Chapter 2

Regression

KWAI CHANG CAINE: A worker is known by his tools. A shovel for a man who digs. An ax for a woodsman. The econometrician runs regressions.

Kung Fu, Season 1, Episode 8

Our Path

When the path to random assignment is blocked, we look for alternate routes to causal knowledge. Wielded skillfully, 'metrics tools other than random assignment can have much of the causality-revealing power of a real experiment. The most basic of these tools is *regression,* which compares treatment and control subjects who have the same observed characteristics. Regression concepts are foundational, paving the way for the more elaborate tools used in the chapters that follow. Regression-based causal inference is predicated on the assumption that when key observed variables have been made equal across treatment and control groups, selection bias from the things we can't see is also mostly eliminated. We illustrate this idea with an empirical investigation of the economic returns to attendance at elite private colleges.

2.1 A Tale of Two Colleges

Students who attended a private four-year college in America paid an average of about $29,000 in tuition and fees in the 2012–2013 school year. Those who went to a public university in their home state paid less than $9,000. An elite private education might be better in many ways: the classes smaller, the athletic facilities newer, the faculty more distinguished, and the students smarter. But $20,000 per year of study is a big difference. It makes you wonder whether the difference is worth it.

The apples-to-apples question in this case asks how much a 40-year-old Massachusetts-born graduate of, say, Harvard, would have earned if he or she had gone to the University of Massachusetts (U-Mass) instead. Money isn't everything, but, as Groucho Marx observed: "Money frees you from doing things you dislike. Since I dislike doing nearly everything, money is handy." So when we ask whether the private school tuition premium is worth paying, we focus on the possible earnings gain enjoyed by those who attend elite private universities. Higher earnings aren't the only reason you might prefer an elite private institution over your local state school. Many college students meet a future spouse and make lasting friendships while in college. Still, when families invest an additional $100,000 or more in human capital, a higher anticipated earnings payoff seems likely to be part of the story.

Comparisons of earnings between those who attend different sorts of schools invariably reveal large gaps in favor of elite-college alumni. Thinking this through, however, it's easy to see why comparisons of the earnings of students who attended Harvard and U-Mass are unlikely to reveal the payoff to a Harvard degree. This comparison reflects the fact that Harvard grads typically have better high school grades and higher SAT scores, are more motivated, and perhaps have other skills and talents. No disrespect intended for the many good students who go to U-Mass, but it's damn hard to get into Harvard, and those who do are a special and select group. In contrast, U-Mass accepts and even awards scholarship money to almost every Massachusetts applicant with decent tenth-grade test scores. We should therefore expect earnings comparisons across alma maters to be contaminated by selection bias, just like the comparisons of health by insurance status discussed in the

previous chapter. We've also seen that this sort of selection bias is eliminated by random assignment. Regrettably, the Harvard admissions office is not yet prepared to turn their admissions decisions over to a random number generator.

The question of whether college selectivity matters must be answered using the data generated by the routine application, admission, and matriculation decisions made by students and universities of various types. Can we use these data to mimic the randomized trial we'd like to run in this context? Not to perfection, surely, but we may be able to come close. The key to this undertaking is the fact that many decisions and choices, including those related to college attendance, involve a certain amount of serendipitous variation generated by financial considerations, personal circumstances, and timing.

Serendipity can be exploited in a sample of applicants on the cusp, who could easily go one way or the other. Does anyone admitted to Harvard really go to their local state school instead? Our friend and former MIT PhD student, Nancy, did just that. Nancy grew up in Texas, so the University of Texas (UT) was her state school. UT's flagship Austin campus is rated "Highly Competitive" in Barron's rankings, but it's not Harvard. UT is, however, much less expensive than Harvard (*The Princeton Review* recently named UT Austin a "Best Value College"). Admitted to both Harvard and UT, Nancy chose UT over Harvard because the UT admissions office, anxious to boost average SAT scores on campus, offered Nancy and a few other outstanding applicants an especially generous financial aid package, which Nancy gladly accepted.

What are the consequences of Nancy's decision to accept UT's offer and decline Harvard's? Things worked out pretty well for Nancy in spite of her choice of UT over Harvard: today she's an economics professor at another Ivy League school in New England. But that's only one example. Well, actually, it's two: Our friend Mandy got her bachelor's from the University of Virginia, her home state school, declining offers from Duke, Harvard, Princeton, and Stanford. Today, Mandy teaches at Harvard.

A sample of two is still too small for reliable causal inference. We'd

like to compare many people like Mandy and Nancy to many other similar people who chose private colleges and universities. From larger group comparisons, we can hope to draw general lessons. Access to a large sample is not enough, however. The first and most important step in our effort to isolate the serendipitous component of school choice is to hold constant the most obvious and important differences between students who go to private and state schools. In this manner, we hope (though cannot promise) to make *other things equal.*

Here's a small-sample numerical example to illustrate the *ceteris paribus* idea (we'll have more data when the time comes for real empirical work). Suppose the only things that matter in life, at least as far as your earnings go, are your SAT scores and where you go to school. Consider Uma and Harvey, both of whom have a combined reading and math score of 1,400 on the SAT.[1] Uma went to U-Mass, while Harvey went to Harvard. We start by comparing Uma's and Harvey's earnings. Because we've assumed that all that matters for earnings besides college choice is the combined SAT score, Uma vs. Harvey is a *ceteris paribus* comparison.

In practice, of course, life is more complicated. This simple example suggests one significant complication: Uma is a young woman, and Harvey is a young man. Women with similar educational qualifications often earn less than men, perhaps due to discrimination or time spent out of the labor market to have children. The fact that Harvey earns 20% more than Uma may be the effect of a superior Harvard education, but it might just as well reflect a male-female wage gap generated by other things.

We'd like to disentangle the pure Harvard effect from these other things. This is easy if the only other thing that matters is gender: replace Harvey with a female Harvard student, Hannah, who also has a combined SAT of 1,400, comparing Uma and Hannah. Finally, because we're after general conclusions that go beyond individual stories, we look for many similar same-sex and same-SAT contrasts across the two schools. That is, we compute the average earnings difference among Harvard and U-Mass students with the same gender and SAT score. The

average of all such group-specific Harvard versus U-Mass differences is our first shot at estimating the causal effect of a Harvard education. This is an econometric *matching* estimator that *controls for*—that is, holds fixed—sex and SAT scores. Assuming that, conditional on sex and SAT scores, the students who attend Harvard and U-Mass have similar earnings potential, this estimator captures the average causal effect of a Harvard degree on earnings.

Matchmaker, Matchmaker

Alas, there's more to earnings than sex, schools, and SAT scores. Since college attendance decisions aren't randomly assigned, we must control for *all* factors that determine both attendance decisions and later earnings. These factors include student characteristics, like writing ability, diligence, family connections, and more. Control for such a wide range of factors seems daunting: the possibilities are virtually infinite, and many characteristics are hard to quantify. But Stacy Berg Dale and Alan Krueger came up with a clever and compelling shortcut.[2] Instead of identifying everything that might matter for college choice and earnings, they work with a key summary measure: the characteristics of colleges to which students applied and were admitted.

Consider again the tale of Uma and Harvey: both applied to, and were admitted to, U-Mass and Harvard. The fact that Uma applied to Harvard suggests she has the motivation to go there, while her admission to Harvard suggests she has the ability to succeed there, just like Harvey. At least that's what the Harvard admissions office thinks, and they are not easily fooled.[3] Uma nevertheless opts for a cheaper U-Mass education. Her choice might be attributable to factors that are not closely related to Uma's earnings potential, such as a successful uncle who went to U-Mass, a best friend who chose U-Mass, or the fact that Uma missed the deadline for that easily won Rotary Club scholarship that would have funded an Ivy League education. If such serendipitous events were decisive for Uma and Harvey, then the two of them make a good match.

Dale and Krueger analyzed a large data set called College and Beyond (C&B). The C&B data set contains information on thousands of students who enrolled in a group of moderately to highly selective U.S. colleges and universities, together with survey information collected from the students at the time they took the SAT, about a year before college entry, and information collected in 1996, long after most had graduated from college. The analysis here focuses on students who enrolled in 1976 and who were working in 1995 (most adult college graduates are working). The colleges include prestigious private universities, like the University of Pennsylvania, Princeton, and Yale; a number of smaller private colleges, like Swarthmore, Williams, and Oberlin; and four public universities (Michigan, The University of North Carolina, Penn State, and Miami University in Ohio). The average (1978) SAT scores at these schools ranged from a low of 1,020 at Tulane to a high of 1,370 at Bryn Mawr. In 1976, tuition rates were as low as $540 at the University of North Carolina and as high as $3,850 at Tufts (those were the days).

Table 2.1 details a stripped-down version of the Dale and Krueger matching strategy, in a setup we call the "college matching matrix." This table lists applications, admissions, and matriculation decisions for a (made-up) list of nine students, each of whom applied to as many as three schools chosen from an imaginary list of six. Three out of the six schools listed in the table are public (All State, Tall State, and Altered State) and three are private (Ivy, Leafy, and Smart). Five of our nine students (numbers 1, 2, 4, 6, and 7) attended private schools. Average earnings in this group are $92,000. The other four, with average earnings of $72,500, went to a public school. The almost $20,000 gap between these two groups suggests a large private school advantage.

TABLE 2.1

The college matching matrix

Applicant group	Student	Private			Public			1996 earnings
		Ivy	Leafy	Smart	All State	Tall State	Altered State	
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

The students in Table 2.1 are organized in four groups defined by the set of schools to which they applied and were admitted. Within each group, students are likely to have similar career ambitions, while they were also judged to be of similar ability by admissions staff at the schools to which they applied. Within-group comparisons should therefore be considerably more apples-to-apples than uncontrolled comparisons involving all students.

