A NON-TECHNICAL INTRODUCTION TO REGRESSIONS

David Romer

University of California, Berkeley

January 2018

CONTENTS

PREFACE

This document provides a brief, very non-technical introduction to the basic econometrics of regressions. It is intended mainly for economics undergraduates who have not yet taken a course in econometrics but who are encountering papers that use regressions. It may also be useful as a "big picture" overview for students who are taking, or have taken, an econometrics course, and to students in other disciplines that use regressions. It is absolutely *not* a substitute for a course in econometrics (such as Economics 140 or 141 at Berkeley). It says nothing about the theory and statistics underlying regressions, and only scratches the surfaces of using and interpreting them. Anyone interested in empirical relationships in economics (or in any other non-experimental setting) should take such a course.

I am indebted to Christina Romer for innumerable helpful conversations that were invaluable in preparing these notes, and to our past Graduate Student Instructors for helpful advice about teaching this material. I am especially grateful to my students for their questions, patience, and encouragement.

I Introduction

Economics is a deeply empirical field. Economists are interested in many questions about the world: What is the impact of monetary policy on the unemployment rate? How does the price of gasoline affect the quantity demanded? How does political instability affect economic growth? What is the effect of wages on the quantity of labor supplied? And much more.

A tool that economists often employ to analyze empirical questions is a *regression*. In essence, a regression is just a way of summarizing the relationship between two or more variables in some set of data. The purpose of these notes is to provide a brief, nontechnical introduction to regressions, with an emphasis on two issues. The most important is when they can and cannot be used to shed light on the *effect* of one variable on another. We will not discuss other possible uses of regressions, such as making predictions. The other is how to interpret the results of a regression.

For concreteness, these notes focus on a specific economic question: What is the effect of the number of years of schooling that an individual obtains on his or her wage?

The implicit assumption behind this question is that an individual's wage depends on his or her years of schooling and other factors. It is helpful to express this assumption in terms of a very simple model of the determinants of individuals' wages:

$$\ln W_i = a + bE_i + u_i. \tag{1}$$

Here *i* indexes individuals, *E* denotes years of education, and *W* denote the wage. Thus, the model says that an individual's wage depends on his or her years of schooling (this is the bE_i piece of [1]). It also says that the wage is affected by other factors (the u_i piece). These other factors might include luck, family connections, inherited characteristics, and abilities acquired through means

other than formal education. Because those other factors are not our focus, the equation lumps them all together.[1]

The fact that the model assumes that the log of the wage is a linear function of E is not critical. We will return to it briefly in Section VII. But for now, one implication is worth noting. Exponentiating both sides of (1) gives us

$$W_i = e^{a}e^{bE_i}e^{u_i}. \tag{2}$$

Thus the model in equation (1) implies that raising E_i by one raises W_i by a factor of e^b. For small values of b, e^b is approximately $1 + b$. For example, if b = 0.1, e^b is 1.105, so that each year of education raises an individual's wage by about 10 percent. b is sometimes referred to as the "rate of return" to education.

If you start to think about it, you are likely to soon find objections to the question, "What is the effect of an individual's years of schooling on his or her wage?" The effects of merely sitting in a classroom for a school year on earnings (as opposed to, for example, paying attention to what the teacher is saying, participating in class, doing the assigned homework, and studying) is probably zero or even negative. Even if we specify that we mean the full range of education provided by schooling rather than just time spent in school, the effects surely vary—perhaps greatly—depending on the quality of the school, the ability and motivation of the student, and the fit between the school and the student.

[1] It would be more precise to say that the effects of the other factors are shown by $a + u_i$ rather than just by u_i. The reason for including a is that in working through the mathematics of regressions, it is often helpful to assume that the u_i term has a mean of zero. Including a allows us to assume this without assuming that the mean effect of other factors ($a + u_i$) is zero. a is referred to as the constant term in the model. These notes generally include constant terms in the equations but say little about them.

Thus, it is helpful to be more precise about our question. A better formulation is: What is the average effect, for students who could plausibly get slightly more or slightly less education, of an additional year of education that is of average quality in the relevant setting (the United States, for example)?

In principle, one way to answer this question would be to run a *randomized experiment*. The first step would be to identify some students who were close to the margin of getting more or less education. For example, suppose we could find some high school seniors who planned to enroll in a four-year college if they were admitted to their local university and in a junior college if they were not, and who were close to the cutoff for admission to the university. Or suppose we could find some 16-year olds agonizing over whether to complete high school, and who could easily be swayed by minor factors, such as whether they happened to talk to a parent or a friend first about the decision. The next step would be to randomly divide the students into two groups, and require the members of one group to get the larger amount of education and the other to get the smaller amount. The final step would be to wait a decade or so and collect data on the students. We could then estimate the effect of an additional year of schooling on wages—the parameter b in equation (1)—by dividing the difference in the average wage of the two groups (or, more precisely, the difference in the average log wage of the two groups) by the difference in the years of schooling in the two groups.

Of course, conducting such an experiment is impossible (as well as grossly immoral). Thus, the challenge that we face in trying to figure out how schooling affects wages is to use data that are not coming from an experiment to answer that question.

