolation to cells without "common support"—in other words, to cells where there is no overlap in the covariate distribution between treatment and controls. Descriptive statistics for samples screened on the score (estimated using the full set of covariates listed in the table) appear in the last two columns of table 3.3.2. The covariate means in the screened CPS-1 and CPS-3 samples are much closer to the NSW means in column 1 than are the covariate means from unscreened samples.

We explored the common support screener further using alternative sets of covariates, but with the same covariates used for both screening and the estimation of treatment effects at each iteration. The resulting estimates are displayed in the final two columns of table 3.3.3. Controlling for demographic variables or lagged earnings alone, these results differ little from those in columns 2 and 3. With both demographic variables and a single lag of earnings as controls, however, the screened CPS-1 estimates are quite a bit closer to the experimental estimates than are the unscreened results. Screened CPS-1 estimates with two lags of earnings are also close to the experimental benchmark. On the other hand, the common support screener improves the CPS-3 results only slightly with a single lag of earnings and seems to be a step backward with two.

This investigation boosts our (already strong) faith in regression. Regression control for the right covariates does a reasonably good job of eliminating selection bias in the CPS-1 sample despite a huge baseline gap. Restricting the sample using our knowledge of program admissions criteria yields even better regression estimates with CPS-3, about as good as Dehejia and Wahba's (1999) propensity score matching results with two lags of earnings. Systematic prescreening to enforce common support seems like a useful adjunct to regression estimation with CPS-1, a large and coarsely selected initial sample. The estimates in screened CPS-1 are as good as unscreened CPS-3. We note, however, that the standard errors for estimates using propensity score–screened samples have not been adjusted to reflect the sampling variance in our estimates of the score. An advantage of prescreening using prior information, as in the step from CPS-1 to CPS-3, is that no such adjustment is necessary.

3.4 Regression Details

3.4.1 Weighting Regression

Few things are as confusing to applied researchers as the role of sample weights. Even now, 20 years post-Ph.D., we read the section of the Stata manual on weighting with some dismay. Weights can be used in a number of ways, and how they are used may well matter for your results. Regrettably, however, the case for or against weighting is often less than clear-cut, as are the specifics of how the weights should be programmed. A detailed discussion of weighting pros and cons is beyond the scope of this book. See Pfefferman (1993) and Deaton (1997) for two perspectives. In this brief subsection, we provide a few guidelines and a rationale for our approach to weighting.

A simple rule of thumb for weighting regression is to use weights when they make it more likely that the regression you are estimating is close to the population target you are trying to estimate. If, for example, the target (or estimand) is the population regression function, and the sample to be used for estimation is nonrandom with sampling weights, w_i, equal to the inverse probability of sampling observation i, then it makes sense to use weighted least squares, weighting by w_i (for this you can use Stata `pweights` or a SAS `weight` statement). Weighting by the inverse sampling probability generates estimates that are consistent for the population regression function even if the sample you have to work with is not a simple random sample.

A related weighting scenario involves grouped data. Suppose you would like to regress Y_i on X_i in a random sample, presumably because you want to learn about the population regression vector $\beta = E[\text{X}_i\text{X}_i']^{-1}E[\text{X}_i\,\text{Y}_i]$. Instead of a random sample, however, you have data grouped at the level of X_i. That is, you have estimates of $E[\,\text{Y}_i|\text{X}_i = x]$ for each x, estimated using data from a random sample. Let this average be denoted $\bar{y}_x$, and suppose you also know n_x, where n_x/N is the relative frequency of the value x in the underlying random sample. As we saw in section 3.1.2, the regression of $\bar{y}_x$ on x, weighted by n_x is the same as the random sample microdata regression. Therefore, if your goal is to get back to the microdata regression, it makes sense to weight by group size. We note, however, that macroeconomists, accustomed to working with published averages (like per capita income) and ignoring the underlying microdata, might disagree, or perhaps take the point in principle but remain disinclined to buck tradition in their discipline, which favors the unweighted analysis of aggregate variables.

On the other hand, if the sole rationale for weighting is heteroskedasticity, as in many textbook discussions of weighting, we are even less sympathetic to weighting than the macroeconomists. The argument for weighting under heteroskedasticity goes roughly like this: suppose you are interested in a linear CEF, $E[\text{Y}_i|\text{X}_i] = \text{X}_i'\beta$. The error term, defined as $e_i \equiv \text{Y}_i - \text{X}_i'\beta$, may be heteroskedastic. That is, the conditional variance function $E[e_i^2|\mathrm{X}_i]$ need not be constant. In this case, while the population regression function is still equal to $E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$, the sample analog is inefficient. A more precise estimator of the linear CEF is WLS—that is, the estimator that minimizes the sum of squared errors weighted by an estimate of $E[e_i^2|\mathrm{X}_i]^{-1}$.

