Chapter 3

Making Regression Make Sense

'Let us think the unthinkable, let us do the undoable.

Let us prepare to grapple with the ineffable itself,

and see if we may not eff it after all.'

Douglas Adams, *Dirk Gently's Holistic Detective Agency* (1990)

Angrist recounts:

> I ran my first regression in the summer of 1979 between my freshman and sophomore years as a student at Oberlin College. I was working as a research assistant for Allan Meltzer and Scott Richard, faculty members at Carnegie-Mellon University, near my house in Pittsburgh. I was still mostly interested in a career in special education, and had planned to go back to work as an orderly in a state mental hospital, my previous summer job. But Econ 101 had got me thinking, and I could also see that at the same wage rate, a research assistant's hours and working conditions were better than those of a hospital orderly. My research assistant duties included data collection and regression analysis, though I did not understand regression or even statistics at the time.
>
> The paper I was working on that summer (Meltzer and Richard, 1983), is an attempt to link the size of governments in democracies, measured as government expenditure over GDP, to income inequality. Most income distributions have a long right tail, which means that average income tends to be way above the median. When inequality grows, more voters find themselves with below-average incomes. Annoyed by this, those with incomes between the median and the average may join those with incomes below the median in voting for fiscal policies which - following Robin Hood - take from the rich and give to the poor. The size of government consequently increases.
>
> I absorbed the basic theory behind the Meltzer and Richards project, though I didn't find it

> all that plausible, since voter turnout is low for the poor. I also remember arguing with Alan Meltzer over whether government expenditure on education should be classified as a public good (something that benefits everyone in society as well as those directly affected) or a private good publicly supplied, and therefore a form of redistribution like welfare. You might say this project marked the beginning of my interest in the social returns to education, a topic I went back to with more enthusiasm and understanding in Acemoglu and Angrist (2000).
>
> Today, I understand the Meltzer and Richard (1983) study as an attempt to use regression to uncover and quantify an interesting causal relation. At the time, however, I was purely a regression mechanic. Sometimes I found the RA work depressing. Days would go by where I didn't talk to anybody but my bosses and the occasional Carnegie-Mellon Ph.D. student, most of whom spoke little English anyway. The best part of the job was lunch with Alan Meltzer, a distinguished scholar and a patient and good-natured supervisor, who was happy to chat while we ate the contents of our brown-bags (this did not take long as Allan ate little and I ate fast). I remember asking Allan whether he found it satisfying to spend his days perusing regression output, which then came on reams of double-wide green-bar paper. Meltzer laughed and said there was nothing he would rather be doing.

Now, we too spend our days (at least, the good ones) happily perusing regression output, in the manner of our teachers and advisors in college and graduate school. This chapter explains why.

3.1 Regression Fundamentals

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

Setting aside the relatively abstract causality problem for the moment, we start with the mechanical properties of regression estimates. These are universal features of the population regression vector and its sample analog that have nothing to do with a researcher's interpretation of his output. This chapter begins by reviewing these properties, which include:

(i) the intimate connection between the population regression function and the conditional expectation function

(ii) how and why regression coefficients change as covariates are added or removed from the model

(iii) the close link between regression and other "control strategies" such as matching

(iv) the sampling distribution of regression estimates

3.1.1 Economic Relationships and the Conditional Expectation Function

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

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

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

If Y_i is discrete, $E\left[\text{Y}_i|\text{X}_i = x\right]$ equals the sum $\sum_t t f_y\left(t|\text{X}_i = x\right)$.