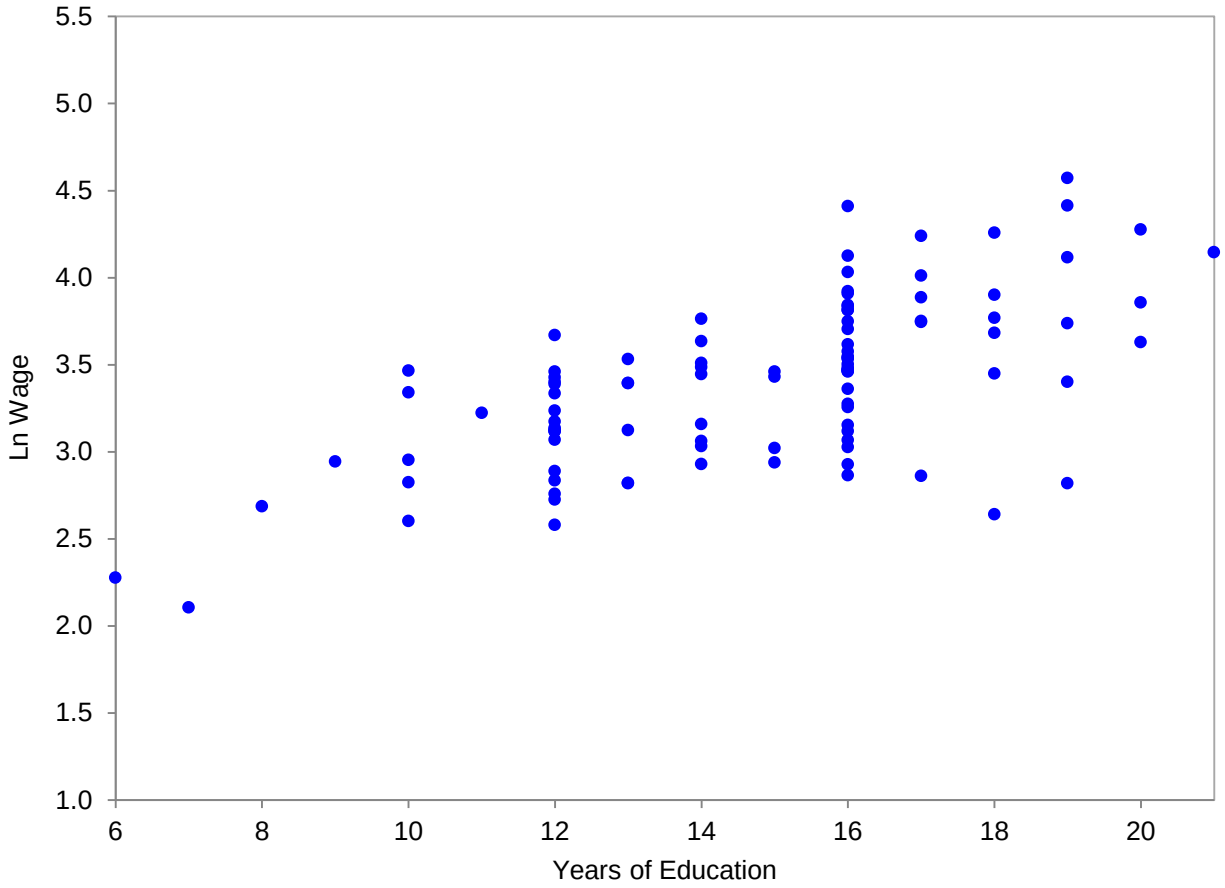

Figure 1. Scatterplot of Log Wages against Years of Education

II Ordinary Least Squares Regression

When we cannot run an experiment, a natural possibility is to ask what the relationship in the data is. Figure 1 is a scatterplot of the log wage against years of education in one dataset. That is, each point shows years of schooling and the log wage for one person. (Importantly, all the "data" presented in these notes are artificially generated, not real data. That is helpful for explaining the basic issues that arise with regressions, but leaves out the complications that always occur with any actual data. Section VII discusses a few of those complications.) Faced with these data, we can ask: Looking across these individuals, on average how much are wages higher when

schooling is one year higher?

A regression—technically, an ordinary least squares, or OLS, regression—is the usual tool that economists use to summarize such relationships. What an OLS regression does is find the straight line that fits the data best.[2]

There is some terminology associated with regression equations such as (1). The variable on the left-hand side of the equation ($\ln W_i$ in our case) is the left-hand side or dependent variable. The variable on the right-hand side (E_i) is the right-hand side or independent variable. The parameters (a and b) are the coefficients. The coefficient corresponding to the intercept of the linear relationship (a) is the constant. And the term capturing the factors that are not being explicitly modeled (u_i) is the residual.

The results of using an OLS regression to estimate equation (1) with the data shown in Figure 1 are:

$$\ln W_i = \underset{(0.19)}{1.87} + \underset{(0.012)}{0.104}\, E_i, \qquad N = 100. \tag{3}$$

Here, the numbers in the top line (1.87 and 0.104) are the coefficient estimates, $\hat{a}$ and $\hat{b}$, from the regression. That is, they are the intercept and slope of the straight line that best fits the data in Figure 1. The numbers in parentheses below the coefficient estimates are the standard errors of $\hat{a}$ and $\hat{b}$. "N" is the number of observations in our sample. For the moment, we will focus just on the

[2] Concretely—but not importantly for our purposes—a straight line in our diagram takes the form $\hat{a} + \hat{b}E$. Given such a line, for each individual in the sample there will generally be some difference between their actual log wage, $\ln W_i$, and the value that the equation predicts for them, $\hat{a} + \hat{b}E_i$. What OLS does is find the values of $\hat{a}$ and $\hat{b}$ that make the sum of the squares of these differences as small as possible (hence the name, "least squares"). To learn about the conditions under which it makes sense to choose $\hat{a}$ and $\hat{b}$ to minimize the sum of squared differences, you will need to take an econometrics course.

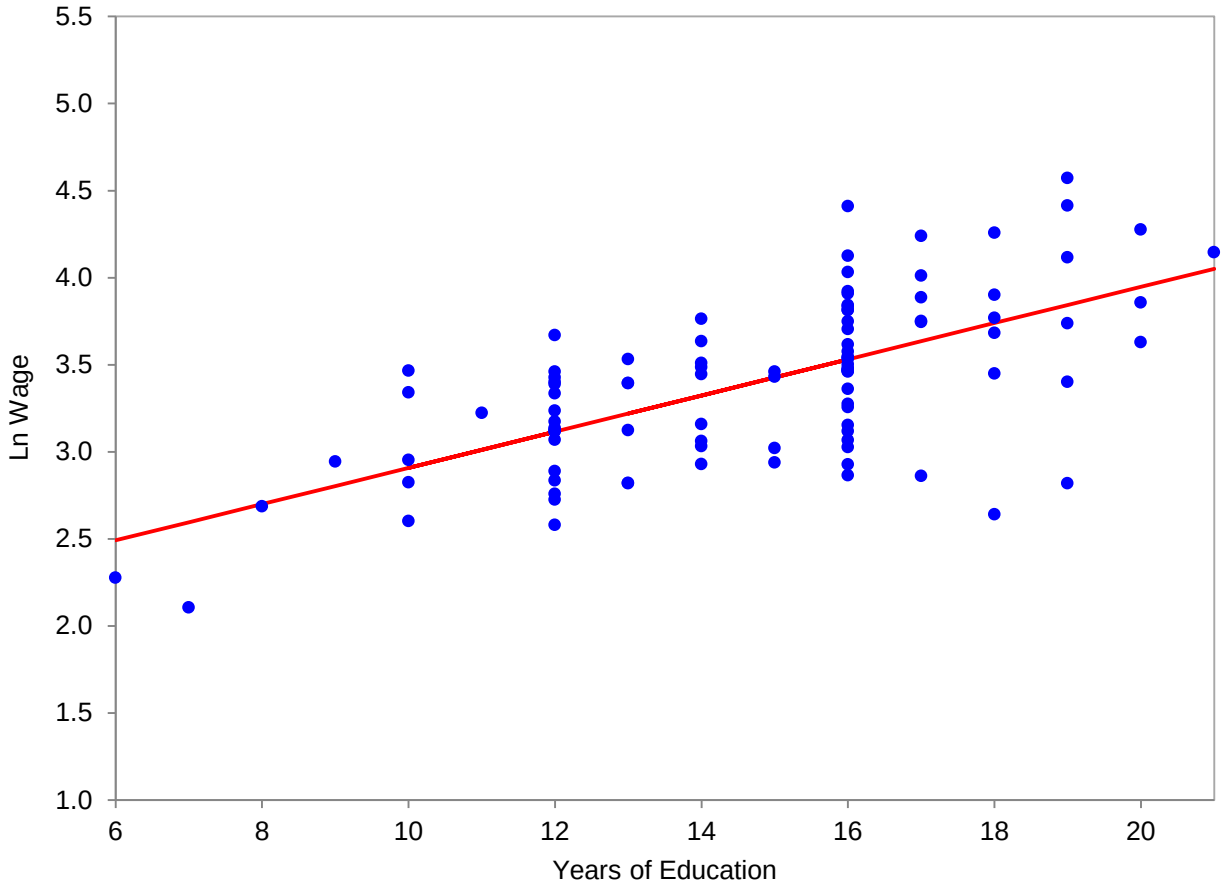

Figure 2. Regression Line and Scatterplot of Log Wages against Years of Education