As noted in section 3.1.3, an inherently heteroskedastic scenario is the LPM, where Y_i is a dummy variable. Assuming the CEF is in fact linear, as it will be if the model is saturated, then $P[\mathrm{Y}_i = 1|\mathrm{X}_i] = \mathrm{X}_i'\beta$ and therefore $E[e_i^2|\mathrm{X}_i] = \mathrm{X}_i'\beta(1 - \mathrm{X}_i'\beta)$, which is obviously a function of X_i. This is an example of model-based heteroskedasticity where estimates of the conditional variance function are easily constructed from estimates of the underlying regression function. The efficient WLS estimator for the LPM—a special case of generalized least squares (GLS)—is to weight by $[\mathrm{X}_i'\beta(1 - \mathrm{X}_i'\beta)]^{-1}$. Because the CEF has been assumed to be linear, these weights can be estimated in a first pass by OLS.

There are two reason why we prefer not to weight in this case (though we would use heteroskedasticity-consistent standard errors). First, in practice, the estimates of $E[e_i^2|\mathrm{X}_i]$ may not be very good. If the conditional variance model is a poor approximation or if the estimates of it are very noisy, WLS estimates may have worse finite-sample properties than unweighted estimates. The inferences you draw based on asymptotic theory may therefore be misleading, and the hoped-for efficiency gain may not materialize.[31] Second, if the CEF is not linear, the WLS estimator is no more likely to estimate it than is the unweighted estimator. On the other hand, the unweighted estimator still estimates something easy to interpret: the MMSE linear approximation to the population CEF.

WLS estimators also provide some sort of approximation, but the nature of this approximation depends on the weights. At a minimum, this makes it harder to compare your results to estimates reported by other researchers, and opens up additional avenues for specification searches when results depend

[31] Altonji and Segal (1996) discuss this point in a generalized method-of-moments context.

on weighting. Finally, an old caution comes to mind: if it ain't broke, don't fix it. The interpretation of the population regression vector is unaffected by heteroskedasticity, so why worry about it? Any efficiency gain from weighting is likely to be modest, and incorrect or poorly estimated weights can do more harm than good.

3.4.2 Limited Dependent Variables and Marginal Effects

Many empirical studies involve dependent variables that take on only a limited number of values. An example is the Angrist and Evans (1998) investigation of the effect of childbearing on female labor supply, also discussed in the chapter on instrumental variables. This study is concerned with the causal effects of childbearing on parents' work and earnings. Because childbearing is likely to be correlated with potential earnings, Angrist and Evans report instrumental variables estimates based on sibling-sex composition and multiple births, as well as OLS estimates. Almost every outcome in this study is either binary (e.g., employment status) or non-negative (e.g., hours worked, weeks worked, and earnings). Should the fact that a dependent variable is limited affect empirical practice? Many econometrics textbooks argue that, while OLS is fine for continuous dependent variables, when the outcome of interest is a limited dependent variable (LDV), linear regression models are inappropriate and nonlinear models such as probit and Tobit are preferred. In contrast, our view of regression as inheriting its legitimacy from the CEF makes LDVness less central.

As always, a useful benchmark is a randomized experiment, where regression generates a simple treatment-control difference. Consider, for example, regressions of various outcome variables on a randomly assigned regressor that indicates one of the treatment groups in the RAND Health Insurance Experiment (HIE; Manning et al. 1987). In this ambitious experiment, probably the most expensive in American social science, the RAND Corporation set up a small health insurance company that charged no premium. Nearly 6,000 participants in the study were randomly assigned to health insurance plans with different features.

One of the most important features of any insurance plan is the portion of health care costs the insured individual is expected to pay. The HIE randomly assigned individuals to many different plans. One plan provided entirely free care, while the others included various combinations of copayments, expenditure caps, and deductibles, so that enrollees paid for some of their health care costs out-of-pocket. The main purpose of the experiment was to learn whether the use of medical care is sensitive to cost and, if so, whether this affects health. The HIE results showed that those offered free or low-cost medical care used more of it but were not, for the most part, any healthier as a result. These findings helped pave the way for cost-sensitive health insurance plans and managed care.

Most of the outcomes in the HIE are LDVs. These include dummies indicating whether an experimental subject incurred any medical expenditures or was hospitalized in a given year, and non-negative outcomes such as the number of face-to-face doctor visits and gross annual medical expenses (whether paid by patient or insurer). The expenditure variable is zero for about 20 percent of the sample. Results for two of the HIE treatment groups are reproduced in table 3.4.1, derived from the estimates reported in table 2 of Manning et al. (1987). Table 3.4.1 shows average outcomes in the free care and individual deductible groups. The latter group faced a deductible of $150 per person or $450 per family per year for outpatient care, after which all costs were covered (there was no charge for inpatient care). The overall sample size in these two groups was a little over 3,000.

To simplify the LDV discussion, suppose that the comparison between free care and deductible plans is the only comparison of interest and that treatment was determined by simple random assignment.[32] Let $\text{D}_i = 1$ denote assignment to the deductible group. By virtue of random assignment, the

[32] The HIE was considerably more complicated than described here. There were 14 different treatments, including assignment to a prepaid HMO-like service. The experimental design did not use simple random assignment but rather a more complicated stratified assignment scheme meant to ensure covariate balance across groups.