The three group A students applied to two private schools, Leafy and Smart, and one public school, Tall State. Although these students were rejected at Leafy, they were admitted to Smart and Tall State. Students 1 and 2 went to Smart, while student 3 opted for Tall State. The students in group A have high earnings, and probably come from upper middle class families (a signal here is that they applied to more private schools than public). Student 3, though admitted to Smart, opted for cheaper Tall State, perhaps to save her family money (like our friends Nancy and Mandy). Although the students in group A have done well, with high average earnings and a high rate of private school attendance, within group A, the private school differential is negative: $(110 + 100)/2 - 110 = -5$, in other words, a gap of $-\$5{,}000$.

The comparison in group A is one of a number of possible matched

comparisons in the table. Group B includes two students, each of whom applied to one private and two public schools (Ivy, All State, and Altered State). The students in group B have lower average earnings than those in group A. Both were admitted to all three schools to which they applied. Number 4 enrolled at Ivy, while number 5 chose Altered State. The earnings differential here is \$30,000 (60 − 30 = 30). This gap suggests a substantial private school advantage.

Group C includes two students who applied to a single school (Leafy), where they were admitted and enrolled. Group C earnings reveal nothing about the effects of private school attendance, because both students in this group attended private school. The two students in group D applied to three schools, were admitted to two, and made different choices. But these two students chose All State and Tall State, both public schools, so their earnings also reveal nothing about the value of a private education. Groups C and D are uninformative, because, from the perspective of our effort to estimate a private school treatment effect, each is composed of either all-treated or all-control individuals.

Groups A and B are where the action is in our example, since these groups include public and private school students who applied to and were admitted to the same set of schools. To generate a single estimate that uses all available data, we average the group-specific estimates. The average of −\$5,000 for group A and \$30,000 for group B is \$12,500. This is a good estimate of the effect of private school attendance on average earnings, because, to a large degree, it controls for applicants' choices and abilities.

The simple average of treatment-control differences in groups A and B isn't the only well-controlled comparison that can be computed from these two groups. For example, we might construct a weighted average which reflects the fact that group B includes two students and group A includes three. The weighted average in this case is calculated as

$$\left(\frac{3}{5} \times -5{,}000\right) + \left(\frac{2}{5} \times 30{,}000\right) = 9{,}000.$$

By emphasizing larger groups, this weighting scheme uses the data more efficiently and may therefore generate a statistically more precise summary of the private-public earnings differential.

The most important point in this context is the apples-to-apples and oranges-to-oranges nature of the underlying matched comparisons. Apples in group A are compared to other group A apples, while oranges in group B are compared only with oranges. In contrast, naive comparisons that simply compare the earnings of private and public school students generate a much larger gap of $19,500 when computed using all nine students in the table. Even when limited to the five students in groups A and B, the uncontrolled comparison generates a gap of $20,000 ($20 = (110 + 100 + 60)/3 - (110 + 30)/2$). These much larger uncontrolled comparisons reflect selection bias: students who apply to and are admitted to private schools have higher earnings wherever they ultimately chose to go.

Evidence of selection bias emerges from a comparison of average earnings across (instead of within) groups A and B. Average earnings in group A, where two-thirds apply to private schools, are around $107,000. Average earnings in group B, where two-thirds apply to public schools, are only $45,000. Our within-group comparisons reveal that much of this shortfall is unrelated to students' college attendance decisions. Rather, the cross-group differential is explained by a combination of ambition and ability, as reflected in application decisions and the set of schools to which students were admitted.

2.2 Make Me a Match, Run Me a Regression

Regression is the tool that masters pick up first, if only to provide a benchmark for more elaborate empirical strategies. Although regression is a many-splendored thing, we think of it as an automated matchmaker. Specifically, regression estimates are weighted averages of multiple matched comparisons of the sort constructed for the groups in our stylized matching matrix (the appendix to this chapter discusses a

closely related connection between regression and mathematical expectation).

The key ingredients in the regression recipe are

- the *dependent variable,* in this case, student *i*'s earnings later in life, also called the *outcome variable* (denoted by Y_i);
- the *treatment variable,* in this case, a dummy variable that indicates students who attended a private college or university (denoted by P_i); and
- a set of *control variables,* in this case, variables that identify sets of schools to which students applied and were admitted.

In our matching matrix, the five students in groups A and B (Table 2.1) contribute useful data, while students in groups C and D can be discarded. In a data set containing those left after discarding groups C and D, a single variable indicating the students in group A tells us which of the two groups the remaining students are in, because those not in group A are in group B. This variable, which we'll call A_i, is our sole control. Note that both P_i and A_i are dummy variables, that is, they equal 1 to indicate observations in a specific state or condition, and 0

otherwise. Dummies, as they are called (no reference to ability here), classify data into simple yes-or-no categories. Even so, by coding many dummies, we get a set of control variables that's as detailed as we like.[4]

The regression model in this context is an equation linking the treatment variable to the dependent variable while holding control variables fixed by including them in the model. With only one control variable, A_i, the regression of interest can be written as

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i. \qquad (2.1)$$

The distinction between the treatment variable, P_i, and the control variable, A_i, in equation (2.1) is conceptual, not formal: there is nothing in equation (2.1) to indicate which is which. Your research question and empirical strategy justify the choice of variables and determine the roles they play.

As in the previous chapter, here we also use Greek letters for parameters to distinguish them from the variables in the model. The regression parameters—called *regression coefficients*—are

- the intercept, α ("alpha");
- the causal effect of treatment, β ("beta");
- and the effect of being a group A student, γ ("gamma").

The last component of equation (2.1) is the *residual,* e_i (also called an error term). Residuals are defined as the difference between the observed Y_i and the *fitted values* generated by the specific regression model we have in mind. These fitted values are written as

$$\hat{Y}_i = \alpha + \beta P_i + \gamma A_i,$$

and the corresponding residuals are given by

$$e_i = Y_i - \hat{Y}_i = Y_i - (\alpha + \beta P_i + \gamma A_i).$$

Regression analysis assigns values to model parameters (α, β, and γ) so as to make $\hat{Y}_i$ as close as possible to Y_i. This is accomplished by choosing values that minimize the sum of squared residuals, leading to the moniker *ordinary least squares* (OLS) for the resulting estimates.[5] Executing this minimization in a particular sample, we are said to be *estimating* regression parameters. 'Metrics masters, who estimate regression models every day, are sometimes said to "run regressions," though often it seems that regressions run us rather than the other way around. The formalities of regression estimation and the statistical theory that goes with it are sketched in the appendix to this chapter.

Running regression (2.1) on data for the five students in groups A and B generates the following estimates (these estimates can be computed using a hand calculator, but for real empirical work, we use professional regression software):

$$\alpha = 40{,}000$$
$$\beta = 10{,}000$$
$$\gamma = 60{,}000.$$

The private school coefficient in this case is 10,000, implying a private-public earnings differential of $10,000. This is indeed a weighted average of our two group-specific effects (recall the group A effect is $-5{,}000$ and the group B effect is 30,000). While this is neither the simple unweighted average (12,500) nor the group-size weighted average (9,000), it's not too far from either of them. In this case, regression assigns a weight of 4/7 to group A and 3/7 to group B. As with these other averages, the regression-weighted average is considerably smaller than the uncontrolled earnings gap between private and public school alumni.[6]

Regression estimates (and the associated standard errors used to quantify their sampling variance) are readily constructed using computers and econometric software. Computational simplicity and the conceptual interpretation of regression estimates as a weighted average of group-specific differences are two of the reasons we regress.

Regression also has two more things going for it. First, it's a convention among masters to report regression estimates in almost every econometric investigation of causal effects, including those involving treatment variables that take on more than two values. Regression estimates provide a simple benchmark for fancier techniques. Second, under some circumstances, regression estimates are efficient in the sense of providing the most statistically precise estimates of average causal effects that we can hope to obtain from a given sample. This technical point is reviewed briefly in the chapter appendix.

Public-Private Face-Off

The C&B data set includes more than 14,000 former students. These students were admitted and rejected at many different combinations of schools (C&B asked for the names of at least three schools students considered seriously, besides the one attended). Many of the possible application/acceptance sets in this data set are represented by only a single student. Moreover, in some sets with more than one student, all schools are either public or private. Just as with groups C and D in Table 2.1, these perfectly homogeneous groups provide no guidance as to the value of a private education.

We can increase the number of useful comparisons by deeming schools to be matched if they are equally selective instead of insisting on identical matches. To fatten up the groups this scheme produces, we'll call schools comparable if they fall into the same Barron's selectivity categories.[7] Returning to our stylized matching matrix, suppose All State and Tall State are rated as Competitive, Altered State and Smart are rated Highly Competitive, and Ivy and Leafy are Most Competitive. In the Barron's scheme, those who applied to Tall State, Smart, and Leafy, and were admitted to Tall State and Smart can be compared with students who applied to All State, Smart, and Ivy, and were admitted to All State and Smart. Students in both groups applied to one Competitive, one Highly Competitive, and one Most Competitive school, and they were admitted to one Competitive and one Highly Competitive school.

In the C&B data, 9,202 students can be matched in this way. But because we're interested in public-private comparisons, our Barron's matched sample is also limited to matched applicant groups that contain both public and private school students. This leaves 5,583 matched students for analysis. These matched students fall into 151 similar-selectivity groups containing both public and private students.

Our operational regression model for the Barron's selectivity-matched sample differs from regression (2.1), used to analyze the matching matrix in Table 2.1, in a number of ways. First, the operational model puts the natural log of earnings on the left-hand side instead of earnings itself. As explained in the chapter appendix, use of a logged dependent variable allows regression estimates to be interpreted as a percent change. For example, an estimated β of .05 implies that private school alumni earn about 5% more than public school alumni, conditional on whatever controls were included in the model.

Another important difference between our operational empirical model and the Table 2.1 example is that the former includes many control variables, while the example controls only for the dummy variable A_i, indicating students in group A. The key controls in the operational model are a set of many dummy variables indicating all Barron's matches represented in the sample (with one group left out as a reference category). These controls capture the relative selectivity of the schools to which students applied and were admitted in the real world, where many combinations of schools are possible. The resulting regression model looks like

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + e_i. \tag{2.2}$$

The parameter β in this model is still the treatment effect of interest, an estimate of the causal effect of attendance at a private school. But this model controls for 151 groups instead of the two groups in our example. The parameters γ_j, for $j = 1$ to 150, are the coefficients on 150

selectivity-group dummies, denoted $GROUP_{ji}$.