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

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

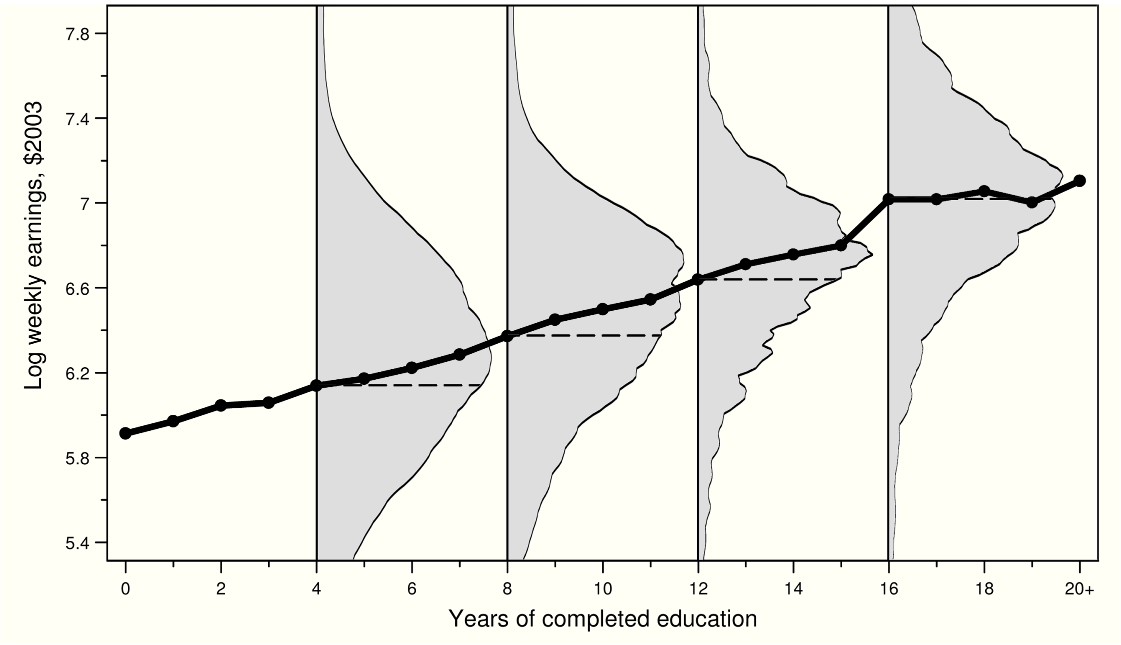


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

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

$$E\left[\mathrm{Y}_i\right] = E\{E\left[\mathrm{Y}_i|\mathrm{X}_i\right]\}, \tag{3.1.1}$$

where the outer expectation uses the distribution of X_i. Here is proof of the law of iterated expectations for continuously distributed $(\mathrm{X}_i, \mathrm{Y}_i)$ with joint density $f_{xy}\left(u,t\right)$, where $f_y\left(t|\mathrm{X}_i = x\right)$ is the conditional

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

distribution of Y_i given $\text{X}_i = x$ and $g_y(t)$ and $g_x(u)$ are the marginal densities:

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

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

The power of the law of iterated expectations comes from the way it breaks a random variable into two pieces.

Theorem 3.1.1 *The CEF-Decomposition Property*

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

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

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

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

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

Theorem 3.1.2 *The CEF-Prediction Property.*

Let $m(\text{X}_i)$ be any function of X_i. The CEF solves

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

so it is the MMSE predictor of Y_i given X_i.

Proof. Write

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

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

A final property of the CEF, closely related to both the CEF decomposition and prediction properties, is the Analysis-of-Variance (ANOVA) Theorem:

Theorem 3.1.3 *The ANOVA Theorem*

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

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

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

$$E\left[\varepsilon_i^2\right] = E\left[E\left[\varepsilon_i^2|\mathrm{X}_i\right]\right] = E[V[\mathrm{Y}_i|\mathrm{X}_i]]$$

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

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

3.1.2 Linear Regression and the CEF

So what's the regression you want to run?

In our world, this question or one like it is heard almost every day. Regression estimates provide a valuable baseline for almost all empirical research because regression is tightly linked to the CEF, and the CEF provides a natural summary of empirical relationships. The link between regression functions – i.e., the best-fitting line generated by minimizing expected squared errors – and the CEF can be explained in at least 3 ways. To lay out these explanations precisely, it helps to be precise about the regression function we have in mind. This chapter is concerned with the vector of *population* regression coefficients, defined as the solution to a population least squares problem. At this point, we are not worried about causality. Rather, we let the $\text{K}\times 1$ regression coefficient vector β be defined by solving

$$\beta = \underset{b}{\arg\min} E\left[\left(\text{Y}_i - \text{X}_i'b\right)^2\right]. \tag{3.1.2}$$

Using the first-order condition,

$$E\left[\text{X}_i\left(\text{Y}_i - \text{X}_i'b\right)\right] = 0.$$

the solution for b can be written $\beta = E\left[\text{X}_i\text{X}_i'\right]^{-1}E\left[\text{X}_i\text{Y}_i\right]$. Note that by construction, $E\left[\text{X}_i\left(\text{Y}_i - \text{X}_i'\beta\right)\right] = 0$. In other words, the population residual, which we *define* as $\text{Y}_i - \text{X}_i'\beta = e_i$, is uncorrelated with the regressors, X_i. It bears emphasizing that this error term does not have a life of its own. It owes its existence and meaning to β.