TABLE 3.4.1
Average outcomes in two of the HIE treatment groups

Plan	Face-to-Face Visits	Outpatient Expenses (1984 $)	Admissions (%)	Prob. Any Medical (%)	Prob. Any Inpatient (%)	Total Expenses (1984 $)
Free	4.55	340	12.8	86.8	10.3	749
	(.17)	(10.9)	(.7)	(.8)	(.5)	(39)
Deductible	3.02	235	11.5	72.3	9.6	608
	(.17)	(11.9)	(.8)	(1.5)	(.6)	(46)
Deductible minus free	−1.53	−105	−1.3	−14.5	−0.7	−141
	(.24)	(16.1)	(1.0)	(1.7)	(.7)	(60)

Notes: Adapted from Manning et al. (1987), table 2. All standard errors (shown in parentheses) are corrected for intertemporal and intrafamily correlations. Amounts are in June 1984 dollars. Visits are face-to-face contacts with health providers; visits solely for radiology, anesthesiology, or pathology services are excluded. Visits and expenses exclude dental care and outpatient psychotherapy.

difference in means between those with $\mathrm{D}_i = 1$ and $\mathrm{D}_i = 0$ gives the unconditional average treatment effect. As in our earlier discussion of experiments (chapter 2):

$$\begin{aligned} &E[\mathrm{Y}_i|\mathrm{D}_i = 1] - E[\mathrm{Y}_i|\mathrm{D}_i = 0] \qquad (3.4.1) \\ &\quad = E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1] \\ &\quad = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] \end{aligned}$$

because D_i is independent of potential outcomes. Also, as before, $E[\mathrm{Y}_i|\mathrm{D}_i = 1] - E[\mathrm{Y}_i|\mathrm{D}_i = 0]$ is the slope coefficient in a regression of Y_i on D_i.

Equation (3.4.1) suggests that the estimation of causal effects in experiments presents no special challenges whether Y_i is binary, non-negative, or continuously distributed. Although the interpretation of the right-hand side changes for different sorts of dependent variables, you do not need to *do* anything special to get the average causal effect. For example, one of the HIE outcomes is a dummy denoting any medical expenditure. Since the outcome here is a Bernoulli trial, we have

$$\begin{aligned} E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] &= E[\mathrm{Y}_{1i}] - E[\mathrm{Y}_{0i}] \\ &= P[\mathrm{Y}_{1i} = 1] - P[\mathrm{Y}_{0i} = 1]. \qquad (3.4.2) \end{aligned}$$

This might affect the language we use to describe results, but not the underlying calculation. In the HIE, for example, comparisons across experimental groups, as on the left-hand side of (3.4.1), show that 87 percent of those assigned to the free-care group used at least some care in a given year, while only 72 percent of those assigned to the deductible plan used care. The relatively modest \$150 deductible therefore had a marked effect on use of care. The difference between these two rates, $-.15$ is an estimate of $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}]$, where Y_i is a dummy indicating any medical expenditure. Because the outcome here is a dummy variable, the average causal effect is also a causal effect on usage rates or probabilities.

Recognizing that the medical usage outcome variable is a probability, suppose instead that you use probit to fit the CEF in this case. No harm in trying! The probit model is usually motivated by the assumption that participation is determined by a latent variable, Y_i^*, that satisfies

$$\mathrm{Y}_i^* = \beta_0^* + \beta_1^* \mathrm{D}_i - \nu_i, \tag{3.4.3}$$

where ν_i is distributed $N(0, \sigma_\nu^2)$. Note that this latent variable cannot be actual medical expenditure since expenditure is nonnegative and therefore non-normal, while normally distributed random variables are continuously distributed on the real line and can therefore be negative. Given the latent index model,

$$\mathrm{Y}_i = 1[\mathrm{Y}_i^* > 0],$$

so the CEF for Y_i can be written

$$E[\mathrm{Y}_i|\mathrm{D}_i] = \Phi\left[\frac{\beta_0^* + \beta_1^* \mathrm{D}_i}{\sigma_\nu}\right],$$

where $\Phi[\cdot]$ is the normal CDF. Therefore

$$E[\mathrm{Y}_i|\mathrm{D}_i] = \Phi\left[\frac{\beta_0^*}{\sigma_\nu}\right] + \left\{\Phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_\nu}\right] - \Phi\left[\frac{\beta_0^*}{\sigma_\nu}\right]\right\} \mathrm{D}_i.$$

This is a linear function of the regressor, D_i, so the slope coefficient in the linear regression, of Y_i on D_i is just the difference in probit fitted values, $\Phi\big[\frac{\beta_0^*+\beta_1^*}{\sigma_\nu}\big]-\Phi\big[\frac{\beta_0^*}{\sigma_\nu}\big]$. But the probit coefficients, $\frac{\beta_0^*}{\sigma_\nu}$ and $\frac{\beta_1^*}{\sigma_\nu}$ do not give us the size of the effect of D_i on participation until we feed them back into the normal CDF (though they do have the right sign). Regression, in contrast, gives us what we need with or without the probit distributional assumptions.

One of the most important outcomes in the HIE is gross medical expenditure, in other words, health care costs. Did subjects who faced a deductible use less care, as measured by the cost? In the HIE, the average difference in expenditures between the deductible and free-care groups was −141 dollars, about 19 percent of the expenditure level in the free-care group. This calculation suggests that making patients pay a portion of costs reduces expenditures quite a bit, though the estimate is not very precise.