It's worth unpacking the notation in equation (2.2), since we'll use it again. The dummy variable $GROUP_{ji}$ equals 1 when student i is in group j and is 0 otherwise. For example, the first of these dummies, denoted $GROUP_{1i}$, might indicate students who applied and were admitted to three Highly Competitive schools. The second, $GROUP_{2i}$, might indicate students who applied to two Highly Competitive schools and one Most Competitive school, and were admitted to one of each type. The order in which the categories are coded doesn't matter as long as we code dummies for all possible combinations, with one group omitted as a reference group. Although we've gone from one group dummy to 150, the idea is as before: controlling for the sets of schools to which students applied and were admitted brings us one giant step closer to a *ceteris paribus* comparison between private and public school students.

A final modification for operational purposes is the addition of two further control variables: individual SAT scores (SAT_i) and the log of parental income (PI_i), plus a few variables we'll relegate to a footnote.[8] The individual SAT and log parental income controls appear in the model with coefficients δ_1 and δ_2 (read as "delta-1" and "delta-2"), respectively. Controls for a direct measure of individual aptitude, like students' SAT scores, and a measure of family background, like parental income, may help make the public-private comparisons at the heart of our model more apples-to-apples and oranges-to-oranges than they otherwise would be. At the same time, conditional on selectivity-group dummies, such controls may no longer matter, a point explored in detail below.

Regressions Run

We start with regression estimates of the private school earnings advantage from models with no controls. The coefficient from a regression of log earnings (in 1995) on a dummy for private school attendance, with no other regressors (right-hand side variables) in the

model, gives the raw difference in log earnings between those who attended a private school and everyone else (the chapter appendix explains why regression on a single dummy variable produces a difference in means across groups defined by the dummy). Not surprisingly, this raw gap, reported in the first column of Table 2.2, shows a substantial private school premium. Specifically, private school students are estimated to have earnings about 14% higher than the earnings of other students.

The numbers that appear in parentheses below the regression estimates in Table 2.2 are the estimated standard errors that go with these estimates. Like the standard errors for a difference in means discussed in the appendix to Chapter 1, these standard errors quantify the statistical precision of the regression estimates reported here. The standard error associated with the estimate in column (1) is .055. The fact that .135 is more than twice the size of the associated standard error of .055 makes it very unlikely the positive estimated private-school gap is merely a chance finding. The private school coefficient is statistically significant.

TABLE 2.2

Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)
Other/missing race			−.074			−.079
			(.157)			(.156)
High school top 10%			.095			.082
			(.027)			(.028)
High school rank missing			.019			.015
			(.033)			(.037)
Athlete			.123			.115
			(.025)			(.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

The large private school premium reported in column (1) of Table 2.2 is an interesting descriptive fact, but, as in our example calculation, some of this gap is almost certainly due to selection bias. As we show below, private school students have higher SAT scores and come from wealthier families than do public school students, and so might be expected to earn more regardless of where they went to college. We therefore control for measures of ability and family background when estimating the private school premium. An estimate of the private school premium from a regression model that includes an individual SAT

control is reported in column (2) of Table 2.2. Every 100 points of SAT achievement are associated with about a 5 percentage point earnings gain. Controlling for students' SAT scores reduces the measured private school premium to about .1. Adding controls for parental income, as well as for demographic characteristics related to race and sex, high school rank, and whether the graduate was a college athlete brings the private school premium down a little further, to a still substantial and statistically significant .086, reported in column (3) of the table.

A substantial effect indeed, but probably still too big, that is, contaminated by positive selection bias. Column (4) reports estimates from a model with no controls for ability, family background, or demographic characteristics. Importantly, however, the regression model used to construct the estimate reported in this column includes a dummy for each matched college selectivity group in the sample. That is, the model used to construct this estimate includes the dummy variables $GROUP_{ji}$, for $j = 1, \ldots, 150$ (the table omits the many estimated γ_j this model produces, but indicates their inclusion in the row labeled "selection controls"). The estimated private school premium with selectivity-group controls included is almost bang on 0, with a standard error of about .04. And that's not all: having killed the private school premium with selectivity-group dummies, columns (5) and (6) show that the premium moves little when controls for ability and family background are added to the model. This suggests that control for college application and admissions selectivity groups takes us a long way toward the apples-to-apples and oranges-to-oranges comparisons at the heart of any credible regression strategy for causal inference.

The results in columns (4)–(6) of Table 2.2 are generated by the subsample of 5,583 students for whom we can construct Barron's matches and generate within-group comparisons of public and private school students. Perhaps there's something special about this limited sample, which contains less than half of the full complement of C&B respondents. This concern motivates a less demanding control scheme that includes only the average SAT score in the set of schools students applied to plus dummies for the number of schools applied to (that is, a

dummy for students who applied to two schools, a dummy for students who applied to three schools, and so on), instead of a full set of 150 selectivity-group dummies. This regression, which can be estimated in the full C&B sample, is christened the "self-revelation model" because it's motivated by the notion that applicants have a pretty good idea of their ability and where they're likely to be admitted. This self-assessment is reflected in the number and average selectivity of the schools to which they apply. As a rule, weaker applicants apply to fewer and to less-selective schools than do stronger applicants.

The self-revelation model generates results remarkably similar to those generated by Barron's matches. The self-revelation estimates, computed in a sample of 14,238 students, can be seen in Table 2.3. As before, the first three columns of the table show that the raw private school premium falls markedly, but remains substantial, when controls for ability and family background are added to the model (falling in this case, from .21 to .14). At the same time, columns (4)–(6) show that models controlling for the number and average selectivity of the schools students apply to generate small and statistically insignificant effects on the order of .03. Moreover, as with the models that control for Barron's matches, models with average selectivity controls generate estimates that are largely insensitive to the inclusion of controls for ability and family background.

Private university attendance seems unrelated to future earnings once we control for selection bias. But perhaps our focus on public-private comparisons misses the point. Students may benefit from attending schools like Ivy, Leafy, or Smart simply because their classmates at such schools are so much better. The synergy generated by a strong peer group may be the feature that justifies the private school price tag.

We can explore this hypothesis by replacing the private school dummy in the self-revelation model with a measure of peer quality. Specifically, as in the original Dale and Krueger study that inspires our analysis, we replace P_i in equation (2.2) with the average SAT score of classmates at the school attended.[9] Columns (1)–(3) of Table 2.4 show that students

who attended more selective schools do markedly better in the labor market, with an estimated college selectivity effect on the order of 8% higher earnings for every 100 points of average selectivity increase. Yet, this effect too appears to be an artifact of selection bias due to the greater ambition and ability of those who attend selective schools. Estimates from models with self-revelation controls, reported in columns (4)–(6) of the table, show average college selectivity to be essentially unrelated to earnings.

TABLE 2.3

Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

TABLE 2.4

School selectivity effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School average SAT score ÷ 100	.109	.071	.076	−.021	−.031	.000
	(.026)	(.025)	(.016)	(.026)	(.026)	(.018)
Own SAT score ÷ 100		.049	.018		.037	.009
		(.007)	(.006)		(.006)	(.006)
Log parental income			.187			.161
			(.024)			(.025)
Female			−.403			−.396
			(.015)			(.014)
Black			−.023			−.034
			(.035)			(.035)
Hispanic			.015			.006
			(.052)			(.053)
Asian			.173			.155
			(.036)			(.037)
Other/missing race			−.188			−.193
			(.119)			(.116)
High school top 10%			.061			.063
			(.018)			(.019)
High school rank missing			.001			−.009
			(.024)			(.022)
Athlete			.102			.094
			(.025)			(.024)
Average SAT score of schools applied to ÷ 100				.138	.116	.089
				(.017)	(.015)	(.013)
Sent two applications				.082	.075	.063
				(.015)	(.014)	(.011)
Sent three applications				.107	.096	.074
				(.026)	(.024)	(.022)
Sent four or more applications				.153	.143	.106
				(.031)	(.030)	(.025)

Notes: This table reports estimates of the effect of alma mater selectivity on earnings. Each column shows coefficients from a regression of log earnings on the average SAT score at the institution attended and controls. The sample size is 14,238. Standard errors are reported in parentheses.

2.3 Ceteris Paribus?

TOPIC: Briefly describe experiences, challenges, and accomplishments that define you as a person.

ESSAY: I am a dynamic figure, often seen scaling walls and crushing ice. I cook Thirty-Minute Brownies in twenty minutes. I am an expert in stucco, a veteran in love, and an outlaw in Peru. On Wednesdays, after school, I repair electrical appliances free of charge.

I am an abstract artist, a concrete analyst, and a ruthless bookie. I wave, dodge, and frolic, yet my bills are all paid. I have won bullfights in San Juan, cliff-diving competitions in Sri Lanka, and spelling bees at the Kremlin. I have played Hamlet, I have performed open-heart surgery, and I have spoken with Elvis.

But I have not yet gone to college.

From an essay by Hugh Gallagher, age 19.
(Hugh later went to New York University.)

Imagine Harvey and Uma on the day admissions letters go out. Both are delighted to get into Harvard (it must be those 20-minute brownies). Harvey immediately accepts Harvard's offer—wouldn't you? But Uma makes a difficult choice and goes to U-Mass instead. What's up with Uma? Is her *ceteris* really *paribus?*

Uma might have good reasons to opt for less-prestigious U-Mass over Harvard. Price is an obvious consideration (Uma won a Massachusetts Adams Scholarship, which pays state school tuition for good students like her but cannot be used at private schools). If price matters more to Uma than to Harvey, it's possible that Uma's circumstances differ from Harvey's in other ways. Perhaps she's poorer. Some of our regression models control for parental income, but this is an imperfect measure of family living standards. Among other things, we don't know how many brothers and sisters the students in the C&B sample had. A larger family at the same income level may find it harder to pay for each child's education. If family size is also related to later earnings (see Chapter 3 for more on this point), our regression estimates of private college premia may not be apples-to-apples after all.