coefficient estimates. We will come back to the standard errors in Section V. Figure 2 redraws Figure 1 with the regression line, $\ln W_i = 1.87 + 0.104E_i$, added.

III Correlation Is Not Causation

Equation (3) tells us that in our dataset, when one individual has one more year of schooling than another, his or her log wage is on average higher by 0.104; that is, he or she earns on average about 10 percent more. We can therefore say that in our data, one more year of education is on average associated with an individual's wage being higher by about 10 percent.

Can we infer from this that these data are telling us that a good estimate of b in equation (1) is 0.104? That is, can we go from the statement that the data indicate that one more year of education *is associated with* the wage being on average 10 percent higher to the statement that the data indicate that one more year of education *causes* the wage to be on average 10 percent higher?

The answer to this question is emphatically no. All the regression does is summarize patterns or correlations in the data. If there is a relationship between two variables, a regression will show the relationship; but it does not tell us that one variable causes the other, or that causation goes in the opposite direction, or that a third variable is causing both—or some combination.

In equation (1), b is the amount that an additional year of education changes an individual's wage, holding other influences on wages the same. But there is no guarantee that in practice, factors other than education that affect wages are not on average systematically different among individuals with more education. In fact, if you think about it, it is easy to think of reasons that other influences on wages might vary systematically with the amount of education people receive. For example, individuals raised by well educated, wealthy parents are likely to obtain more education, but to have other advantages in the labor market—perhaps from more learning at home, or from more job opportunities through family connections. Individuals who are naturally more disciplined may get more education, but also be more productive workers for a given amount of education. Individuals who are healthier may be able to stay in school longer, but also to work harder when they get a job. And so on.

To see the problem more fully, go back to our model of what determines wages, equation (1), $\ln W_i = a + bE_i + u_i$. Recall that u_i reflects all factors other than the number of years of schooling that affect an individual's wage. In the examples we just discussed, such as individuals

who are more disciplined both getting more education and earning more for a given amount of education, there is a positive correlation between the right-hand side variable in equation (1), E_i, and the residual, u_i. That is, u_i is on average higher when E_i is higher. When there is such positive correlation, the coefficient estimate from an OLS regression will overstate the true effect of the right-hand side variable on the left-hand side one. The technical term for this is that the estimate of b from the regression is a *biased* estimate of the true value of b—that is, of the effect of education on wages.

To understand why positive correlation between the right-hand side variable and the residuals leads OLS to overstate the effect of the right-hand side variable on the left-hand side one, suppose that the true value of b is zero—that is, that schooling has no impact at all on earnings—but that u_i is on average higher among individuals with more years of schooling. Thus, individuals who have more years of education on average earn more not because they get any benefit in the labor market from their additional education, but because they on average have other characteristics that cause them to earn more. What OLS does is choose the coefficient estimates, $\hat{a}$ and $\hat{b}$, to fit the data as well as possible. Since individuals with more education on average earn more, OLS will choose a positive value of $\hat{b}$. That is, the coefficient estimate will overstate the true effect of education on wages. And this same logic carries over to cases where the true value of b is not zero. If the true value of b is positive, for example, positive correlation between E_i and u_i will cause OLS to give a value of $\hat{b}$ that is larger than the true value of b.

The name for this problem is *omitted variable bias*. All the factors that influence wages that are not included in the model are put into the u_i term in (1); they are "omitted variables." If their overall impact on the dependent variable (that is, u_i) is systematically positively correlated with the right-hand side variable (E_i), the regression coefficient, $\hat{b}$, will systematically overstate

the true effect of the right-hand side variable on the left-hand side variable, *b*—that is, the coefficient estimate from the regression will be a biased estimate of the true coefficient.[3]

Note that the source of the problem is *not* just that there are omitted variables. Any empirical model will be incomplete, and so will have omitted variables. The problem arises only if there is systematic correlation between the omitted variables and the right-hand side variables. For example, think about the randomized experiment described in Section I. That experiment leaves out a huge amount of information about the determinants of individuals' wages—all it considers is the variation coming from the fact we intervened to require a few individuals to get more education than they might have gotten otherwise, and a few to get less. But because that intervention was random, we know that it is not systematically correlated with the other influences on education. Thus, the experiment does not suffer from omitted variable bias.[4]

Note also it is possible for omitted variable bias to cause the regression coefficient to systematically understate rather than overstate the true effect of schooling on wages. Consider again the case where the true value of *b* is zero. But suppose we are in a setting where the individuals who get more education are ones that have fewer skills that are valued in the labor market; for example, they might lack social skills, or be interested mainly in arcane questions with little practical relevance. In that environment, we might find that even though education had no effect on wages, individuals with more education on average earned less. That is, in this case the correlation between the residual (u_i) and the right-hand side variable (E_i) is negative, and so the

[3] "Systematic" correlation means correlation that is not just due to chance (technically, correlation that would not approach zero as the sample size became larger and larger). For now, we are discussing only systematic correlation, not chance correlation. We will discuss some issues raised by chance correlation in Section V.

[4] Again, our focus for the moment is on issues raised by systematic correlation. Because the assignment is random, it could happen to be correlated, either positively or negatively, with other influences just by chance. As a result, the estimate of *b* we get from the experiment will generally not be exactly equal to the true value of *b*. But it will not be systematically biased either above or below the true value. Again, we will come back to the issue of chance correlation in Section V.

regression estimate systematically understates the true effect of education.

IV Instrumental Variables

Omitted variable bias is a central challenge—indeed, it is *the* central challenge—to using regressions to determine the effect of one variable on another. The fundamental solution is to use *instrumental variables* (or *IV*) estimation.

The basics of instrumental variables. To understand IV, return to our model of wage determination, $\ln W_i = a + bE_i + u_i$. But suppose that in addition to data on years of education and wages, we have data on a third variable (an "instrument"), *Z*, that has two characteristics. First, *Z* is correlated (either positively or negatively) with *E*. Second, it is not systematically correlated with factors other than *E* that affect wages—that is, with *u*.