Because expenditure outcomes are non-negative random variables, and sometimes equal to zero, their expectation can be written

$$E[\text{Y}_i|\text{D}_i]=E[\text{Y}_i|\text{Y}_i>0,\text{D}_i]P[\text{Y}_i>0|\text{D}_i].$$

The difference in expenditure outcomes across treatment groups is

$$\begin{aligned}
&E[\text{Y}_i|\text{D}_i=1]-E[\text{Y}_i|\text{D}_i=0] \qquad (3.4.4)\\
&\quad=E[\text{Y}_i|\text{Y}_i>0,\text{D}_i=1]P[\text{Y}_i>0|\text{D}_i=1]\\
&\qquad-E[\text{Y}_i|\text{Y}_i>0,\text{D}_i=0]P[\text{Y}_i>0|\text{D}_i=0]\\
&\quad=\underbrace{\{P[\text{Y}_i>0|\text{D}_i=1]-P[\text{Y}_i>0|\text{D}_i=0]\}}_{\text{Participation effect}}E[\text{Y}_i|\text{Y}_i>0,\text{D}_i=1]\\
&\qquad+\underbrace{\{E[\text{Y}_i|\text{Y}_i>0,\text{D}_i=1]-E[\text{Y}_i|\text{Y}_i>0,\text{D}_i=0]\}}_{\text{COP effect}}\\
&\qquad\times P[\text{Y}_i>0|\text{D}_i=0].
\end{aligned}$$

So, the overall difference in average expenditure can be broken up into two parts: the difference in the probability that expenditures are positive (often called a participation effect) and the difference in means conditional on participation, a conditional-on-positive (COP) effect. Again, however, this has no special implications for the estimation of causal effects; equation (3.4.1) remains true: the regression of Y_i on D_i gives the unconditional average treatment effect for expenditures.

Good COP, Bad COP: Conditional-on-Positive Effects

Because causal effects on a non-negative random variable such as expenditure have two parts, some applied researchers feel they should look at these parts separately. In fact, many use a two-part model, in which the first part is an evaluation of the effect on participation and the second part looks at COP effects (see, e.g., Duan et al., 1983 and 1984, for such models applied to the HIE). The first part of (3.4.4) raises no special issues, because, as noted above, the fact that Y_i is a dummy means only that average treatment effects are also differences in probabilities. The problem with the two-part model is that the COP effects do not have a causal interpretation, even in a randomized trial. This complication can be understood as the same selection problem described in section 3.2.3, on bad control.

To analyze the COP effect further, write

$$\begin{aligned} &E[Y_i|Y_i > 0, D_i = 1] - E[Y_i|Y_i > 0, D_i = 0] \qquad (3.4.5)\\ &\quad = E[Y_{1i}|Y_{1i} > 0] - E[Y_{0i}|Y_{0i} > 0]\\ &\quad = \underbrace{E[Y_{1i} - Y_{0i}|Y_{1i} > 0]}_{\text{Causal effect}} + \underbrace{\{E[Y_{0i}|Y_{1i} > 0] - E[Y_{0i}|Y_{0i} > 0]\}}_{\text{Selection bias}}, \end{aligned}$$

where the second line uses the random assignment of D_i. This decomposition shows that the COP effect is composed of two terms: a causal effect for the subpopulation that uses medical care with a deductible and the difference in Y_{0i} between those who use medical care when they have to pay something and when it is free. This second term is a form of selection bias, though it is more subtle than the selection bias in chapter 2.

Here selection bias arises because the experiment changes the composition of the group with positive expenditures. The $Y_{0i} > 0$ population probably includes some low-cost users who would opt out of care if they had to pay a deductible. In other words, this group is larger and probably has lower costs on average than the $Y_{1i} > 0$ group. The selection bias term is therefore positive, with the result that COP effects are closer to zero than the presumably negative causal effect, $E[Y_{1i} - Y_{0i}|Y_{1i} > 0]$. This is a version of the bad control problem from section 3.2.3: in a causal effects setting, $Y_i > 0$ is an outcome variable and therefore unkosher for conditioning unless the treatment has no effect on the likelihood that Y_i is positive.

One resolution of the noncausality of COP effects relies on censored regression models like Tobit. These models postulate a latent expenditure outcome for nonparticipants (e.g., Hay and Olsen, 1984). A traditional Tobit formulation for the expenditure problem stipulates that the observed Y_i is generated by

$$Y_i = 1[Y_i^* > 0]Y_i^*,$$

where Y_i^* is a normally distributed latent expenditure variable that can take on negative values. Because Y_i^* is not an LDV, Tobit proponents feel comfortable linking this to D_i using a traditional linear model, say, equation (3.4.3). In this case, β_1^* is the causal effect of D_i on latent expenditure, Y_i^*. This equation is defined for everyone, whether Y_i is positive or not. There is no COP-style selection problem if we are happy to study effects on Y_i^*.