In the simple bivariate case where the regression vector includes only the single regressor, x_i, and a constant, the slope coefficient is $\beta_1 = \frac{Cov(\text{Y}_i, x_i)}{V(x_i)}$, and the intercept is $\alpha = E\left[\text{Y}_i\right] - \beta_1 E\left[\text{X}_i\right]$. In the multivariate case, i.e., with more than one non-constant regressor, the slope coefficient for the k-th regressor is given below:

REGRESSION ANATOMY

$$\beta_k = \frac{Cov\left(\text{Y}_i, \tilde{x}_{ki}\right)}{V\left(\tilde{x}_{ki}\right)}, \tag{3.1.3}$$

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

In other words, $E\left[\text{X}_i\text{X}_i'\right]^{-1}E\left[\text{X}_i\text{Y}_i\right]$ is the $\text{K}\times 1$ vector with k-th element $\frac{Cov(\text{Y}_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$. This important formula is said to describe the “anatomy of a multivariate regression coefficient” because it reveals much more than the matrix formula $\beta = E\left[\text{X}_i\text{X}_i'\right]^{-1}E\left[\text{X}_i\text{Y}_i\right]$. It shows us that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after "partialling out" all the other variables in the model.

To verify the regression-anatomy formula, substitute

$$\text{Y}_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + ... + \beta_\text{K} x_{\text{K}i} + e_i$$

in the numerator of (3.1.3). Since $\tilde{x}_{ki}$ is a linear combination of the regressors, it is uncorrelated with e_i. Also, since $\tilde{x}_{ki}$ is a residual from a regression on all the other covariates in the model, it must be uncorrelated these covariates. Finally, for the same reason, the covariance of $\tilde{x}_{ki}$ with x_{ki} is just the variance of $\tilde{x}_{ki}$. We therefore have that $Cov\left(\text{Y}_i, \tilde{x}_{ki}\right) = \beta_k V\left(\tilde{x}_{ki}\right)$.[2]

The regression-anatomy formula is probably familiar to you from a regression or statistics course, perhaps with one twist: the regression coefficients defined in this section are not estimators, but rather they are non-stochastic features of the joint distribution of dependent and independent variables. The joint distribution is what you would observe if you had a complete enumeration of the population of interest (or knew the stochastic process generating the data). You probably don't have such information. Still, it's kosher—even desirable—to think about what a set of population parameters might mean, without initially worrying about how to estimate them.

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

Theorem 3.1.4 *The Linear CEF Theorem (Regression-justification I)*

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

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

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

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

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

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

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

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

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

Theorem 3.1.5 *The Best Linear Predictor Theorem (Regression-justification II)*

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

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

In other words, just as the CEF, $E\left[\mathrm{Y}_i|\mathrm{X}_i\right]$, is the best (i.e., MMSE) predictor of Y_i given X_i in the class of *all* functions of X_i, the population regression function is the best we can do in the class of *linear* functions.