If we have such a variable, we can use a two-step procedure to estimate the impact of schooling on wages. The first stage is to run an OLS regression of schooling on the instrument:

$$E_i = c + dZ_i + v_i. \tag{4}$$

The purpose of this regression is *not* to find the *effect* of *Z* on *E*. Rather, it is just to find the portion of *E* that is correlated (for whatever reason) with *Z*. Let $\hat{c}$ and $\hat{d}$ be the estimates of *c* and *d* from this regression, and let $\hat{E}_i$ be the "fitted value" for observation *i*—that is, $\hat{c} + \hat{d}Z_i$—from the regression.

The second stage is an OLS regression of the wage not on years of education, but on the fitted value of *E* from the first-stage regression. To understand the reason for doing this, write E_i as $\hat{E}_i + \hat{v}_i$ (where $\hat{v}_i$ is defined as $E_i - \hat{E}_i$). Substituting this for E_i in our model of wage

determination, $\ln W_i = a + bE_i + u_i$, gives us:

(5)
$$\begin{aligned} \ln W_i &= a + b\left(\hat{E}_i + \hat{v}_i\right) + u_i \\ &= a + b\hat{E}_i + (b\hat{v}_i + u_i) \\ &\equiv a + b\hat{E}_i + \eta_i, \end{aligned}$$

where η_i is defined as $b\hat{v}_i + u_i$.

The second stage is a regression of $\ln W_i$ on $\hat{E}_i$. Thus, to determine whether it will give us a valid estimate of *b* we need to know whether η is systematically correlated with $\hat{E}$. $\hat{E}$ is a linear function of *Z*. Thus, asking whether η is correlated with $\hat{E}$ is the same as asking whether it is correlated with *Z*. Now, η has two pieces, $b\hat{v}$ and u. $\hat{v}$ is the residual from the OLS regression of *E* on *Z*. By construction, it is therefore uncorrelated with *Z*.[5] And one of our two assumptions about *Z* is that it is not systematically correlated with *u*. Thus, there is no systematic correlation between either piece of η and $\hat{E}$, and so the regression gives us a valid estimate of *b*—that is, of the true effect of schooling on wages.[6]

Intuitively, what IV does is to estimate the relationship between education and wages using only a *portion* of the variation in education. Our worry is that for some of the things that cause education to vary across individuals, such as differences in parental resources or in personal discipline, forces that cause wages to be higher for a given amount of education are on average stronger when education is greater. As a result, if we estimate the relationship between education

[5] A basic fact about OLS regression is that the coefficients that minimize the sum of squared differences between the actual values of the dependent variable and the fitted values (see n. 2) make the regression residual uncorrelated with the right-hand side variable.

[6] This exposition sweeps some important issues under the rug (mainly ones involving "consistency" and "finite-sample bias" and the standard errors). They are covered in econometrics courses.

and wages using all the variation in education, we get an estimate that exceeds the true effect of education. But suppose we can find something that is associated with variation in education where the forces that cause wages to be higher for a given amount of education are not on average stronger or weaker when education is greater. Then if we ask how wages vary with that specific component of education, we will get a good estimate of the actual effect of education on wages.

***Valid instruments*.** So far, this discussion probably seems like the famous joke about the economist and the can opener. We have assumed the existence of a wonderful tool—in this case, the instrument—that allows us to solve our problem. But we have not said anything about where to find such a tool.[7] And indeed, for IV to work, we need a valid instrument—in our case, a variable correlated with E and not systematically correlated with u. And in most situations, finding a valid instrument is not easy.

To understand what a valid instrument might be like, return to our example from Section I of a group of individuals who were on the margin of getting additional education, and who we could randomly allocate into one group that that did get the extra education and one that did not. Now define a variable that is +1 for the individuals who were randomly allocated to get more education and −1 for those who were randomly allocated to not get more education. And if there are any individuals in the sample who were not on the verge of getting more education, let the variable equal 0 for them. The variable would be a valid instrument for our regression. The individuals for whom it is +1 on average get more education than those for whom it is −1. Thus it is correlated with E. And because it is chosen randomly, it is not systematically correlated with other factors influencing education (that is, with u).

Unfortunately (from the point of view of estimating the effects of one variable on another),

[7] If you do not know the joke, type "economist can opener" into any search engine.

true randomization is rare in economics. Economists have therefore devoted a great deal of effort to trying to find cases where allocations are similar to what they would be with randomization. One famous study used the fact that the amount of education students are legally required to obtain varies with their birthdate.[8] For example, in a state where students start kindergarten in September if they turn five before September 1, a student born on August 31 will start school almost a year younger than one born on September 1. If students are legally required to stay in school until they turn sixteen, this means that the student born on August 31 is legally required to obtain one more year of schooling than the one born on September 1. As a result, individuals born in late August typically receive slightly more education than ones born in early September. And since whether a child is born in late August or early September is effectively random, this difference in the amount of education these individuals receive is almost surely not due to other differences between the two groups. Another well known study focused on a major program in Indonesia in the 1970s of building schools in areas where they were previously lacking. Because the timing of the program varied greatly across regions of the country, relatively small differences across individuals in when and where they were born led to substantial differences in the amount of education they received.[9]

***Example*.** Consider again our example of wages and schooling, but now suppose we have data on an instrument, *Z*, that is correlated with years of education and that we are confident is not systematically correlated with the residual. for concreteness, we assume that the instrument is similar in spirit to the examples we discussed in the introduction: among some high school students who are on the margin of going to community college, a few are pushed more or less at random in to not going, and so to getting less education than they would otherwise, and a few are

[8] Joshua D. Angrist and Alan B. Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (November 1991): 979–1014.
[9] Esther Duflo, "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review* 91 (September 2001): 795–813.