This is more than a campfire story. Regression is a way to make other things equal, but equality is generated only for variables included as

controls on the right-hand side of the model. Failure to include enough controls or the right controls still leaves us with selection bias. The regression version of the selection bias generated by inadequate controls is called *omitted variables bias* (OVB), and it's one of the most important ideas in the 'metrics canon.

To illustrate OVB, we return to our five-student example and the bias from omitting control for membership in applicant group A. The "long regression" here includes the dummy variable, A_i, which indicates those in group A. We write the regression model that includes A_i as

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l. \qquad (2.3)$$

This is equation (2.1) rewritten with superscript l on parameters and the residual to remind us that the intercept and private school coefficient are from the long model, and to facilitate comparisons with the short model to come.

Does the inclusion of A_i matter for estimates of the private school effect in the regression above? Suppose we make do with a short regression with no controls. This can be written as

$$Y_i = \alpha^s + \beta^s P_i + e_i^s.$$

Because the single regressor here is a dummy variable, the slope coefficient in this model is the difference in average Y_i between those with P_i switched on and those with P_i switched off. As we noted in Section 2.1, $\beta^s = 20{,}000$ in the short regression, while the long regression parameter, β^l, is only 10,000. The difference between β^s and β^l is the OVB due to omission of A_i in the short regression. Here, OVB amounts to $10,000, a figure worth worrying about.

Why does the omission of the group A dummy change the private college effect so much? Recall that the average earnings of students in group A exceeds the average earnings of those in group B. Moreover, two-thirds of the students in high-earning group A attended a private

school, while lower-earning group B is only half private. Differences in earnings between private and public alumni come in part from the fact that the mostly private students in group A have higher earnings anyway, regardless of where they enrolled. Inclusion of the group A dummy in the long regression controls for this difference.

As this discussion suggests, the formal connection between short and long regression coefficients has two components:

(i) The relationship between the omitted variable (A_i) and the treatment variable (P_i); we'll soon see how to quantify this with an additional regression.

(ii) The relationship between the omitted variable (A_i) and the outcome variable (Y_i). This is given by the coefficient on the omitted variable in the long regression, in this case, the parameter γ in equation (2.3).

Together, these pieces produce the *OVB formula*. We start with the fact that

$$\begin{aligned} \textit{Effect of } P_i \textit{ in short} = \textit{Effect of } P_i \textit{ in long} \\ + \big(\{\textit{Relationship between omitted and included}\} \\ \times \{\textit{Effect of omitted in long}\}\big). \end{aligned}$$

To be specific, when the omitted variable is A_i and the treatment variable is P_i, we have

$$\begin{aligned} \textit{Effect of } P_i \textit{ in short} = \textit{Effect of } P_i \textit{ in long} \\ + \big(\{\textit{Relationship between } A_i \textit{ and } P_i\} \\ \times \{\textit{Effect of } A_i \textit{ in long}\}\big). \end{aligned}$$

Omitted variables bias, defined as the difference between the coefficient on P_i in the short and long models, is a simple rearrangement of this equation:

$$OVB = \{Relationship\ between\ A_i\ and\ P_i\} \times \{Effect\ of\ A_i\ in\ long\}.$$

We can refine the OVB formula using the fact that both terms in the formula are themselves regression coefficients. The first term is the coefficient from a regression of the omitted variable A_i on the private school dummy. In other words, this term is the coefficient π_1 (read "pi-1") in the regression model

$$A_i = \pi_0 + \pi_1 P_i + u_i,$$

where u_i is a residual. We can now write the OVB formula compactly in Greek:

$$\begin{aligned} OVB &= Effect\ of\ P_i\ in\ short - Effect\ of\ P_i\ in\ long \\ &= \beta^s - \beta^l = \pi_1 \times \gamma, \end{aligned}$$

where γ is the coefficient on A_i in the long regression. This important formula is derived in the chapter appendix.

Among students who attended private school, two are in group A and one in group B, while among those who went to public school, one is in group A and one in group B. The coefficient π_1 in our five-student example is therefore $2/3 - 1/2 = .1667$. As noted in Section 2.2, the coefficient γ is 60,000, reflecting the higher earnings of group A. Putting the pieces together, we have

$$\begin{aligned} OVB &= Short - Long \\ &= \beta^s - \beta^l \\ &= 20{,}000 - 10{,}000 = 10{,}000 \end{aligned}$$

and

$$\begin{aligned} OVB &= \{Regression\ of\ omitted\ on\ included\} \\ &\quad \times \{Effect\ of\ omitted\ in\ long\} \\ &= \pi_1 \times \gamma = .1667 \times 60{,}000 = 10{,}000. \end{aligned}$$

Phew! The calculation suggested by the OVB formula indeed matches the direct comparison of short and long regression coefficients.

The OVB formula is a mathematical result that explains differences between regression coefficients in any short-versus-long scenario, irrespective of the causal interpretation of the regression parameters. The labels "short" and "long" are purely relative: The short regression need not be particularly short, but the long regression is always longer, since it includes the same regressors plus at least one more. Often, the additional variables that make the long regression long are hypothetical, that is, unavailable in our data. The OVB formula is a tool that allows us to consider the impact of control for variables we *wish* we had. This in turn helps us assess whether *ceteris* is indeed *paribus*. Which brings us back to Uma and Harvey.

Suppose an omitted variable in equation (2.2) is family size, FS_i. We've included parental income as a control variable, but not the number of brothers and sisters who might also go to college, which is not available in the C&B data set. When the omitted variable is FS_i, we have

$$\begin{aligned} OVB &= Short - Long \\ &= \{Relationship\ between\ FS_i\ and\ P_i\} \\ &\quad \times \{Effect\ of\ FS_i\ in\ long\}. \end{aligned}$$

Why might the omission of family size bias regression estimates of the private college effect? Because differences in earnings between Harvard and U-Mass graduates arise in part from differences in family size between the two groups of students (this is the relationship between FS_i and P_i) *and* from the fact that smaller families are associated with higher earnings, even after controlling for the variables included in the short regression (this is the effect of FS_i in the long regression, which includes these same controls as well). The long regression controls for the fact

that students who go to Harvard come from smaller families (on average) than do students who went to U-Mass, while the short regression that omits FS_i does not.

The first term in this application of the OVB formula is the coefficient in a regression of omitted (FS_i) on included (P_i) variables and everything else that appears on the right-hand side of equation (2.2). This regression—which is sometimes said to be "auxiliary" because it helps us interpret the regression we care about—can be written as

$$FS_i = \pi_0 + \pi_1 P_i + \sum_j \theta_j GROUP_{ji} + \pi_2 SAT_i + \pi_3 \ln PI_i + u_i. \tag{2.4}$$

Most of the coefficients in equation (2.4) are of little interest. What matters here is π_1, since this captures the relationship between the omitted variable, FS_i, and the variable whose effect we're after, P_i, after controlling for other variables that appear in both the short and long regression models.[10]

To complete the OVB formula for this case, we write the long regression as

$$\ln Y_i = \alpha^l + \beta^l P_i + \sum_j \gamma_j^l GROUP_{ji} + \delta_1^l SAT_i + \delta_2^l \ln PI_i + \lambda FS_i + e_i^l, \tag{2.5}$$

again using superscript l for "long." The regressor FS_i appears here with coefficient λ.[11] The OVB formula is therefore

$$OVB = Short - Long = \beta - \beta^l = \pi_1 \times \lambda,$$

where β is from equation (2.2).

Continuing to think of equation (2.2) as the short regression, while the long regression includes the control variables that appear in this model plus family size, we see that OVB here is probably positive. Private school students tend to come from smaller families on average, even

after conditioning on family income. If so, the regression coefficient linking family size and private college attendance is negative ($\pi_1 < 0$ in equation (2.4)). Students from smaller families are also likely to earn more no matter where they go to school, so the effect of omitting family size controls in a long regression is also negative ($\lambda < 0$ in equation (2.5)). The product of these two negative terms is positive.

Careful reasoning about OVB is an essential part of the 'metrics game. We can't use data to check the consequences of omitting variables that we don't observe, but we can use the OVB formula to make an educated guess as to the likely consequences of their omission. Most of the control variables that might be omitted from equation (2.2) are similar to family size in that the sign of the OVB from their omission is probably positive. From this we conclude that, as small as the estimates of the effects of private school attendance in columns (4)–(6) of Tables 2.2–2.3 are, they could well be too big. These estimates therefore weigh strongly against the hypothesis of a substantial private school earnings advantage.

Regression Sensitivity Analysis

Because we can never be sure whether a given set of controls is enough to eliminate selection bias, it's important to ask how sensitive regression results are to changes in the list of controls. Our confidence in regression estimates of causal effects grows when treatment effects are insensitive—masters say "robust"—to whether a particular variable is added or dropped as long as a few core controls are always included in the model. This desirable pattern is illustrated by columns (4)–(6) in Tables 2.2–2.3, which show that estimates of the private school premium are insensitive to the inclusion of students' ability (as measured by own SAT scores), parental income, and a few other control variables, once we control for the nature of the schools to which students applied.

The OVB formula explains this remarkable finding. Start with Table 2.5, which reports coefficients from regressions like equation (2.4), except that instead of FS_i, we put SAT_i on the left-hand side to produce the estimates in columns (1)–(3) while ln PI_i on the left-hand side

generates columns (4)–(6). These auxiliary regressions assess the relationship between private school attendance and two of our controls, SAT_i and ln PI_i, conditional on other controls in the model. Not surprisingly, private school attendance is a strong predictor of students' own SAT scores and family income, relationships documented in columns (1) and (4) in the table. The addition of demographic controls, high school rank, and a dummy for athletic participation does little to change this, as can be seen in columns (2) and (5). But control for the number of applications and the average SAT score of schools applied to, as in the self-revelation model, effectively eliminates the relationship between private school attendance and these important background variables. This explains why the estimated private school coefficients in columns (4), (5), and (6) of Table 2.3 are essentially the same.