Theorem 3.1.6 *The Regression-CEF Theorem (Regression-justification III)*

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

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

Proof. Write

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

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

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

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

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

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

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

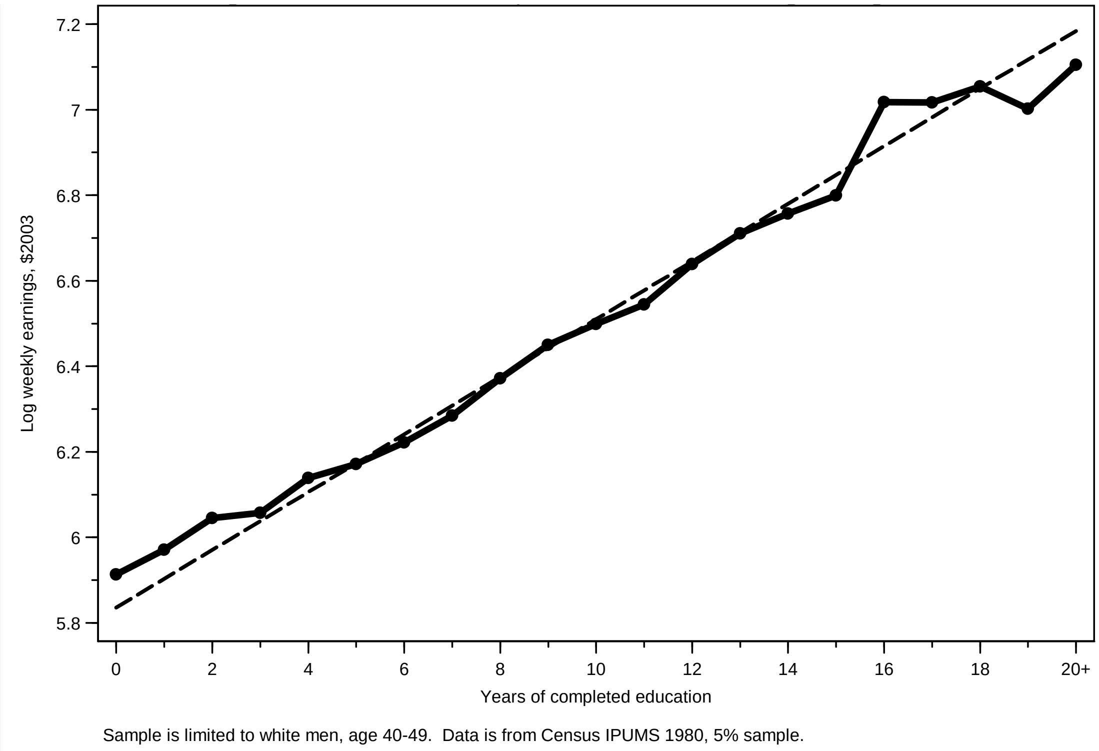
$$\beta = E[\text{X}_i\text{X}_i']^{-1}E[\text{X}_i\text{Y}_i] = E[\text{X}_i\text{X}_i']^{-1}E[\text{X}_iE(\text{Y}_i|\text{X}_i)]. \tag{3.1.5}$$

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

An illustration of the grouped-data approach to regression appears below. We estimated the schooling coefficient in a wage equation using 21 conditional means, the sample CEF of earnings given schooling. As the Stata output reported here shows, a grouped-data regression, weighted by the number of individuals at each schooling level in the sample, produces coefficients *identical* to what would be obtained using the underlying microdata sample with hundreds of thousands of observations. Note, however, that the standard errors from the grouped regression do not correctly reflect the asymptotic sampling variance of the slope estimate in repeated *micro-data* samples; for that you need an estimate of the variance of $\text{Y}_i - \text{X}_i'\beta$. This variance depends on the microdata, in particular, the second-moments of $W_i \equiv \begin{bmatrix} \text{Y}_i; & \text{X}_i' \end{bmatrix}'$, a point we elaborate on in the next section.

3.1.3 Asymptotic OLS Inference

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



Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

estimates should be interpreted. Whatever a regression coefficient may mean, it has a sampling distribution that is easy to describe and use for statistical inference.[3]