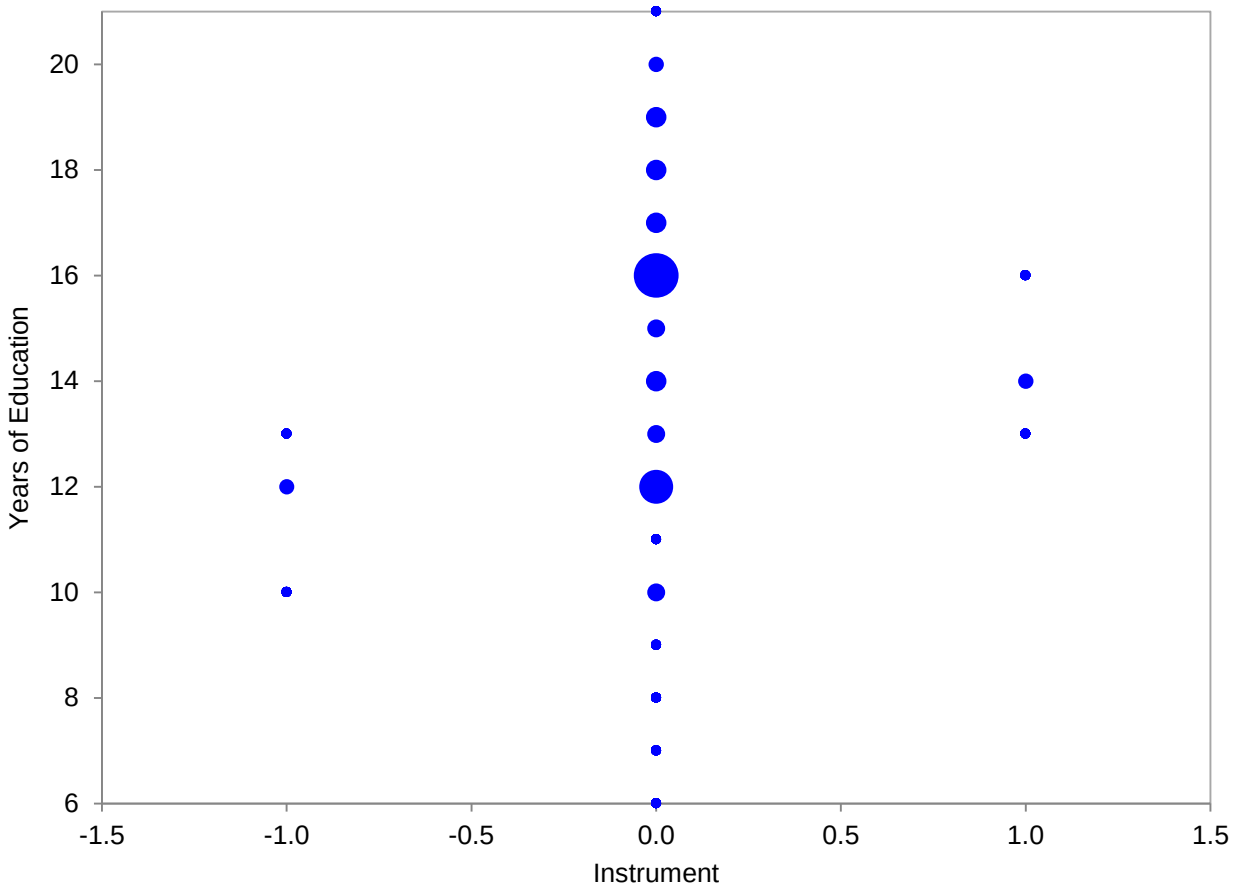

Figure 3. Scatterplot of Years of Education against Instrument

pushed into going. Thus in the example, the instrument takes on only three possible values: 0 (for the individuals who are not subject to the intervention, which is a large majority of the sample), −1 (for the students who are pushed to get less education), and 1 (for the students who are pushed to get more).

Figure 3 shows the first stage of IV: it is a scatterplot of *E* against *Z*, together with the regression line from a regression of *E* on a constant and *Z*. Three things are worth noting. First, the scatterplot is somewhat peculiar-looking; this reflects the facts that the instrument has only three possible values and that years of schooling only takes on integer values. (In the figure, the size of

each point is proportional to the number of observations.) Second, and more importantly, there is a clear relationship between the two variables, which is necessary for *Z* to be useful as an instrument. Third, and also importantly, the relationship is not particularly tight: there is a lot of variation in *E* that is not correlated with *Z*. That makes sense. The reason there is a need to use IV is that there are a lot of factors that affect how much education individuals get that are likely to be correlated with other factors that affect their wages. Thus there is a lot of variation in *E* that we do not want to use to estimate the impact schooling on wages, because we think that it would lead to a biased estimate. In addition, it is very unlikely that we will have an instrument that captures all the variation in *E* that is not correlated with the residual. This too will reduce the amount of variation in *E* that is correlated with *E*. In actual empirical work (such as the two studies cited above), it is common for the instruments to capture only a very small portion of the variation in the independent variable. That does not make them invalid instruments; the only problem it creates is that it tends to mean that we need a large sample to learn a lot from the estimates.

Figure 4 is the scatterplot corresponding to the second stage of the IV estimation. That is, it is a scatterplot of ln *W* not against years of schooling (*E*), but against the portion of years of schooling correlated with the instrument (the fitted values from the first-stage regression, $\hat{E}$). The figure shows a positive relationship, but one with a slope that is smaller than the simple relationship shown in Figure 1. This is what one would expect if there is omitted variable bias that causes OLS to overstate the effect of schooling on wages and if we have found an instrument that solves the problem.

The results of the IV estimation in our example are:

(6) $$\ln W_i = \underset{(0.28)}{2.44} + \underset{(0.019)}{0.070}\, E_i, \qquad N = 2500.$$