TABLE 2.5

Private school effects: Omitted variables bias

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165	1.130	.066	.128	.138	.028
	(.196)	(.188)	(.112)	(.035)	(.037)	(.037)
Female		−.367			.016	
		(.076)			(.013)	
Black		−1.947			−.359	
		(.079)			(.019)	
Hispanic		−1.185			−.259	
		(.168)			(.050)	
Asian		−.014			−.060	
		(.116)			(.031)	
Other/missing race		−.521			−.082	
		(.293)			(.061)	
High school top 10%		.948			−.066	
		(.107)			(.011)	
High school rank missing		.556			−.030	
		(.102)			(.023)	
Athlete		−.318			.037	
		(.147)			(.016)	
Average SAT score of schools applied to ÷ 100			.777			.063
			(.058)			(.014)
Sent two applications			.252			.020
			(.077)			(.010)
Sent three applications			.375			.042
			(.106)			(.013)
Sent four or more applications			.330			.079
			(.093)			(.014)

Notes: This table describes the relationship between private school attendance and personal characteristics. Dependent variables are the respondent's SAT score (divided by 100) in columns (1)–(3) and log parental income in columns (4)–(6). Each column shows the coefficient from a regression of the dependent variable on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

The OVB formula is the Prime Directive of applied econometrics, so let's rock it with our numbers and see how it works out. For illustration, we'll take the short model to be a regression of log wages on P_i with no controls and the long model to be the regression that adds individual SAT scores. The short (no controls) coefficient on P_i in column (1) of Table 2.3 is .212, while the corresponding long coefficient (controlling

for SAT_i) in column (2) is .152. As can also be seen in column (2) of the table, the effect of SAT_i in the long regression is .051. The first column in Table 2.5 shows that the regression of omitted SAT_i on included P_i produces a coefficient of 1.165. Putting these together, we have OVB, two ways:

$$\begin{aligned} OVB &= Short - Long = .212 - .152 = .06 \\ OVB &= \{Regression\ of\ omitted\ on\ included\} \\ &\quad \times \{Effect\ of\ omitted\ in\ long\} \\ &= 1.165 \times .051 = .06. \end{aligned}$$

Compare this with the parallel calculation taking us from column (4) to column (5) in Table 2.3. These columns report results from models that include self-revelation controls. Here, *Short* − *Long* is small: .034 − .031 = .003, to be precise. Both the short and long regressions include selectivity controls from the self-revelation model, as does the relevant auxiliary regression of own SAT scores on P_i. With self-revelation controls included in both models, we have

$$\begin{aligned} OVB &= \{Regression\ of\ omitted\ on\ included\} \\ &\quad \times \{Effect\ of\ omitted\ in\ long\} \\ &= .066 \times .036 = .0024. \end{aligned}$$

(Rounding error with small numbers pushes us off of the target of .003.) The effect of the omitted SAT_i in the long regression falls here from .051 to .036, while the regression of omitted on included goes from a hefty 1.165 to something an order of magnitude smaller at .066 (shown in column (3) of Table 2.5). This shows that, conditional on the number and average selectivity of schools applied to, students who chose private and public schools aren't very different, at least as far as their own SAT scores go. Consequently, the gap between short and long estimates disappears.

Because our estimated private school effect is insensitive to the inclusion of the available ability and family background variables once

the self-revelation controls are included, other control variables, including those for which we have no data, might matter little as well. In other words, any remaining OVB due to uncontrolled differences is probably modest.[12] This circumstantial evidence for modest OVB doesn't guarantee that the regression results discussed in this chapter have the same causal force as results from a randomized trial—we'd still rather have a real experiment. At a minimum, however, these findings call into question claims for a substantial earnings advantage due to attendance at expensive private colleges.

MASTER STEVEFU: In a nutshell, please, Grasshopper.

GRASSHOPPER: Causal comparisons compare like with like. In assessing the effects of college choice, we focus on students with similar characteristics.

MASTER STEVEFU: Each is different in a thousand ways. Must all ways be similar?

GRASSHOPPER: Good comparisons eliminate systematic differences between those who chose one path and those who choose another, when such differences are associated with outcomes.

MASTER STEVEFU: How is this accomplished?

GRASSHOPPER: The method of matching sorts individuals into groups with the same values of control variables, like measures of ability and family background. Matched comparisons within these groups are then averaged to get a single overall effect.

MASTER STEVEFU: And regression?

GRASSHOPPER: Regression is an automated matchmaker. The regression estimate of a causal effect is also an average of within-group comparisons.

MASTER STEVEFU: What is the Tao of OVB?

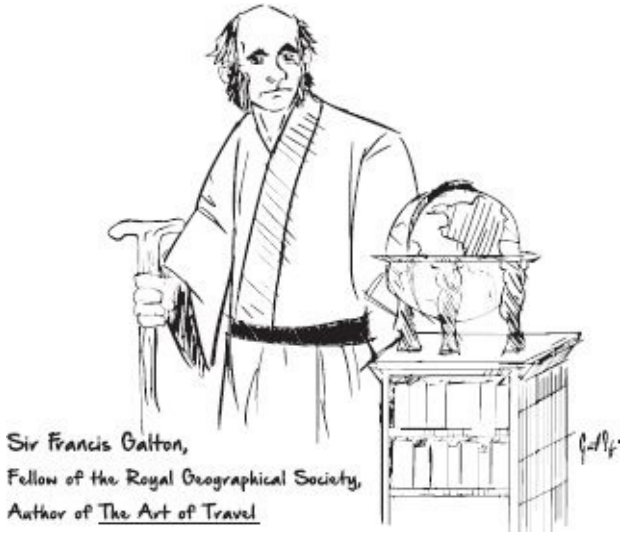
GRASSHOPPER: OVB is the difference between short and long regression coefficients. The long regression includes additional controls, those omitted from the short. Short equals long plus the

effect of omitted in long times the regression of omitted on included.

MASTER JOSHWAY: Nothing omitted here, Grasshopper.

Masters of 'Metrics: Galton and Yule

The term "regression" was coined by Sir Francis Galton, Charles Darwin's half-cousin, in 1886. Galton had many interests, but he was gripped by Darwin's masterpiece, *The Origin of Species*. Galton hoped to apply Darwin's theory of evolution to variation in human traits. In the course of his research, Galton studied attributes ranging from fingerprints to beauty. He was also one of many British intellectuals to use Darwin in the sinister service of eugenics. This regrettable diversion notwithstanding, his work in theoretical statistics had a lasting and salutary effect on social science. Galton laid the statistical foundations for quantitative social science of the sort that grips us.



Galton discovered that the average heights of fathers and sons are linked by a regression equation. He also uncovered an interesting implication of this particular regression model: the average height of sons is a weighted average of their fathers' height and the average height in the population from which the fathers and sons were sampled. Thus, parents who are taller than average will have children who are not quite

as tall, while parents who are shorter than average will have children who are a bit taller. To be specific, Master Stevefu, who is 6′3″, can expect his children to be tall, though not as tall as he is. Thankfully, however, Master Joshway, who is 5′6″ on a good day, can expect his children to attain somewhat grander stature.

Galton explained this averaging phenomenon in his celebrated 1886 paper "Regression towards Mediocrity in Hereditary Stature."[13] Today, we call this property "regression to the mean." Regression to the mean is not a causal relationship. Rather, it's a statistical property of correlated pairs of variables like the heights of fathers and sons. Although fathers' and sons' heights are never exactly the same, their frequency distributions are essentially unchanging. This distributional stability generates the Galton regression.

We see regression as a statistical procedure with the power to make comparisons more equal through the inclusion of control variables in models for treatment effects. Galton seems to have been uninterested in regression as a control strategy. The use of regression for statistical control was pioneered by George Udny Yule, a student of statistician Karl Pearson, who was Galton's protégé. Yule realized that Galton's regression method could be extended to include many variables. In an 1899 paper, Yule used this extension to link the administration of the English Poor Laws in different counties to the likelihood county residents were poor, while controlling for population growth and the age distribution in the county.[14] The poor laws provided subsistence for the indigent, usually by offering shelter and employment in institutions called workhouses. Yule was particularly interested in whether the practice of outdoor relief, which provided income support for poor people without requiring them to move to a workhouse, increased poverty rates by making pauperism less onerous. This is a well-defined causal question much like those that occupy social scientists today.

Appendix: Regression Theory

Conditional Expectation Functions

Chapter 1 introduces the notion of mathematical expectation, called "expectation" for short. We write $E[Y_i]$ for the expectation of a variable, Y_i. We're also concerned with *conditional expectations,* that is, the expectation of a variable in groups (also called "cells") defined by a second variable. Sometimes this second variable is a dummy, taking on only two values, but it need not be. Often, as in this chapter, we're interested in conditional expectations in groups defined by the values of variables that aren't dummies, for example, the expected earnings for people who have completed 16 years of schooling. This sort of conditional expectation can be written as

$$E[Y_i|X_i = x],$$

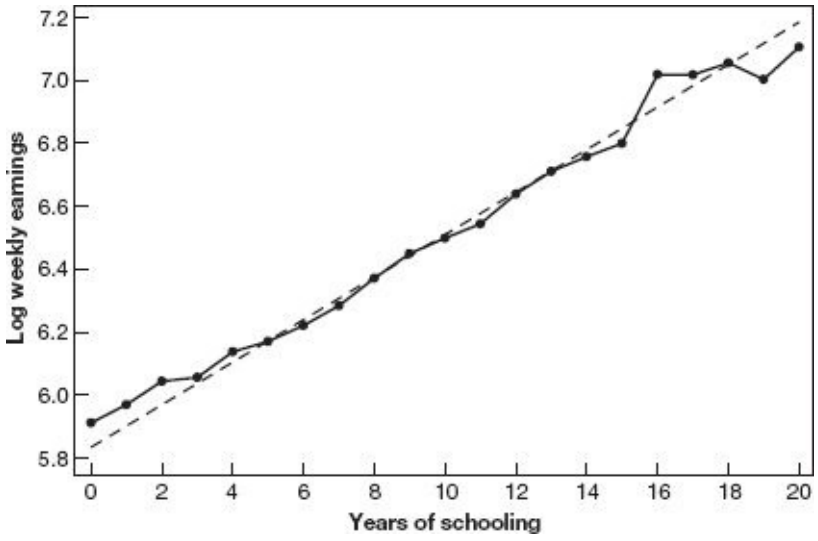
and it's read as "The conditional expectation of Y_i given that X_i equals the particular value x."