We are interested in the distribution of the sample analog of

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

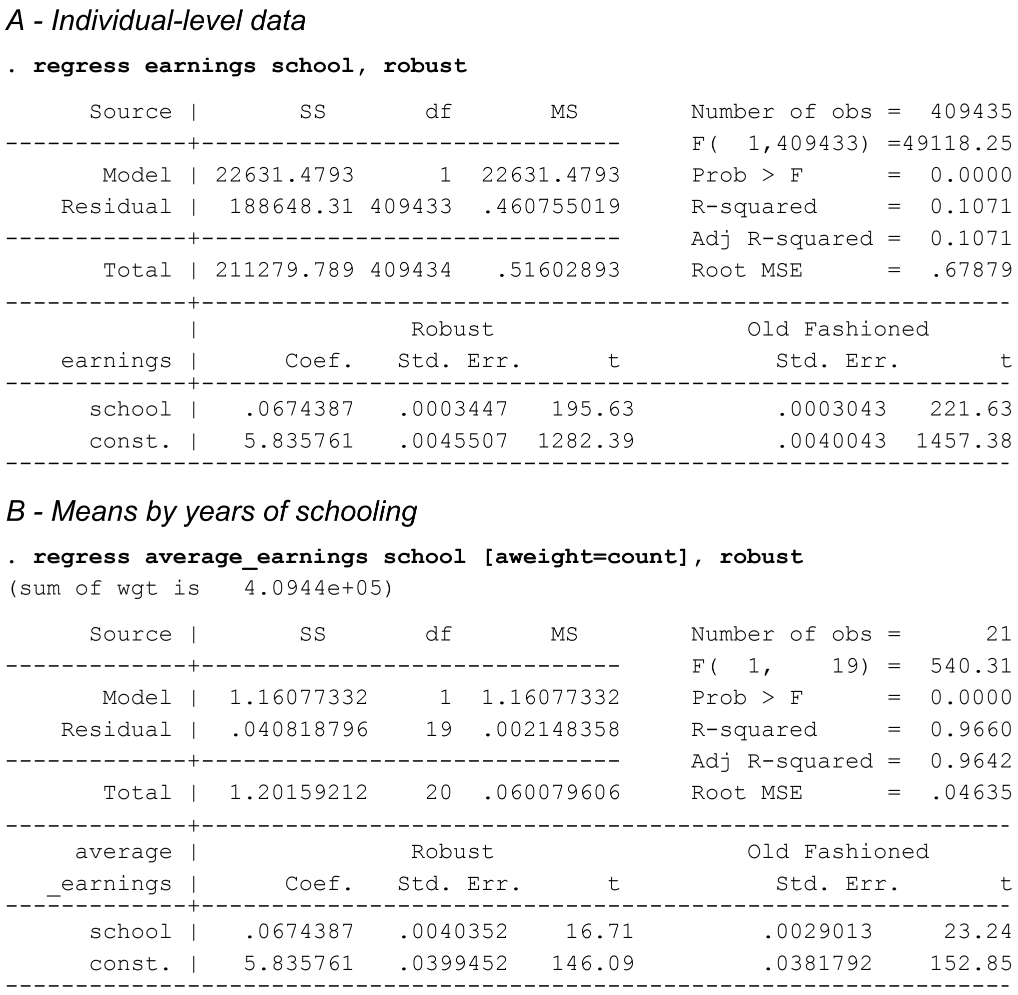
in repeated samples. Suppose the vector $W_i \equiv \begin{bmatrix} \mathrm{Y}_i; & \mathrm{X}_i' \end{bmatrix}'$ is independently and identically distributed in a sample of size N. A natural estimator of the first population moment, $E[W_i]$, is the sum, $\frac{1}{N}\sum_{i=1}^{N} W_i$. By the law of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

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

Although we derived $\hat{\beta}$ as a method of moments estimator, it is called the OLS estimator of β because it solves the sample analog of the least-squares problem described at the beginning of Section 3.1.2.[4]

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

[4] Econometricians like to use matrices because the notation is so compact. Sometimes (not very often) we do too. Suppose