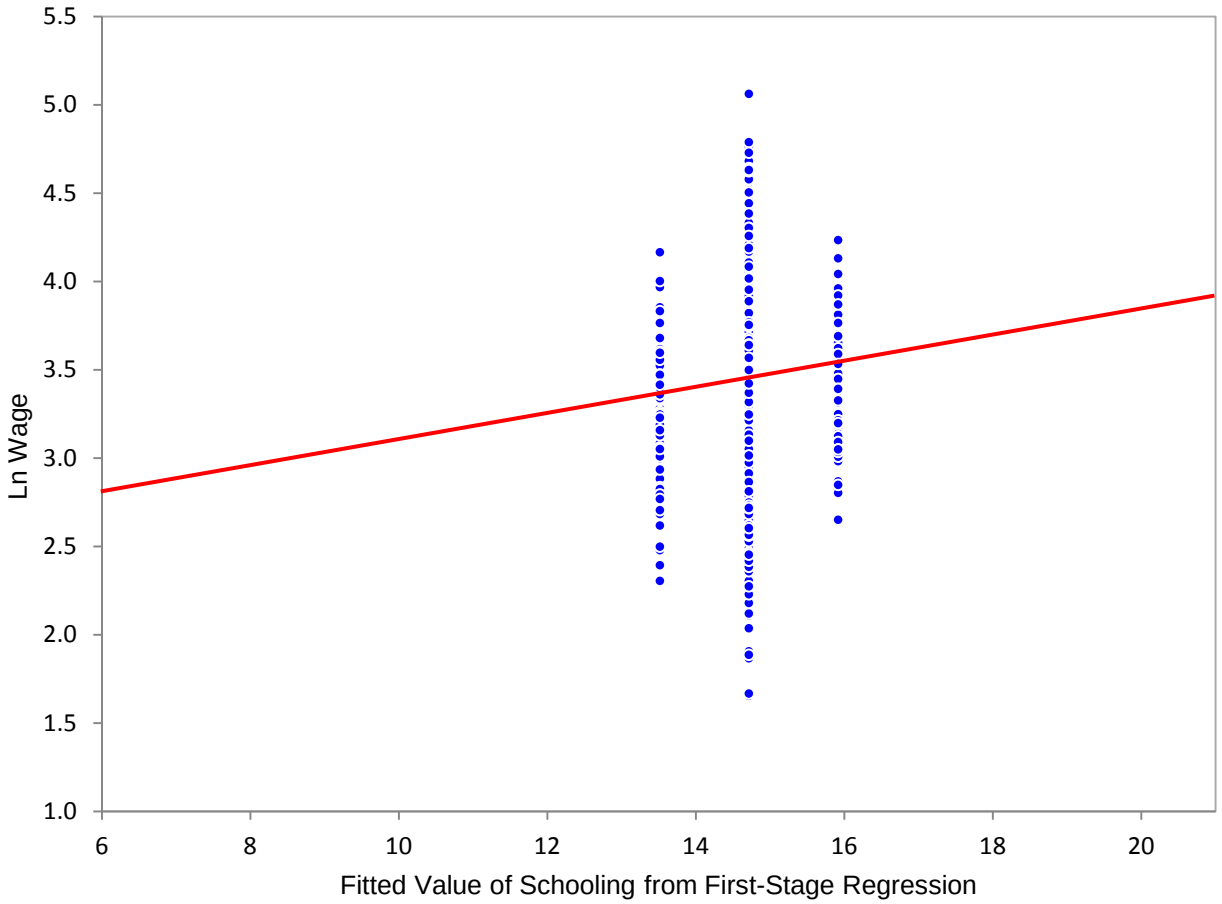

Figure 4. Regression Line and Scatterplot of Log Wages against Fitted Value from First-Stage Regression

As before, the numbers in the top line are the coefficient estimates, and the numbers in parentheses below the coefficient estimates are the standard errors of those estimates. Consistent with what we have been saying, the IV estimate of the impact of an additional year of schooling on wages (0.070) is quite a bit less than the OLS estimate (0.104). Figure 4 also shows the line corresponding to this regression, $\ln W_i = 2.44 + 0.070\hat{E}_i$.

V Interpreting Regressions

***Point estimates and standard errors*.** Suppose we have a regression we believe estimates the impact of the right-hand side variable on the left-hand side one. That is, suppose that we have either an OLS regression that we believe does not suffer from omitted variable bias or an IV estimate where we believe that the instrument is valid (in the sense of not being systematically correlated with the residual). How should we interpret the results of the regression?

For concreteness, we will discuss this issue in the context of the IV estimates reported in equation (6). There are two key numbers in (6). The first is 0.070 in the top line. This is $\hat{b}$—the estimate of *b* in our simple model of wage determination in equation (1). 0.070 is the number most consistent with the statistical relationship we are estimating. $\hat{b}$ is referred to as the "point estimate" of the effect of education on wages. It tells us that our best estimate of *b* is 0.070—that is, that on average an additional year of schooling raises wages by about 7 percent.

The other important number in (6) is 0.019 in the second line. Recall that the key characteristic of a good instrument is that it is not *systematically* correlated with the omitted variables (that is, with u_i). A more formal way of stating this characteristic is that the correlation will approach zero as the sample size becomes larger and larger. But in any given sample, there will be some correlation just by chance. Consider the example of looking at children born right around September 1. In any sample, it could be that the children born just after September 1 just happen to earn somewhat more or somewhat less for a given amount of education than those born just before September 1.

The result of this random variation in the correlation between the instrument and the omitted variables is that even when we have a valid instrument, the estimated coefficient, $\hat{b}$, will differ randomly from the true effect, *b*. The standard error—in our case, 0.019—is an estimate of

the size of those random differences. Specifically, under certain assumptions (the details of which are covered in econometrics courses), the standard error is a good estimate of the standard deviation of the difference between $\hat{b}$ and *b*.

***Confidence intervals*.** The point estimate and standard error can be combined to form a *confidence interval* for *b*. If the instrument is valid, $\hat{b}$ will on average equal *b*; that is, the mean of $\hat{b} - b$ is zero. The chances of $\hat{b} - b$ being more than two standard errors away from its mean are small—about 5 percent. For that reason, researchers often focus on the *two-standard error confidence interval*—that is, the range from $\hat{b}$ minus two times the standard error to $\hat{b}$ plus two times the standard error. In equation (6), the two-standard error confidence interval is from 0.070 − 2•0.019 to 0.070 + 2•0.019, or (0.032,0.108). The two-standard error confidence interval is sometimes referred to as the 95 percent confidence interval.[10]

It is important to be clear about what the two-standard error confidence interval means. What it shows is the range of values of *b* for which (under the assumption that the instrument is valid) it would not be surprising to obtain the estimate of $\hat{b}$ that we did. For example, suppose the true value of *b* is 0.045. Then a value of $\hat{b}$ is about 1.3 standard deviations away from *b*. A departure of this size or larger would occur about 20 percent of the time, and so would not be particularly surprising. Thus the regression results do not provide strong evidence against the view that *b* is 0.045. More generally, the data provide strong evidence against any proposed value of *b* that is less than 0.032 or greater than 0.108 (in the sense that we would be quite unlikely to get a value of $\hat{b}$ that far from the true value of *b*), but they do not provide such strong evidence against proposed values of *b* between 0.032 and 0.108. We say that the data *reject* the hypothesis ("at the 5

[10] Once again, we have swept some details under the rug. As described in econometrics courses, the width of the 95 percent confidence interval is not exactly plus or minus twice the standard error, and the factor we need to multiply the standard error by to construct the 95 percent confidence interval differs slightly depending on the number of observations and the number of right-hand side variables.

percent level") that, say, $b = 0.01$ or $b = 0.12$, but they *fail to reject* (again, at the 5 percent level) the hypothesis that, say, $b = 0.05$ or $b = 0.10$.

There are two common errors in interpreting confidence intervals. The first is to jump from the statement that the data fail to reject a hypothesis to the statement that they accept a hypothesis. Even if some value of b that we are especially interested in (perhaps zero, or perhaps the estimate from the OLS regression) lies in the 95 percent confidence interval, so are many other values. Thus regression results can never point to one specific value of the parameter as being the correct one.

The second error is to jump from the statement that 95 percent of the time $\hat{b}$ will lie within two standard errors of the true value of b to the statement that there is a 95 percent chance that the true value of b is in the 95 percent confidence interval. In fact, however, the likelihood of various possible b's depends not just on the results of the regression, but on all the information we have about various possible values of b.