Conditional expectations tell us how the population average of one variable changes as we move the conditioning variable over the values this variable might assume. For every value of the conditioning variable,

we might get a different average of the dependent variable, Y_i. The collection of all such averages is called the *conditional expectation function* (CEF for short). $E[Y_i|X_i]$ is the CEF of Y_i given X_i, without specifying a value for X_i, while $E[Y_i|X_i = x]$ is one point in the range of this function.

A favorite CEF of ours appears in Figure 2.1. The dots in this figure show the average log weekly wage for men with different levels of schooling (measured by highest grade completed), with schooling levels arrayed on the X-axis (data here come from the 1980 U.S. Census). Though it bobs up and down, the earnings-schooling CEF is strongly upward-sloping, with an average slope of about .1. In other words, each year of schooling is associated with wages that are about 10% higher on average.

FIGURE 2.1
The CEF and the regression line



Notes: This figure shows the conditional expectation function (CEF) of log weekly wages given years of education, and the line generated by regressing log weekly wages on years of education (plotted as a broken line).

Many of the CEFs we're interested in involve more than one conditioning variable, each of which takes on two or more values. We write

$$E[Y_i|X_{1i}, \ldots, X_{Ki}]$$

for a CEF with K conditioning variables. With many conditioning variables, the CEF is harder to plot, but the idea is the same. $E[Y_i|X_{1i} = x_1, \ldots, X_{Ki} = x_K]$ gives the population average of Y_i with these K other variables held fixed. Instead of looking at average wages conditional only on schooling, for example, we might also condition on cells defined by age, race, and sex.

Regression and the CEF

Table 2.1 illustrates the matchmaking idea by comparing students who attended public and private colleges, after sorting students into cells on the basis of the colleges to which they applied and were admitted. The body of the chapter explains how we see regression as a quick and easy way of automating such matched comparisons. Here, we use the CEF to make this interpretation of regression more rigorous.[15]

The regression estimates of equation (2.2) reported in Table 2.3 suggest that private school attendance is unrelated to average earnings once individual SAT scores, parental income, and the selectivity of colleges applied and admitted to are held fixed. As a simplification, suppose that the CEF of log wages is a linear function of these conditioning variables. Specifically, assume that

$$\begin{aligned} &E[\ln Y_i|P_i, GROUP_i, SAT_i, \ln PI_i] \qquad (2.6)\\ &= \alpha + \beta P_i + \sum_j \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i, \end{aligned}$$

where Greek letters, as always, are parameters. When the CEF of $\ln Y_i$ is a linear function of the conditioning variables as in equation (2.6), the regression of $\ln Y_i$ on these same conditioning variables recovers this linear function. (We skip a detailed proof of this fact, though it's not hard to show.) In particular, given linearity, the coefficient on P_i in equation (2.2) will be equal to the coefficient on P_i in equation (2.6).

With a linear CEF, regression estimates of private school effects based on equation (2.2) are also identical to those we'd get from a strategy that (i) matches students by values of $GROUP_i$, SAT_i, and ln PI_i; (ii) compares the average earnings of matched students who went to private ($P_i = 1$) and public ($P_i = 0$) schools for each possible combination of the conditioning variables; and (iii) produces a single average by averaging all of these cell-specific contrasts. To see this, it's enough to use equation (2.6) to write cell-specific comparisons as

$$E[\ln Y_i | P_i = 1, GROUP_i, SAT_i, \ln PI_i] - E[\ln Y_i | P_i = 0, GROUP_i, SAT_i, \ln PI_i] = \beta.$$

Because our linear model for the CEF assumes that the effect of private school attendance is equal to the constant β in every cell, any weighted average of cell-specific private-attendance contrasts is also equal to β.

Linear models help us understand regression, but regression is a wonderfully flexible tool, useful regardless of whether the underlying CEF is linear. Regression inherits this flexibility from the following pair of closely related theoretical properties:

- If $E\ E[Y_i|X_{1i}, \ldots, X_{Ki}] = a + \sum_{k=1}^{K} b_k X_{ki}$ for some constants a and b_1, …, b_K, then the regression of Y_i on X_{1i}, …, X_{Ki} has intercept a and slopes b_1, …, b_K. In other words, if the CEF of Y_i on X_{1i}, …, X_{Ki} is linear, then the regression of Y_i on X_{1i}, …, X_{Ki} is it.
- If $E[Y_i|X_{1i}, \ldots, X_{Ki}]$ is a nonlinear function of the conditioning variables, then the regression of Y_i on X_{1i}, …, X_{Ki} gives the best linear approximation to this nonlinear CEF in the sense of minimizing the expected squared deviation between the fitted values from a linear model and the CEF.

To summarize: if the CEF is linear, regression finds it; if not linear, regression finds a good approximation to it. We've just used the first theoretical property to interpret regression estimates of private school

effects when the CEF is linear. The second property tells us that we can expect regression estimates of a treatment effect to be close to those we'd get by matching on covariates and then averaging within-cell treatment-control differences, even if the CEF isn't linear.

Figure 2.1 documents the manner in which regression approximates the nonlinear CEF of log wages conditional on schooling. Although the CEF bounces around the regression line, this line captures the strong positive relationship between schooling and wages. Moreover, the regression slope is close to $E\{E[Y_i|X_i] - E[Y_i|X_i - 1]\}$; that is, the regression slope also comes close to the expected effect of a one-unit change in X_i on $E[Y_i|X_i]$.[16]

Bivariate Regression and Covariance

Regression is closely related to the statistical concept of *covariance*. The covariance between two variables, X_i and Y_i, is defined as

$$C(X_i, Y_i) = E\big[\big(X_i - E[X_i]\big)\big(Y_i - E[Y_i]\big)\big].$$

Covariance has three important properties:

(i) The covariance of a variable with itself is its variance; $C(X_i, X_i) = \sigma_X^2$.

(ii) If the expectation of either X_i or Y_i is 0, the covariance between them is the expectation of their product; $C(X_i, Y_i) = E[X_iY_i]$.

(iii) The covariance between linear functions of variables X_i and Y_i —written $W_i = a + bX_i$ and $Z_i = c + dY_i$ for constants a, b, c, d—is given by

$$C(W_i, Z_i) = bdC(X_i, Y_i).$$

The intimate connection between regression and covariance can be seen in a *bivariate regression model,* that is, a regression with one regressor, X_i, plus an intercept.[17] The bivariate regression slope and

intercept are the values of a and b that minimize the associated *residual sum of squares,* which we write as

$$RSS(a, b) = E[Y_i - a - bX_i]^2.$$

The term *RSS* references a *sum* of squares because, carrying out this minimization in a particular sample, we replace expectation with a sample average or sum. The solution for the bivariate case is

$$b = \beta = \frac{C(Y_i, X_i)}{V(X_i)} \qquad (2.7)$$
$$a = \alpha = E[Y_i] - \beta E[X_i].$$

An implication of equation (2.7) is that when two variables are *uncorrelated* (have a covariance of 0), the regression of either one on the other generates a slope coefficient of 0. Likewise, a bivariate regression slope of 0 implies the two variables involved are uncorrelated.

Fits and Residuals

Regression breaks any dependent variable into two pieces. Specifically, for dependent variable Y_i, we can write

$$Y_i = \hat{Y}_i + e_i.$$

The first term consists of the fitted values, $\hat{Y}_i$, sometimes said to be the part of Y_i that's "explained" by the model. The second part, the residuals, e_i, is what's left over.

Regression residuals and the regressors included in the model that produced them are uncorrelated. In other words, if e_i is the residual from a regression on X_{1i}, …, X_{Ki}, then the regression of e_i on these same variables produces coefficients that are all 0. Because fitted values are a linear combination of regressors, they're also uncorrelated with residuals. We summarize these important properties here.

PROPERTIES OF RESIDUALS Suppose that α and $\beta_1, \ldots, \beta_K$ are the intercept and slope coefficients from a regression of Y_i on $X_{1i}, \ldots, X_{Ki}$. The *fitted values* from this regression are

$$\hat{Y}_i = \alpha + \sum_{k=1}^{K} \beta_k X_{ki},$$

and the associated regression *residuals* are

$$e_i = Y_i - \hat{Y}_i = Y_i - \alpha - \sum_{k=1}^{K} \beta_k X_{ki}.$$

Regression residuals

(i) have expectation and sample mean 0: $E[e_i] = \sum_{i=1}^{n} e_i = 0$;

(ii) are uncorrelated in both population and sample with all regressors that made them and with the corresponding fitted values. That is, for each regressor, X_{ki},

$$E[X_{ki}e_i] = \sum_{i=1}^{n} X_{ki}e_i = 0 \text{ and } E[\hat{Y}_i e_i] = \sum_{i=1}^{n} \hat{Y}_i e_i = 0.$$

You can take these properties on faith, but for those who know a little calculus, they're easy to establish. Start with the fact that regression parameters and estimates minimize the residual sum of squares. The first-order conditions for this minimization problem amount to statements equivalent to (i) and (ii).