A - Individual-level data

```
. regress earnings school, robust

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

B - Means by years of schooling

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

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

Figure 3.1.3: Micro-data and grouped-data estimates of returns to schooling. Source: 1980 Census - IPUMS, 5 percent sample. Sample is limited to white men, age 40-49. Derived from Stata regression output. Old-fashioned standard errors are the default reported. Robust standard errors are heteroscedasticity-consistent. Panel A uses individual-level data. Panel B uses earnings averaged by years of schooling.

The asymptotic sampling distribution of $\hat{\beta}$ depends solely on the definition of the estimand (i.e., the nature of the thing we're trying to estimate, β) and the assumption that the data constitute a random sample. Before deriving this distribution, it helps to record the general asymptotic distribution theory that covers our needs. This basic theory can be stated mostly in words. For the purposes of these statements, we assume the reader is familiar with the core terms and concepts of statistical theory (e.g., moments, mathematical expectation, probability limits, and asymptotic distributions). For definitions of these terms and a formal mathematical statement of the theoretical propositions given below, see, e.g., Knight (2000).

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

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

SLUTSKY'S THEOREM

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

(b) Consider the product of two random variables, one of which converges in distribution and the other converges in probability to a constant: the asymptotic distribution of this product is unaffected by replacing the one that converges to a constant by this constant. This allows us to replaces some sample moments by population moments (i.e., by their probability limits) when deriving distributions. Formally, let a_N be a statistic with a limiting distribution and let b_N be a statistic with probability limit b. Then $a_N b_N$ and $a_N b$ have the same asymptotic distribution.

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

X is the matrix whose rows are given by X_i' and y is the vector with elements y_i, for $i = 1, ..., N$. The sample moment $\frac{1}{N}\sum X_i X_i'$ is $X'X/N$ and the sample moment $\frac{1}{N}\sum X_i y_i$ is $X'y/N$. Then we can write $\hat{\beta} = (X'X)^{-1} X'y$, a familiar matrix formula.

THE DELTA METHOD Consider a vector-valued random variable that is asymptotically Normally distributed. Most scalar functions of this random variable are also asymptotically Normally distributed, with covariance matrix given by a quadratic form with the covariance matrix of the random variable on the inside and the gradient of the function evaluated at the probability limit of the random variable on the outside. Formally, the asymptotic distribution of $h(b_N)$ is Normal with covariance matrix $\nabla h(b)'\Omega\nabla h(b)$ where $plim\ b_N = b$, $h(\cdot)$ is continuously differentiable at b with gradient $\nabla h(b)$, and b_N has asymptotic covariance matrix Ω.[5]

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

$$\textsc{y}_i = \mathrm{X}_i'\beta + [\textsc{y}_i - \mathrm{X}_i'\beta] \equiv \mathrm{X}_i'\beta + e_i, \tag{3.1.6}$$

where the residual e_i is *defined* as the difference between the dependent variable and the population regression function, as before. This is as good a place as any to point out that these residuals are uncorrelated with the regressors *by definition of* β. In other words, $E[\mathrm{X}_i e_i] = 0$ is a consequence of $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\textsc{y}_i]$ and $e_i = \textsc{y}_i - \mathrm{X}_i'\beta$, and not an assumption about an underlying economic relation. We return to this important point in the discussion of causal regression models in Section 3.2.[6]

Substituting the identity 3.1.6 for $\textsc{y}_i$ in the formula for $\hat{\beta}$, we have

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

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

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

The standard errors used to construct t-statistics are the square roots of the diagonal elements of this

[5] For a derivation of the the delta method formula using the Slutsky and continuous mapping theorems, see, e.g., Knight, 2000, pp. 120-121.

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

matrix. In practice these standard errors are estimated by substituting sums for expectations, and using the estimated residuals, $\hat{e}_i = \mathrm{Y}_i - \mathrm{X}_i'\hat{\beta}$ to form the empirical fourth moment, $\sum[\mathrm{X}_i\mathrm{X}_i\hat{e}_i^2]/N$.

Asymptotic standard errors computed in this way are known as heteroskedasticity-consistent standard errors, White (1980a) standard errors, or Eicker-White standard errors in recognition of Eicker's (1967) derivation. They are also known as "robust" standard errors (e.g., in Stata). These standard errors are said to be robust because, in large enough samples, they provide accurate hypothesis tests and confidence intervals given minimal assumptions about the data and model. In particular, our derivation of the limiting distribution makes no assumptions other than those needed to ensure that basic statistical results like the central limit theorem go through. These are not, however, the standard errors that you get by default from packaged software. Default standard errors are derived under a homoskedasticity assumption, specifically, that $E[e_i^2|\mathrm{X}_i] = \sigma^2$, a constant. Given this assumption, we have

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

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

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

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

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

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

Therefore, even if $V[\mathrm{Y}_i|\mathrm{X}_i]$ is constant, the residual variance increases with the square of the gap between the regression line and the CEF, a fact noted in White (1980b).[7]

In the same spirit, it's also worth noting that while a linear CEF makes homoskedasticity possible, this is

[7] The cross-product term resulting from an expansion of the quadratic in the middle of 3.1.9 is zero because $\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]$ is mean-independent of X_i.

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