But we are not happy with effects on Y_i^*. The first problem is that "latent health care expenditure" is a puzzling construct. Health care expenditure really is zero for some people; this is not a statistical artifact or due to some kind of censoring. So the notion of latent and potentially negative Y_i^* is hard to grasp. There are no data on Y_i^* and there never will be. A second problem is that the link between the parameter β_1^* in the latent model and causal effects on the observed outcome, Y_i, turns on distributional assumptions about the latent variable. To establish this link we evaluate the expectation of Y_i given D_i to find

$$\begin{aligned} E[Y_i|D_i] = {} & \Phi\left[\frac{\beta_0^* + \beta_1^* D_i}{\sigma_\nu}\right][\beta_0^* + \beta_1^* D_i] \\ & + \sigma_\nu \phi\left[\frac{\beta_0^* + \beta_1^*\ D_i}{\sigma_\nu}\right], \end{aligned} \qquad (3.4.6)$$

(see, e.g., McDonald and Moffitt, 1980). This expression is derived using the normality and homoskedasticity of ν_i and the assumption that Y_i can be represented as $1[\text{Y}_i^* > 0]\text{Y}_i^*$.

The Tobit CEF provides us with an expression for the average treatment effect on observed expenditure. Specifically,

$$\begin{aligned} E[\text{Y}_i|\text{D}_i = 1] - E[\text{Y}_i|\text{D}_i = 0] \\ = \left\{\Phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_\nu}\right][\beta_0^* + \beta_1^*] + \sigma\phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_\nu}\right]\right\} \\ - \left\{\Phi\left[\frac{\beta_0}{\sigma_\nu}\right][\beta_0^*] + \sigma_\nu\phi\left[\frac{\beta_0^*}{\sigma_\nu}\right]\right\} \end{aligned} \quad (3.4.7)$$

a rather daunting formula. But since the only regressor is a dummy variable, D_i, none of this is necessary for the estimation of $E[\text{Y}_i|\text{D}_i = 1] - E[\text{Y}_i|\text{D}_i = 0]$. The slope coefficient from an OLS regression of Y_i on D_i recovers the CEF difference on the left-hand side of (3.4.7) whether or not you adopt a Tobit model to explain the underlying structure.[33]

COP effects are sometimes motivated by a researcher's sense that when the outcome distribution has a mass point—that is, when it piles up on a particular value, such as zero—or has a heavily skewed distribution, or both, then an analysis of effects on averages misses something. Analyses of effects on averages indeed miss some things, such as changes in the probability of specific values or a shift in quantiles away from the median. But why not look at these distribution effects directly? Distribution outcomes include the likelihood that annual medical expenditures exceed zero, 100 dollars, 200 dollars, and so on. In other words, put $1[\text{Y}_i > c]$ for different choices of c on the left-hand side of the regression of interest. Econometrically, these outcomes are all in the category of (3.4.2). The idea of looking directly at distribution effects with linear probability models is illustrated by Angrist (2001), in an analysis of the effects of childbearing on hours worked. Alternatively, if

[33] A generalization of Tobit is the sample selection model, where the latent variable determining participation differs from the latent expenditure variable. See, for example, Maddala (1983). The same conceptual problems related to the interpretation of effects on latent variables arise in the sample selection model as with Tobit.

quantiles provide a focal point, we can use quantile regression to model them. Chapter 7 discusses this idea in detail.

Do Tobit-type latent variable models ever make sense? Yes, if the data you are working with are truly censored. True censoring means the latent variable has an empirical counterpart that is the outcome of primary interest. A leading example from labor economics is CPS earnings data, which topcodes (censors) very high values of earnings to protect respondent confidentiality. Typically, we're interested in the causal effect of schooling on earnings as it appears on respondents' tax returns, not their CPS-topcoded earnings. Chamberlain (1994) shows that in some years, CPS topcoding reduces the measured returns to schooling considerably, and proposes an adjustment for censoring based on a Tobit-style adaptation of quantile regression. The use of quantile regression to model censored data is also discussed in chapter 7.[34]

Covariates Lead to Nonlinearity

True censoring as with the CPS topcode is rare, a fact that leaves limited scope for constructive applications of Tobit-type models in applied work. At this point, however, we have to hedge a bit. Part of the neatness in the discussion of experiments comes from the fact that $E[\textsc{y}_i|\textsc{d}_i]$ is necessarily a linear function of $\textsc{d}_i$, so that regression and the CEF are one and the same. In fact, this CEF is linear for any function of $\textsc{y}_i$, including the distribution indicators, $1[\textsc{y}_i > c]$. In practice, of course, the explanatory variable of interest isn't always a dummy, and there are usually additional covariates in the CEF, in which case $E[\textsc{y}_i|\mathrm{X}_i,\textsc{d}_i]$ for LDVs is almost certainly nonlinear. Intuitively, as predicted means get close to the dependent variable boundaries, the derivatives of the CEF

[34] We should note that our favorite regression example, a regression of log wages on schooling, may have a COP problem since the sample of log wages naturally omits those with zero earnings. This leads to COP-style selection bias if education affects the probability of working. In practice, therefore, we focus on samples of prime-age males, whose participation rates are high and reasonably stable across schooling groups (e.g., white men aged 40–49 in figure 3.1.1).

for LDVs get smaller (think, for example, of how the normal CDF flattens at extreme values).