Regression for Dummies

An important regression special case is bivariate regression with a dummy regressor. The conditional expectation of Y_i given a dummy variable, Z_i, takes on two values. Write them in Greek, like this:

$$E[Y_i|Z_i = 0] = \alpha$$
$$E[Y_i|Z_i = 1] = \alpha + \beta,$$

so that

$$\beta = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

is the difference in expected Y_i with the dummy regressor, Z_i, switched on and off.

Using this notation, we can write

$$\begin{aligned} E[Y_i|Z_i] &= E[Y_i|Z_i = 0] + \big(E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]\big)Z_i \\ &= \alpha + \beta Z_i. \end{aligned} \tag{2.8}$$

This shows that $E[Y_i|Z_i]$ is a linear function of Z_i, with slope β and intercept α. Because the CEF with a single dummy variable is linear, regression fits this CEF perfectly. As a result, the regression slope must also be $\beta = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$, the difference in expected Y_i with Z_i switched on and off.

Regression for dummies is important because dummy regressors crop up often, as in our analyses of health insurance and types of college attended.

Regression Anatomy and the OVB Formula

The most interesting regressions are multiple; that is, they include a causal variable of interest, plus one or more control variables. Equation (2.2), for example, regresses log earnings on a dummy for private college attendance in a model that controls for ability, family background, and the selectivity of schools that students have applied to and been admitted to. We've argued that control for covariates in a regression model is much like matching. That is, the regression coeffiicent on a private school dummy in a model with controls is similar to what we'd get if we divided students into cells based on these controls, compared

public school and private school students within these cells, and then took an average of the resulting set of conditional comparisons. Here, we offer a more detailed "regression anatomy" lesson.

Suppose the causal variable of interest is X_{1i} (say, a dummy for private school) and the control variable is X_{2i} (say, SAT scores). With a little work, the coefficient on X_{1i} in a regression controlling for X_{2i} can be written as

$$\beta_1 = \frac{C(Y_i, \tilde{X}_{1i})}{V(\tilde{X}_{1i})},$$

where $\tilde{X}_{1i}$ is the residual from a regression of X_{1i} on X_{2i}:

$$X_{1i} = \pi_0 + \pi_1 X_{2i} + \tilde{X}_{1i}.$$

As always, residuals are uncorrelated with the regressors that made them, and so it is for the residual $\tilde{X}_{1i}$. It's not surprising, therefore, that the coefficient on X_{1i} in a multivariate regression that controls for X_{2i} is the bivariate coefficient from a model that includes only the part of X_{1i} that is uncorrelated with X_{2i}. This important regression anatomy formula shapes our understanding of regression coefficients from around the world.

The regression anatomy idea extends to models with more than two regressors. The multivariate coefficient on a given regressor can be written as the coefficient from a bivariate regression on the residual from regressing this regressor on all others. Here's the anatomy of the *k*th coefficient in a model with *K* regressors:

REGRESSION ANATOMY

$$\beta_k = \frac{C(Y_i, \tilde{X}_{ki})}{V(\tilde{X}_{ki})},$$

where $\tilde{X}_{ki}$ is the residual from a regression of X_{ki} on the $K - 1$ other

covariates included in the model.

Regression anatomy is especially revealing when the controls consist of dummy variables, as in equation (2.2). For the purposes of this discussion, we simplify the model of interest to have only dummy controls, that is,

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + e_i. \qquad (2.9)$$

Regression anatomy tells us that the coefficient on P_i controlling for the set of 150 $GROUP_{ji}$ dummies is the bivariate coefficient from a regression on $\tilde{P}_i$, where this is the residual from a regression of P_i on a constant and the set of 150 $GROUP_{ji}$ dummies.

It's helpful here to add a second subscript to index groups as well as individuals. In this scheme, $\ln Y_{ij}$ is the log earnings of college graduate i in selectivity group j, while P_{ij} is this graduate's private school enrollment status. What is the residual, $\tilde{P}_{ij}$, from the auxiliary regression of P_{ij} on the set of 150 selectivity-group dummies? Because the auxiliary regression that generates $\tilde{P}_{ij}$ has a parameter for every possible value of the underlying CEF, this regression captures the CEF of P_{ij} conditional on selectivity group perfectly. (Here we're extending the dummy-variable result described by equation (2.8) to regression on dummies describing a categorical variable that takes on many values instead of just two.) Consequently, the fitted value from a regression of P_{ij} on the full set of selectivity-group dummies is the mean private school attendance rate in each group. For applicant i in group j, the auxiliary regression residual is therefore $\tilde{P}_{ij} = P_{ij} - \bar{P}_j$, where $\bar{P}_j$ is shorthand for the mean private school enrollment rate in the selectivity group to which i belongs.

Finally, putting the pieces together, regression anatomy tells us that the multivariate β in the model described by equation (2.9) is

$$\beta = \frac{C(\ln Y_{ij}, \tilde{P}_{ij})}{V(\tilde{P}_{ij})} = \frac{C(\ln Y_{ij}, P_{ij} - \bar{P}_j)}{V(P_{ij} - \bar{P}_j)}. \quad (2.10)$$

This expression reveals that, just as if we were to manually sort students into groups and compare public and private students within each group, regression on private school attendance with control for selectivity-group dummies is also a within-group procedure: variation across groups is removed by subtracting $\bar{P}_j$ to construct the residual, $\tilde{P}_{ij}$. Moreover, as for groups C and D in Table 2.1, equation (2.10) implies that applicant groups in which everyone attends either a public or private institution are uninformative about the effects of private school attendance because $P_{ij} - \bar{P}_j$ is 0 for everyone in such groups.

The OVB formula, used at the end of this chapter (in Section 2.3) to interpret estimates from models with different sets of controls, provides another revealing take on regression anatomy. Call the coefficient on X_{1i} in a multivariate regression model controlling for X_{2i} the long regression coefficient, β^l:

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l.$$

Call the coefficient on X_{1i} in a bivariate regression (that is, without X_{2i}) the short regression coefficient, β^s:

$$Y_i = \alpha^s + \beta^s X_{1i} + e_i^s.$$

The OVB formula describes the relationship between short and long coefficients as follows.

OMITTED VARIABLES BIAS (OVB) FORMULA

$$\beta^s = \beta^l + \pi_{21}\gamma,$$

where γ is the coefficient on X_{2i} in the long regression, and π_{21} is the coefficient on X_{1i} in a regression of X_{2i} on X_{1i}. In words: *short equals*

long plus the effect of omitted times the regression of omitted on included.

This central formula is worth deriving. The slope coefficient in the short model is

$$\beta^s = \frac{C(Y_i, X_{1i})}{V(X_{1i})}. \qquad (2.11)$$

Substituting the long model for Y_i in equation (2.11) gives

$$\begin{aligned}
&\frac{C(\alpha^l + \beta_1^l X_{1i} + \gamma X_{2i} + e_i^l, X_{1i})}{V(X_{1i})} \\
&= \frac{\beta_1^l V(X_{1i}) + \gamma C(X_{2i}, X_{1i}) + C(e_i^l, X_{1i})}{V(X_{1i})} \\
&= \beta_1^l + \frac{C(X_{2i}, X_{1i})}{V(X_{1i})}\gamma = \beta_1^l + \pi_{21}\gamma.
\end{aligned}$$

The first equals sign comes from the fact that the covariance of a linear combination of variables is the corresponding linear combination of covariances after distributing terms. Also, the covariance of a constant with anything else is 0, and the covariance of a variable with itself is the variance of that variable. The second equals sign comes from the fact that $C(e_i^l, X_{1i}) = 0$, because residuals are uncorrelated with the regressors that made them (e_i^l is the residual from a regression that includes X_{1i}). The third equals sign defines π_{21} to be the coefficient on X_{1i} in a regression of X_{2i} on X_{1i}.[18]

Often, as in the discussion of equations (2.2) and (2.5), we're interested in short vs. long comparisons across regression models that include a set of controls common to both models. The OVB formula for this scenario is a straightforward extension of the one above. Call the coefficient on X_{1i} in a multivariate regression controlling for X_{2i} and X_{3i} the long regression coefficient, β^l; call the coefficient on X_{1i} in a multivariate regression controlling only for X_{3i} (that is, without X_{2i}) the short regression coefficient, β^s. The OVB formula in this case can still be

written

$$\beta^s = \beta^l + \pi_{21}\gamma, \qquad (2.12)$$

where γ is the coefficient on X_{2i} in the long regression, but that regression now includes X_{3i} as well as X_{2i}, and π_{21} is the coefficient on X_{1i} in a regression of X_{2i} on both X_{1i} and X_{3i}. Once again, we can say: *short equals long plus the effect of omitted times the regression of omitted on included.* We leave it to the reader to derive equation (2.12); this derivation tests your understanding (and makes an awesome exam question).

Building Models with Logs

The regressions discussed in this chapter look like

$$\ln Y_i = \alpha + \beta P_i + \sum_j \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + e_i,$$

a repeat of equation (2.2). What's up with $\ln Y_i$ on the left-hand side? Why use logs and not the variable Y_i itself? The answer is easiest to see in a bivariate regression, say,

$$\ln Y_i = \alpha + \beta P_i + e_i, \qquad (2.13)$$

where P_i is a dummy for private school attendance. Because this is a case of regression for dummies, we have

$$E[\ln Y_i | P_i] = \alpha + \beta P_i.$$

In other words, regression in this case fits the CEF perfectly.

Suppose we engineer a *ceteris paribus* change in P_i for student i. This reveals potential outcome Y_{0i} when $P_i = 0$ and Y_{1i} when $P_i = 1$. Thinking now of equation (2.13) as a model for the log of these potential outcomes, we have

$$\ln Y_{0i} = \alpha + e_i$$
$$\ln Y_{1i} = \alpha + \beta + e_i.$$

The difference in potential outcomes is therefore

$$\ln Y_{1i} - \ln Y_{0i} = \beta. \qquad (2.14)$$

Rearranging further gives

$$\begin{aligned}\beta &= \ln\frac{Y_{1i}}{Y_{0i}} = \ln\left\{1 + \frac{Y_{1i} - Y_{0i}}{Y_{0i}}\right\} \\ &= \ln\{1 + \Delta\%Y_p\} \\ &\approx \Delta\%Y_p,\end{aligned}$$

where $\Delta\%Y_p$ is shorthand for the percentage change in potential outcomes induced by P_i. Calculus tells us that $\ln\{1 + \Delta\%Y_p\}$ is close to $\Delta\%Y_p$, when the latter is small. From this, we conclude that the regression slope in a model with $\ln Y_i$ on the left-hand side gives the approximate percentage change in Y_i generated by changing the corresponding regressor.