These points of principle notwithstanding, as an empirical matter, heteroskedasticity may matter little. In the micro-data schooling regression depicted in Figure 3.1.3, the robust standard error is .0003447, while the old-fashioned standard error is .0003043, only slightly smaller. The standard errors from the grouped-data regression, which are necessarily heteroskedastic if group sizes differ, change somewhat more; compare the .004 robust standard to the .0029 conventional standard error. Based on our experience, these differences are typical. If heteroskedasticity matters too much, say, more than a 30% increase or any marked decrease in standard errors, you should worry about possible programming errors or other problems (for example, robust standard errors below conventional may be a sign of finite-sample bias in the robust calculation; see Chapter 8, below.)

3.1.4 Saturated Models, Main Effects, and Other Regression Talk

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

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

$$\mathrm{Y}_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + ... + \beta_\tau d_{\tau i} + \varepsilon_i,$$

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

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

while $\beta_0 = E[\mathrm{Y}_i|\mathrm{S}_i = 0]$. In practice, you can pick any value of S_i for the reference group; a regression model is saturated as long as it has one parameter for every possible j in $E[\mathrm{Y}_i|\mathrm{S}_i = j]$. Saturated models fit the

CEF perfectly because the CEF is linear in the dummy regressors used to saturate. This is an important special case of the regression-CEF theorem.

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

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

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

We can label these using the following scheme:

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

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

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

a parameterization with two main effects and one interaction term.[8] The saturated regression equation becomes

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

Finally, we can combine the multi-valued schooling variable with sex to produce a saturated model that

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

has τ main effects for schooling, one main effect for sex, and τ sex-schooling interactions:

$$\text{Y}_i = \beta_0 + \sum_{j=1}^{\tau} \beta_j d_{ji} + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji} x_{2i}) + \varepsilon_i. \tag{3.1.10}$$

The interaction terms, δ_j, tell us how each of the schooling effects differ by sex. The CEF in this case takes on $2(\tau + 1)$ values while the regression has this many parameters.

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

$$\text{Y}_i = \beta_0 + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji} x_{2i}) + \varepsilon_i. \tag{3.1.11}$$

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

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

3.2 Regression and Causality

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

3.2.1 The Conditional Independence Assumption

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

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

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

For starters, think of schooling as a binary decision, like whether Angrist goes to college. Denote this by a dummy variable, C_i. The causal relationship between college attendance and a future outcome like earnings can be described using the same potential-outcomes notation we used to describe experiments in

Chapter 2. To address this question, we imagine two potential earnings variables:

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

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

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

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

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

$$\begin{aligned} \underbrace{E\left[\text{Y}_i|\text{C}_i = 1\right] - E[\text{Y}_i|\text{C}_i = 0]}_{\text{Observed difference in earnings}} \quad = \quad & \underbrace{E[\text{Y}_{1i} - \text{Y}_{0i}|\text{C}_i = 1]}_{\text{average treatment effect on the treated}} \\ & + \underbrace{E\left[\text{Y}_{0i}|\text{C}_i = 1\right] - E\left[\text{Y}_{0i}|\text{C}_i = 0\right]}_{\text{selection bias}}. \end{aligned} \tag{3.2.1}$$

It seems likely that those who go to college would have earned more anyway. If so, selection bias is positive, and the naive comparison, $E\left[\text{Y}_i|\text{C}_i = 1\right] - E[\text{Y}_i|\text{C}_i = 0]$, exaggerates the benefits of college attendance.

The CIA asserts that conditional on observed characteristics, X_i, selection bias disappears. In this example, the CIA says,

$$\{\text{Y}_{0i},\text{Y}_{1i}\} \amalg \text{C}_i|\text{X}_i. \tag{3.2.2}$$

Given the CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation. In other words,

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

Now, we'd like to expand the conditional independence assumption to causal relations that involve variables that can take on more than two values, like years of schooling, S_i. The causal relationship between schooling and earnings is likely to be different for each person. We therefore use the individual-specific notation,

$$\text{Y}_{si} \equiv f_i(s)$$

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

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

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

The CIA in this more general setup becomes

$$Y_{si} \amalg \text{S}_i|\text{X}_i \tag{CIA}$$

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

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

$$\begin{aligned} & E\left[\text{Y}_i|\text{X}_i, \text{S}_i = s\right] - E\left[\text{Y}_i|\text{X}_i, \text{S}_i = s-1\right] \\ = \;& E\left[f_i(s) - f_i(s-1)|\text{X}_i\right] \end{aligned}$$

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

$$E\left[\text{Y}_i|\text{X}_i, \text{S}_i = 12\right] - E\left[\text{Y}_i|\text{X}_i, \text{S}_i = 11\right] = E\left[f_i(12)|\text{X}_i, \text{S}_i = 12\right] - E\left[f_i(11)|\text{X}_i, \text{S}_i = 11\right].$$

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

$$E\left[f_i(12)|\text{X}_i, \text{S}_i = 12\right] - E\left[f_i(11)|\text{X}_i, \text{S}_i = 11\right] = E\left[f_i(12) - f_i(11)|\text{X}_i, \text{S}_i = 12\right].$$

Here, the selection bias term is the average difference in the potential dropout-earnings of high school graduates and dropouts. Given the CIA, however, high school graduation is independent of potential earnings conditional on X_i, so the selection-bias vanishes. Note also that in this case, the causal effect of graduating high school on high school graduates is the population average high school graduation effect:

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

This is important . . . but less important than the elimination of selection bias in (3.2.1).

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

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

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

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

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

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