The upshot is that in LDV models with covariates, regression need not fit the CEF perfectly. It remains true, however, that the underlying CEF has a causal interpretation if the CIA holds. And if the CEF has a causal interpretation, it seems fair to say that regression has a causal interpretation as well, because it still provides the MMSE approximation to the CEF. Moreover, if the model for covariates is saturated, then regression also estimates a weighted average treatment effect similar to (3.3.1) and (3.3.3). Likewise, if the regressor of interest is multivalued or continuous, we get a weighted average derivative, as described by the formulas at the end of subsection 3.3.1.

And yet, we may not have enough data for the saturated-covariate regression specification to be very attractive. Regression will therefore miss some features of the CEF. For one thing, it may generate fitted values outside the LDV boundaries. This fact bothers some researchers and has generated a lot of bad press for the linear probability model. One attractive feature of nonlinear models like probit and Tobit is that they produce CEFs that respect LDV boundaries. In particular, probit fitted values are always between zero and one, while Tobit fitted values are positive (this is not obvious from equation (3.4.6)). We might therefore prefer nonlinear models on simple curve-fitting grounds.

Point conceded. It's important to emphasize, however, that the output from nonlinear models must be converted into *marginal effects* to be useful. Marginal effects are the (average) changes in CEF implied by a nonlinear model. Without marginal effects, it's hard to talk about the impact on observed dependent variables. If we continue to assume the regressor of interest is the dummy variable, D_i, marginal effects can be constructed either by differencing

$$E\{E[\text{Y}_i|X_i, \text{D}_i = 1] - E[\text{Y}_i|X_i, \text{D}_i = 0]\},$$

or, by differentiation, $E\left\{\frac{\partial E[\text{Y}_i|X_i,\text{D}_i]}{\partial \text{D}_i}\right\}$. Most people use derivatives when dealing with continuous or multivalued regressors.

How close do OLS regression estimates come to the marginal effects induced by a nonlinear model like probit or Tobit? We first derive the marginal effects, and then show an empirical example. The probit CEF for a model with covariates is

$$E[\text{Y}_i|\text{X}_i, \text{D}_i] = \Phi\left[\frac{\text{X}_i'\beta_0^* + \beta_1^*\text{D}_i}{\sigma_\nu}\right].$$

The average finite difference is therefore

$$E\left\{\Phi\left[\frac{\text{X}_i'\beta_0^* + \beta_1^*}{\sigma_\nu}\right] - \Phi\left[\frac{\text{X}_i'\beta_0^*}{\sigma_\nu}\right]\right\}. \tag{3.4.8}$$

In practice, this can be approximated by the average derivative,

$$E\left\{\phi\left[\frac{\text{X}_i'\beta_0^* + \beta_1^*\text{D}_i}{\sigma_\nu}\right]\right\} \cdot \left(\frac{\beta_1^*}{\sigma_\nu}\right)$$

(Stata computes marginal effects both ways but defaults to (3.4.8) for dummy regressors).

Similarly, generalizing equation (3.4.6) to a model with covariates, we have

$$\begin{aligned} E[\text{Y}_i|\text{X}_i, \text{D}_i] = {} & \Phi\left[\frac{\text{X}_i'\beta_0^* + \beta_1^*\text{D}_i}{\sigma_\nu}\right][\text{X}_i'\beta_0^* + \beta_1^*\text{D}_i] \\ & + \sigma_\nu\phi\left[\frac{\text{X}_i'\beta_0^* + \beta_1^*\text{D}_i}{\sigma_\nu}\right] \end{aligned}$$

for a non-negative LDV. Tobit marginal effects are almost always cast in terms of the average derivative, which can be shown to be the surprisingly simple expression

$$E\left\{\Phi\left[\frac{\text{X}_i'\beta_0^* + \beta_1^*\text{D}_i}{\sigma_\nu}\right]\right\} \cdot \beta_1^* \tag{3.4.9}$$

(see, e.g., Wooldridge, 2006). One immediate implication of (3.4.9) is that the Tobit coefficient, β_1^*, is always too big relative to the effect of D_i on Y_i. Intuitively, this is because, given the linear model for latent Y_i^*, the latent outcome always changes when D_i switches on or off. But real Y_i need not change: for many people, it's zero either way.

Table 3.4.2 compares OLS estimates and nonlinear marginal effects for regressions of female employment and hours of work, both LDVs, on measures of fertility. These estimates were constructed using one of the 1980 census samples used by Angrist and Evans (1998). This sample includes married women aged 21–35 with at least two children. The childbearing variables consist of a dummy indicating women with more than two children or the total number of births. The covariates include linear terms in mother's age, age at first birth, race dummies (black and Hispanic), and mother's education (dummies for high school graduates, some college, and college graduates). The covariate model is not saturated; rather, there are additive terms and no interactions, though the underlying CEF in this example is surely nonlinear.

Probit marginal effects for the impact of a dummy variable indicating more than two children are indistinguishable from OLS estimates of the same relation. This can be seen in columns 2, 3, and 4 of table 3.4.2, the first row of which compares the estimates from different methods for the full 1980 sample. The OLS estimate of the effect of a third child is $-.162$, while the corresponding probit marginal effects are $-.163$ and $-.162$. These were estimated using (3.4.8) in the first case and

$$E\left\{\Phi\left[\frac{X_i'\beta_0^* + \beta_1^*}{\sigma_\nu}\right] - \Phi\left[\frac{X_i'\beta_0^*}{\sigma_\nu}\right] \middle| \, \mathrm{D}_i = 1\right\}$$

in the second (hence, a marginal effect on the treated).