To calculate the exact percentage change generated by changing P_i, exponentiate both sides of equation (2.14)

$$\frac{Y_{1i}}{Y_{0i}} = \exp(\beta),$$

so

$$\frac{Y_{1i} - Y_{0i}}{Y_{0i}} = \exp(\beta) - 1.$$

When β is less than about .2, $\exp(\beta) - 1$ and β are close enough to justify reference to the latter as percentage change.[19]

You might hear masters describe regression coefficients from a log-linear model as measuring "log points." This terminology reminds

listeners that the percentage change interpretation is approximate. In general, log points underestimate percentage change, that is,

$$\beta < \exp(\beta) - 1,$$

with the gap between the two growing as β increases. For example, when $\beta = .05$, $\exp(\beta) - 1 = .051$, but when $\beta = .3$, $\exp(\beta) - 1 = .35$.

Regression Standard Errors and Confidence Intervals

Our regression discussion has largely ignored the fact that our data come from samples. As we noted in the appendix to the first chapter, sample regression estimates, like sample means, are subject to sampling variance. Although we imagine the underlying relationship quantified by a regression to be fixed and nonrandom, we expect estimates of this relationship to change when computed in a new sample drawn from the same population. Suppose we're after the relationship between the earnings of college graduates and the types of colleges they've attended. We're unlikely to have data on the entire population of graduates. In practice, therefore, we work with samples drawn from the population of interest. (Even if we had a complete enumeration of the student population in one year, different students will have gone to school in other years.) The data set analyzed to produce the estimates in Tables 2.2–2.5 is one such sample. We would like to quantify the sampling variance associated with these estimates.

Just as with a sample mean, the sampling variance of a regression coefficient is measured by its standard error. In the appendix to Chapter 1, we explained that the standard error of a sample average is

$$SE(\bar{Y}_n) = \frac{\sigma_Y}{\sqrt{n}}.$$

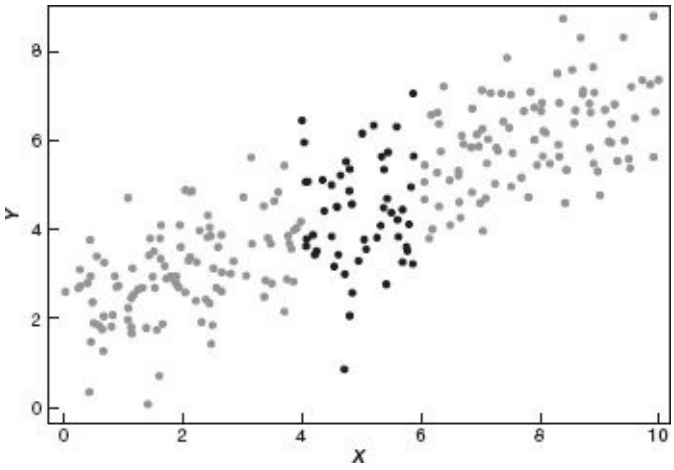
The standard error of the slope estimate in a bivariate regression ($\hat{\beta}$) looks similar and can be written as

$$SE(\hat{\beta}) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_X},$$

where σ_e is the standard deviation of the regression residuals, and σ_X is the standard deviation of the regressor, X_i.

Like the standard error of a sample average, regression standard errors decrease with sample size. Standard errors increase (that is, regression estimates are less precise) when the residual variance is large. This isn't surprising, since a large residual variance means the regression line doesn't fit very well. On the other hand, variability in regressors is good: as σ_X increases, the slope estimate becomes more precise. This is illustrated in Figure 2.2, which shows how adding variability in X_i (specifically, adding the observations plotted in gray) helps pin down the slope linking Y_i and X_i.

FIGURE 2.2
Variance in X is good



The regression anatomy formula for multiple regression carries over to standard errors. In a multivariate model like this,

$$Y_i = \alpha + \sum_{k=1}^{K} \beta_k X_{ki} + e_i,$$

the standard error for the *k*th sample slope, $\hat{\beta}_k$, is

$$SE(\hat{\beta}_k) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_{\tilde{X}_k}}, \qquad (2.15)$$

where $\sigma_{\tilde{X}_k}$ is the standard deviation of $\tilde{X}_{ki}$, the residual from a regression of X_{ki} on all other regressors. The addition of controls has two opposing effects on $SE(\hat{\beta}_k)$. The residual variance (σ_e in the numerator of the standard error formula) falls when covariates that predict Y_i are added to the regression. On the other hand, the standard deviation of $\tilde{X}_{ki}$ in the denominator of the standard error formula is less than the standard deviation of X_{ki}, increasing the standard error. Additional covariates explain some of the variation in other regressors, and this variation is removed by virtue of regression anatomy. The upshot of these changes to top and bottom can be either an increase or decrease in precision.

Standard errors computed using equation (2.15) are nowadays considered old-fashioned and are not often seen in public. The old-fashioned formula is derived assuming the variance of residuals is unrelated to regressors—a scenario that masters call *homoskedasticity*. Homoskedastic residuals can make regression estimates a statistically efficient matchmaker. However, because the homoskedasticity assumption may not be satisfied, kids today rock a more complicated calculation known as *robust standard errors*.

The robust standard error formula can be written as

$$RSE(\hat{\beta}) = \frac{1}{\sqrt{n}} \frac{V(\tilde{X}_{ki} e_i)}{(\sigma^2_{\tilde{X}_k})^2}. \qquad (2.16)$$

Robust standard errors allow for the possibility that the regression line fits more or less well for different values of X_i, a scenario known as *heteroskedasticity*. If the residuals turn out to be homoskedastic after all, the robust numerator simplifies:

$$V(\tilde{X}_{ki} e_i) = V(\tilde{X}_{ki}) V(e_i) = \sigma^2_{\tilde{X}_k} \sigma^2_e.$$

In this case, estimates of $RSE(\hat{\beta})$ should be close to estimates of $SE(\hat{\beta})$, since the theoretical standard errors are then identical. But if residuals are indeed heteroskedastic, estimates of $RSE(\hat{\beta})$ usually provide a more accurate (and typically somewhat larger) measure of sampling variance.[20]

[1] SAT scores here are from the pre-2005 SAT. Pre-2005 total scores add math and verbal scores, each of which range from 0 to 800, so the combined maximum is 1,600.

[2] Stacy Berg Dale and Alan B. Krueger, "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," *Quarterly Journal of Economics,* vol. 117, no. 4, November 2002, pages 1491–1527.

[3] Which isn't to say they are never fooled. Adam Wheeler faked his way into Harvard with doctored transcripts and board scores in 2007. His fakery notwithstanding, Adam managed to earn mostly As and Bs at Harvard before his scheme was uncovered (John R. Ellement and Tracy Jan, "Ex-Harvard Student Accused of Living a Lie," *The Boston Globe,* May 18, 2010).

[4] When data fall into one of J groups, we need $J - 1$ dummies for a full description of the groups. The category for which no dummy is coded is called the *reference group*.

[5] "Ordinary-ness" here refers to the fact that OLS weights each observation in this sum of squares equally. We discuss weighted least squares estimation in Chapter 5.

[6] Our book, *Mostly Harmless Econometrics* (Princeton University Press, 2009), discusses regression-weighting schemes in more detail.

[7] Barron's classifies colleges as Most Competitive, Highly Competitive, Very Competitive, Competitive, Less Competitive, and Noncompetitive, according to the class rank of enrolled students and the proportion of applicants admitted.

[8] Other controls in the empirical model include dummies for female students, student race, athletes, and a dummy for those who graduated in the top 10% of their high school class. These variables are not written out in equation (2.2).

[9] Dale and Krueger, "Estimating the Payoff to Attending a More Selective College," *Quarterly Journal of Economics,* 2002.

[10] The group dummies in (2.4), θ_j, are read "theta-j."

[11] This coefficient is read "lambda."

[12] Joseph Altonji, Todd Elder, and Christopher Taber formalize the notion that the OVB associated with the regressors you have at hand provides a guide to the OVB generated by those you don't. For details, see their study "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, vol. 113, no. 1, February 2005, pages 151–184.

[13] Francis Galton, "Regression towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute of Great Britain and Ireland,* vol. 15, 1886, pages 246–263.

[14] George Udny Yule, "An Investigation into the Causes of Changes in Pauperism in England, Chiefly during the Last Two Intercensal Decades," *Journal of the Royal Statistical Society,* vol. 62, no. 2, June 1899, pages 249–295.

[15] For a more detailed explanation, see Chapter 3 of Angrist and Pischke, *Mostly Harmless*

Econometrics, 2009.

[16] The thing inside braces here, $E[Y_i|X_i] - E[Y_i|X_i - 1]$, is a function of X_i, and so, like the variable X_i, it has an expectation.

[17] The term "bivariate" comes from the fact that two variables are involved, one dependent, on the left-hand side, and one regressor, on the right. *Multivariate* regression models add regressors to this basic setup.

[18] The regression anatomy formula is derived similarly, hence we show the steps only for OVB.

[19] The percentage change interpretation of regression models built with logs does not require a link with potential outcomes, but it's easier to explain in the context of models with such a link.

[20] The distinction between robust and old-fashioned standard errors for regression estimates parallels the distinction (noted in the appendix to Chapter 1) between standard error estimators for the difference in two means that use separate or common estimates of σ_Y^2 for the variance of data from treatment and control groups.