This point is subtle, so it may help to consider a simpler example. Suppose you get a 50-cent coin at the bank. You are interested in the likelihood of a flip of the coin coming up heads; call this likelihood q. You flip the coin 6 times, and it comes up heads each time. The chance of getting 6 heads in a row is less than 5 percent if q is less than about 0.0607, and more than 5 percent or more is q is greater than that. Should you therefore think there is a 95 percent chance that the true value of q is greater than 60.7 percent—that is, that there is a 95 percent chance you have gotten a very unfair coin? Surely not! Before you did the coin flips, you had good reasons to think that the true value of q was very close to 50 percent. The coin looks pretty symmetric, and even though you have never flipped that particular coin before, you have flipped many other coins, and seen or heard about many other flips, and all that evidence points to the outcomes of flips being very close to 50-50. Thus, even though getting 6 heads in a row would not be very surprising if the

true value of q were, say, 0.8, and would be quite surprising if the true value of q were 0.5, the totality of evidence that you have will almost certainly lead you to think that it is much more likely that q is 0.5 than that it is 0.8. That is, you would be making a big mistake (and potentially a very costly one if you bet on the outcome of future coin flips) if you decided that you were 95 percent sure that the coin was very unfair.

***t-statistics*.** Another statistic about regression coefficients that is often reported is the *t-statistic*, which is the ratio of the point estimate to the standard error. In fact, in many papers the numbers shown in parentheses underneath the point estimates are *t*-statistics rather than standard errors.

To see the usefulness of a *t*-statistic, suppose the true value of some coefficient is zero. Then (under the assumption that the regression does not suffer from omitted variable bias), the point estimate from your regression will on average be zero, but it can be either positive or negative just by chance. It turns out that when the true coefficient is zero (and, again, there is no omitted variable bias), the *t*-statistic will have a distribution that is approximately normal with a mean of zero and a standard deviation of 1. Thus if the true coefficient is zero, there is only about a 5 percent chance that the *t*-statistic will be outside the range from −2 to +2.

The terminology that goes with this is that if the *t*-statistic is below −2 or greater than +2, we say that the coefficient is *statistically significant*; if it is between −2 and +2, we say that the coefficient is *statistically insignificant*. That is, saying that a coefficient is statistically significant is shorthand for saying that the data reject the hypothesis that the coefficient is zero; saying that it is statistically insignificant is shorthand for saying that the data fail to reject that the coefficient is zero.

As with confidence intervals, there are two common errors in interpreting *t*-statistics. The

first is to jump from a statement that a coefficient is statistically insignificant to the statement that the data indicate that the coefficient is zero. This is an example of the error of confusing failing to reject a hypothesis with accepting a hypothesis. If a coefficient is statistically insignificant, then (by the definition of statistical insignificance) we cannot reject the null hypothesis that it is zero. But there is a range of hypotheses about the coefficient that we cannot reject, not just the hypothesis that the coefficient is zero: we cannot reject the hypothesis that the coefficient equals any value in the two-standard error confidence interval.

The other common error is to jump from a finding that a coefficient is *statistically* significant to the conclusion that we have found something that is *economically* important. A finding that a coefficient is statistically significant provides support for the view that the true value of the coefficient is different from zero. But it tells us nothing about how large the difference is. To do that, we need to look at the point estimate and the standard error—that is, we need to look at the confidence interval.

Again, because these issues are often a source of confusion, it may help to consider an example. As usual, consider the relationship between wages and years of education. To keep the focus on the issues we are discussing here, suppose we have a regression that we are confident does not suffer from omitted variable bias. Now consider two possible results from the regression. In the first, the point estimate is 0.0010, with a standard error of 0.0001. In the second, the point estimate is 0.2, with a standard error of 0.5. These numbers imply that in the first case, the *t*-statistic is 10, while in the second, it is 0.2. Thus the coefficient in the first regression is extremely statistically significant, and that from the second is extremely insignificant. But notice that it would be completely wrong to conclude that the first regression shows an important effect of schooling on wages and that the second shows no effect. The two-standard error confidence

interval from the first regression is 0.0010 − 2•0.0001 to 0.010 + 2•0.001, or (0.0008,0.0012). Thus, while the data provide strong evidence against the view that the impact of an additional year of education on wages is exactly zero, they also provide strong evidence against the hypothesis that the true effect equals any value greater than 0.0012—that is, that an additional year of schooling raises wages by more than about 0.12 percent. Thus the correct conclusion to draw from the first regression is that it provides strong evidence for the view that the impact of schooling on wages is very small.

Conversely, the two-standard error confidence interval from the second regression is (−0.8,1.2). Thus in this case, while we cannot reject a coefficient of zero, we also cannot reject the hypothesis that the coefficient is 1 (so that each additional year of schooling leads to more than a doubling of wages)—or the hypothesis that it is −0.7 for that matter (so that each additional of schooling causes wages to fall more than in half).[11] Thus in this case, the correct conclusion to draw from the regression is not that it shows no impact of schooling on wages, but that these data are close to completely uninformative about any economically interesting hypothesis about the impact of schooling on wages

A good way to avoid these errors is to *always focus on point estimates and confidence intervals and their quantitative interpretation, not on t-statistics and statistical significance*. In the examples we just discussed, focusing on confidence intervals (from −0.0008 to 0.0012 for the first regression, and from −0.8 to 1.2 for the second) and thinking about the economic interpretation of those numbers (for example, that a coefficient of 0.0012 implies an impact of schooling on wages that is very small for all practical purposes, and a coefficient of 1 implies an impact that is

[11] Recall that the regression is of the natural log of the wage on years of education, and that its implications for the level of the wage are given by $W_i = e^{a}e^{bE_i}e^{u_i}$ (equation [2]). $e^{1} \simeq 2.72$, so a coefficient of 1 implies that an additional year of schooling raises wages by a factor of 2.72, or 170 percent. $e^{-0.7} \simeq 0.497$, so a coefficient of −0.7 implies that a year of schooling lowers wages by slightly more than 50 percent.

enormous) would have led immediately to the correct interpretation of the regressions.

A corollary of this discussion is that in reporting regression results, it is better to report standard errors rather than *t*-statistics in parentheses with the point estimates. In keeping with this, most modern papers report standard errors, not *t*-statistics.

VI Multivariate Regressions

So far, we have only discussed regressions with one independent variable. But regressions with multiple independent variables are often useful.

To see the potential value of multivariate regressions, return yet again to our discussion of schooling and wages. Suppose the amount of education an individual gets depends mainly on two factors: the economic circumstances of the family they grew up in, and happenstance (that is, a set of factors that are effectively random). The economic circumstances of the individual's family are likely to be correlated with forces other than schooling that influence how much the individual earns. For example, someone coming from a wealthier family is likely to live in an area with higher quality public schools, to get better health care, to have a stronger network of connections that are helpful in finding a job, and so on. On the other hand, the random factors that affect the amount of schooling an individual gets—for example, whether they happen to be talking to someone who had a good or a bad experience with college just as they are deciding whether to apply—are likely to be much less correlated with other influences on the individual's wages.