Tobit marginal effects for the relation between fertility and hours worked are quite close to the corresponding OLS estimates, though not indistinguishable. This can be seen in columns 5 and 6. Compare, for example, the Tobit estimates of -6.56 and -5.87 with the OLS estimate of -5.92 in column 2. Although one Tobit estimate is 10 percent larger in absolute value, this seems unlikely to be of substantive importance. The remaining columns of the table compare OLS estimates to marginal effects for an ordinal childbearing variable instead of a dummy. These calculations all use derivatives to compute marginal effects (labeled MFX). Here, too, the OLS and

Table 3.4.2
Comparison of alternative estimates of the effect of childbearing on LDVs

		Right-Hand-Side Variable								
		More than Two Children					Number of Children			
			Probit		Tobit			Probit MFX	Tobit MFX	
Dependent Variable	Mean (1)	OLS (2)	Avg. Effect, Full Sample (3)	Avg. Effect on Treated (4)	Avg. Effect, Full Sample (5)	Avg. Effect on Treated (6)	OLS (7)	Avg. Effect, Full Sample (8)	Avg. Effect, Full Sample (9)	Avg. Effect on Treated (10)
A. Full sample										
Employment	.528	−.162	−.163	−.162	—	—	−.113	−.114	—	—
	(.499)	(.002)	(.002)	(.002)			(.001)	(.001)		
Hours worked	16.7	−5.92	—	—	−6.56	−5.87	−4.07	—	−4.66	−4.23
	(18.3)	(.074)			(.081)	(.073)	(.047)		(.054)	(.049)
B. Nonwhite college attenders over age 30, first birth before age 20										
Employment	.832	−.061	−.064	−.070	—	—	−.054	−.048	—	—
	(.374)	(.028)	(.028)	(.031)			(.016)	(.013)		
Hours worked	30.8	−4.69	—	—	−4.97	−4.90	−2.83	—	−3.20	−3.15
	(16.0)	(1.18)			(1.33)	(1.31)	(.645)		(.670)	(.659)

Notes: The table reports OLS estimates, average treatment effects, and marginal effects (MFX) for the effect of childbearing on mothers' labor supply. The sample in panel A includes 254,654 observations and is the same as the 1980 census sample of married women used by Angrist and Evans (1998). Covariates include age, age at first birth, and dummies for boys at first and second birth. The sample in panel B includes 746 nonwhite women with at least some college aged over 30 whose first birth was before age 20. Standard deviations are reported in parentheses in column 1. Standard errors are shown in parentheses in other columns. The sample used to estimate average effects on the treated in columns 4, 6, and 10 includes women with more than two children.

nonlinear marginal effects estimates are similar for both probit and Tobit.

It is sometimes said that probit models can be expected to generate marginal effects close to OLS when the predicted probabilities are close to .5 because the underlying nonlinear CEF is roughly linear in the middle. With predictions close to zero or one, however, we might expect a larger gap. We therefore replicated the comparison of OLS and marginal effects in a subsample with relatively high average employment rates, nonwhite women over age 30 who attended college and whose first birth was before age 20. Although the average employment rate is 83 percent in this group, the OLS estimates and marginal effects are again similar.

The upshot of this discussion is that while a nonlinear model may fit the CEF for LDVs more closely than a linear model, when it comes to marginal effects, this probably matters little. This optimistic conclusion is not a theorem, but, as in the empirical example here, it seems to be fairly robustly true.

Why, then, should we bother with nonlinear models and marginal effects? One answer is that the marginal effects are easy enough to compute now that they are automated in packages like Stata. But there are a number of decisions to make along the way (e.g., the weighting scheme, derivatives versus finite differences), while OLS is standardized. Nonlinear life also gets considerably more complicated when we work with instrumental variables and panel data. Finally, extra complexity comes into the inference step as well, since we need standard errors for marginal effects. The principle of Occam's razor advises, "Entities should not be multiplied unnecessarily." In this spirit, we quote our former teacher, Angus Deaton (1997), pondering the nonlinear regression function generated by Tobit-type models:

> Absent knowledge of *F* [the distribution of the errors], this regression function does not even identify the β's [Tobit coefficients]—see Powell (1989)—but more fundamentally, we should ask how it has come about that we have to deal with such an awkward, difficult, and non-robust object.

3.4.3 Why Is Regression Called Regression, and What Does Regression to the Mean Mean?

The term *regression* originates with Francis Galton's (1886) study of height. Galton, who is pictured visiting his tailor on page 26, worked with samples of roughly normally distributed data on parents and children. He noted that the CEF of a child's height conditional on his parents' height is linear, with parameters given by the bivariate regression slope and intercept. Since height is stationary (its distribution does not change much over time), the bivariate regression slope is also the correlation coefficient, that is, between zero and one.