In practice, there are many details to worry about when implementing a matching strategy. We fill in some of the technical details on the mechanics of matching in Section 3.3.1, below. Here we note that a global drawback of the matching approach is that it is not "automatic," rather it requires two steps, matching and averaging. Estimating the standard errors of the resulting estimates may not be straightforward, either.

A third consideration is that the two-way contrast at the heart of this subsection (high school or college completers versus dropouts) does not do full justice to the problem at hand. Since s_i takes on many values, there are separate average causal effects for each possible increment in s_i, which also must be summarized in some way.[9] These considerations lead us back to regression.

Regression provides an easy-to-use empirical strategy that automatically turns the CIA into causal effects. Two routes can be traced from the CIA to regression. One assumes that $f_i(s)$ is both linear in s and the same for everyone except for an additive error term, in which case linear regression is a natural tool to estimate the features of $f_i(s)$. A more general but somewhat longer route recognizes that $f_i(s)$ almost certainly differs for different people, and, moreover, need not be linear in s. Even so, allowing for random variation in $f_i(s)$ across people, and for non-linearity for a given person, regression can be thought of as strategy for the estimation of a weighted average of the individual-specific difference, $f_i(s) - f_i(s-1)$. In fact, regression can be seen as a particular sort of matching estimator, capturing an average causal effect much like 3.2.3 or 3.2.5.

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

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

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

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

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

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

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

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

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

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

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

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

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

Because mean-independence implies orthogonality, the residual in the linear causal model

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

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

3.2.2 The Omitted Variables Bias Formula

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

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

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

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

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

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

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

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

This formula is easy to derive: plug the long regression into the short regression formula, $\frac{Cov(\text{Y}_i,\text{s}_i)}{V(\text{s}_i)}$. Not surprisingly, the OVB formula is closely related to the regression anatomy formula, 3.1.3, from Section 3.1.2. Both the OVB and regression anatomy formulas tell us that short and long regression coefficients are the same whenever the omitted and included variables are uncorrelated.[10]

We can use the OVB formula to get a sense of the likely consequences of omitting ability for schooling coefficients. Ability variables have positive effects on wages, and these variables are also likely to be positively correlated with schooling. The short regression coefficient may therefore be "too big" relative to what we want. On the other hand, as a matter of economic theory, the direction of the correlation between schooling and ability is not entirely clear. Some omitted variables may be negatively correlated with schooling, in which case the short regression coefficient will be too small.[11]

[10] Here is the multivariate generalization of OVB: Let β_1^s denote the coefficient vector on a $\text{K}_1 \times 1$ vector of variables, X_{1i} in a (short) regression that has no other variables and let β_1^l denote the coefficient vector on these variables in a (long) regression that includes a $\text{K}_2 \times 1$ vector of control variables, X_{2i}, with coefficient vector β_2^l. Then $\beta_1^s = \beta_1^l + E[X_{1i}X_{1i}']^{-1}E[X_{1i}X_{2i}']\beta_2^l$.

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

Table 3.2.1 illustrates these points using data from the NLSY. The first three entries in the table show that the schooling coefficient decreases from .132 to .114 when family background variables—in this case, parents' education—as well as a few basic demographic characteristics (age, race, census region of residence) are included as controls. Further control for individual ability, as proxied by the Armed Forces Qualification Test (AFQT) test score, reduces the schooling coefficient to .087 (AFQT is used by the military to select soldiers). The omitted variables bias formula tells us that these reductions are a result of the fact that the additional controls are positively correlated with both wages and schooling.[12]

Table 3.2.1: Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupation dummies
	0.132 (0.007)	0.131 (0.007)	0.114 (0.007)	0.087 (0.009)	0.066 (0.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2434.

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

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

At this point, it's worth considering when the CIA is most likely to give a plausible basis for empirical work. The best-case scenario is random assignment of s_i , conditional on X_i, in some sort of (possibly natural) experiment. An example is the study of a mandatory re-training program for unemployed workers by Black, *et al.* (2003). The authors of this study were interested in whether the re-training program succeeded in raising earnings later on. They exploit the fact that eligibility for the training program they study was determined on the basis of personal characteristics and past unemployment and job histories. Workers were divided up into groups on the basis of these characteristics. While some of these groups of workers were ineligible for training, those in other groups were required to take training if they did not take

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

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

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

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

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

3.2.3 Bad Control

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

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

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

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