The single regressor in Galton's setup, x_i, is average parent height and the dependent variable, Y_i, is the height of adult children. The regression slope coefficient, as always, is $\beta_1 = \frac{Cov(\text{Y}_i, x_i)}{V(x_i)}$, and the intercept is $\alpha = E[\text{Y}_i] - \beta_1 E[\text{X}_i]$. But because height is not changing across generations, the mean and variance of Y_i and x_i are the same. Therefore,

$$\begin{aligned}\beta_1 &= \frac{Cov(\text{Y}_i, x_i)}{V(x_i)} = \frac{Cov(\text{Y}_i, x_i)}{\sqrt{V(x_i)}\sqrt{V(\text{Y}_i)}} = \rho_{xy}\\ \alpha &= E[\text{Y}_i] - \beta_1 E[\text{X}_i] = \mu(1-\beta_1) = \mu(1-\rho_{xy}),\end{aligned}$$

where ρ_{xy} is the intergenerational correlation coefficient in height and $\mu = E[\text{Y}_i] = E[\text{X}_i]$ is the population average height. From this we get the linear CEF

$$E[\text{Y}_i | x_i] = \mu(1-\rho_{xy}) + \rho_{xy} x_i,$$

so the height of a child given his parents' height is a weighted average of his parents' height and the population average height. The child of tall parents will therefore not be as tall as they are, on average. Likewise, for the short. To be specific, Pischke, who is six feet three inches tall, can expect his children to be tall, though not as tall as he is. Thankfully, however, Angrist, who is five feet six inches tall, can expect his children to be taller than he is. Galton called this property "regression toward mediocrity in hereditary stature." Today we call it regression to the mean.

Galton, who was Charles Darwin's cousin, is also remembered for having founded the Eugenics Society, dedicated to breeding better people. Indeed, his interest in regression came largely from this quest. We conclude from this that the value of scientific ideas should not be judged by their author's politics.

Galton does not seem to have shown much interest in multiple regression, our chief concern in this chapter. The regressions in Galton's work are mechanical features of distributions of stationary random variables; they work just as well for the regression of parents' height on childrens' height and are certainly not causal. Galton would have said so himself, because he objected to the Lamarckian idea (later promoted in Stalin's Russia) that acquired traits can be inherited.

The idea that regression can be used for statistical control in pursuit of causality satisfyingly originates in an inquiry into the determinants of poverty rates by George Udny Yule (1899). Yule, a statistician and student of Karl Pearson (Pearson was Galton's protégé), realized that Galton's regression coefficient could be extended to multiple variables by solving the least squares normal equations that had been derived long before by Legendre and Gauss. Yule's (1899) paper appears to be the first publication containing multivariate regression estimates. His model links changes in poverty rates in an area to changes in the local administration of the English Poor Laws, while controlling for population growth and the age distribution in the area. He was particularly interested in whether out-relief, the practice of providing income support for poor people without requiring them to move to the poorhouse, did not itself contribute to higher poverty rates. This is a well-defined causal question of a sort that still occupies us today.[35]

Finally, we note that the history of regression is beautifully detailed in the book by Steven Stigler (1986). Stigler is a famous statistician at the University of Chicago, but not quite

[35] Yule's first applied paper on the poor laws was published in 1895 in the *Economic Journal*, where Pischke is proud to serve as co-editor. The theory of multiple regression that goes along with this appears in Yule (1897).

as famous as his father, the economist and Nobel laureate, George Stigler.

3.5 Appendix: Derivation of the Average Derivative Weighting Function

Begin with the regression of Y_i on S_i:

$$\frac{Cov(\text{Y}_i, \text{S}_i)}{V(\text{S}_i)} = \frac{E[h(\text{S}_i)(\text{S}_i - E[\text{S}_i])]}{E[\text{S}_i(\text{S}_i - E[\text{S}_i])]}.$$

Let $\kappa_{-\infty} = \lim_{t \to -\infty} h(t)$, which we assume exists. By the fundamental theorem of calculus, we have:

$$h(\text{S}_i) = \kappa_{-\infty} + \int_{-\infty}^{\text{S}_i} h'(t)dt.$$

Substituting for $h(\text{S}_i)$, the numerator becomes

$$E[h(\text{S}_i)(\text{S}_i - E[\text{S}_i])] = \int_{-\infty}^{+\infty} \int_{-\infty}^{u} h'(t)(u - E[\text{S}_i])g(u)dtdu,$$

where $g(u)$ is the density of S_i at u. Reversing the order of integration, we have

$$E[h(\text{S}_i)(\text{S}_i - E[\text{S}_i])] = \int_{-\infty}^{+\infty} h'(t) \int_{t}^{+\infty} (u - E[\text{S}_i])g(u)dudt.$$

The inner integral is equal to $\mu_t \equiv \{E[\text{S}_i|\text{S}_i \geq t] - E[\text{S}_i|\text{S}_i < t]\}$ $\{P(\text{S}_i \geq t)[1 - P(\text{S}_i \geq t)\}$, the weighting function in (3.3.9), which is clearly non-negative. Setting $\text{S}_i = \text{Y}_i$, the denominator can similarly be shown to be the integral of these weights. We therefore have a weighted average derivative representation of the bivariate regression coefficient, $\frac{Cov(\text{Y}_i, \text{S}_i)}{V(\text{S}_i)}$. A similar formula for a regression with covariates is derived in the appendix to Angrist and Krueger (1999).

This page intentionally left blank