This discussion implies that if we had a measure of happenstance, it would make a good instrument for years of education: it is correlated with the amount of education individuals get, and it is not likely to be systematically correlated with other influences on their wages. Unfortunately, we are unlikely to have a good measure of happenstance. But we may have reasonably good

measures of families' economic circumstances, such as their incomes. This suggests another approach. In our example, the portion of the amount of education that individuals get that is *not* the result of economic circumstances would be a good instrument for schooling. Thus, we could first run a regression of years of schooling on a constant term and family income. We could then find the residuals that came out of the regression, and use that variable as an instrument for schooling. Since the residual is not correlated with family income, it may be a reasonably good proxy for the influence of happenstance on schooling—and thus a good instrument.[12]

This approach is perfectly reasonable. But there is a much easier way to do the same thing: we can just run a regression of wages on *both* years of schooling and family income. That is, we can estimate, simply by ordinary least squares with no use of instrumental variables, the regression:

$$\ln W_i = a + b_1 E_i + b_2 Y_i + u_i, \tag{7}$$

where Y_i is a measure of the income of the family that individual *i* grew up in. It turns out (and one can prove) that the estimate of b_1 from this regression is *identical* to the coefficient estimate you would get from the instrumental variables approach we just described. Intuitively, when we run the regression in (7), we are asking how ln *W* varies with *E*, holding *Y* fixed. The answer to this question is determined by the relation between ln *W* and the variation in *E* in our data that is not associated with variation in *Y*. Likewise, if we took the instrumental variables approach, we would be finding the relation between ln *W* and the variation in *E* that is not correlated with *Y*. Thus it

[12] In symbols, the starting point would be to run the regression $E_i = f + qY_i + e_i$ by ordinary least squares, where Y_i is a measure of the income of the family individual *i* grew up in. Then, letting $\hat{f}$ and $\hat{q}$ denote the coefficient estimates from the regression, one would compute $\hat{e}_i = E_i - (\hat{f} + \hat{q}Y_i)$, and use $\hat{e}_i$ as an instrument in our original regression of wages on years of schooling.

makes sense that the two methods give the same answer. But the regression of wages on schooling and family income is much simpler—rather than three steps (the first regression to construct the instrument plus the two regressions of instrumental variables), it has only one. For this and other reasons, it is the preferred approach.

A regression with more than one independent variable is known as a *multivariate* or *multiple* regression. The main purpose of multivariate regressions, as with instrumental variables, is to focus on a subset of the variation in the key variable (in our case, years of schooling) in order to estimate the causal effect of the key variable on the dependent variable (in our case, wages). Any independent variables other than the key one are known as *control* variables. In our case, the only control variable is family income. Thus we would say that we are estimating the relationship between wages and schooling *controlling* for family income.

One important implication of this discussion is that the estimate of the coefficient on the key variable and the estimates of the coefficients on the control variables should be interpreted very differently. The estimate of the coefficient on the key variable (in our case, the estimate of b_1) is intended to be an estimate of the key variable's causal impact on the dependent variable. The estimates of the coefficients on any control variables (in our case, the estimate of b_2), on the other hand, are not intended to be estimates of the control variables' causal impact on the dependent variable. Rather, as we have been discussing, the purpose of including the controls is only to affect what portion of the variation in the key variable is used to estimate its relationship with the dependent variable. That is, the purpose of including the additional variables is to help us get a better estimate of the key variable's causal impact. There is therefore no particular reason to think that the values of the coefficients we get on those variables will be good estimates of their causal impact on the dependent variable. In the case of our example, family income is likely to be

correlated with many other factors that affect wages. That is, as a way of estimating the effect of family income on wages, equation (7) almost certainly suffers from omitted variable bias, and is thus a poor approach. But again, that is not why we included family income in the regression.

A related implication of this discussion is that the purpose of a multivariate regression is not to "explain" as much of the variation in the dependent variable as possible. In deciding whether to include a potential control variable, the crucial question is whether it is likely to reduce omitted variable bias, not whether it makes the regression "better" in some other way. If, for example, we had some imperfect measure of the "chance" influences on how much education individuals received, it would be a terrible idea to include it as a control variable in (7). Controlling for some of the effects of chance on schooling would mean that a larger portion of the variation in schooling that was left to use to estimate b_1 was the result of factors other than chance, and thus would probably make omitted variable bias worse.

A final implication of this discussion is that including control variables is unlikely to fully solve omitted variable bias. The problem is that we almost certainly will not be able to control for all the factors that cause the omitted variable bias. In the case of schooling and wages, although family income affects the amount of schooling individuals get and is correlated with many factors other than schooling that affect that earnings, it does not completely capture all those factors. We saw back in Section III that there is a wide range of possible sources of omitted variable bias in a regression of wages on schooling, including some that involve things that are very difficult to measure. Thus a realistic goal in choosing control variables is to reduce omitted variable bias, not eliminate it. For that reason, when one is available, a persuasive instrument is generally the best way to address omitted variable bias and obtain reliable estimates of causal effects.

VII A Little Bit about Some Other Complications

Actual empirical work is more complicated than the picture we have painted so far. Here we briefly discuss a few of the more important issues.

***Actual data are messy*.** The "empirical" results presented in these notes are based on artificially created data, not any actual datasets. There is a simple reason for this: all actual datasets are messy. There are often typos and other outright errors. When the value of a variable is missing for an observation, in some datasets it is entered as zero, or as some crazy value, such as −9999. There may be ambiguities in how variables are defined. And more.

In addition, virtually all datasets have *outliers*—that is, extreme values of some variables. Extreme observations can be a powerful source of evidence. For example, the experience of North Korea provides almost incontrovertible evidence that bad institutions and bad government are harmful to an economy. But outliers should be treated with caution. Ordinary least square regressions put very large weight on them. And the forces that create an outlier may also cause the residual for that observation to have an especially high variance, implying that the observation should get even less weight. And an outlier may reflect a data error rather than a genuine data point.

The main message here is simple: treat your data with care. Read the data definitions and documentation attentively. Plot your data in some way. Dig into the specifics of any extreme observations to try to figure out if they are data errors or genuine—and if they are genuine, how important they are to your results. One useful rule is that whenever you have an empirical finding (for example, a result suggesting a quantitatively large, statistically significant impact of schooling on wages), you should be able to explain what it is in the data that is producing that finding based on simple summary statistics (such as means, standard deviations, and correlations) and graphs

(such as scatterplots).

Another message is that just as these notes are no substitute for an econometrics course, learning the theory of econometrics is no substitute for actually doing empirical work—either on your own, or even better, as a research assistant for an experienced empirical researcher.

***Standard errors are hard too*.** These notes focus on the challenges of estimating parameters, such as the impact of schooling on wages. All we have said about standard errors is that "under certain assumptions," they are estimates of the amount of variation in the parameter estimates that could arise just by chance. That simple statement sidesteps a very important issue: just as there are difficulties in estimating parameter values, there are difficulties in estimating standard errors.

Going into the specifics of what those difficulties are and how they can be addressed would take us further into econometric theory than is appropriate for these notes. One issue involving standard errors is sufficiently important and comes up sufficiently often, however, that it is worth mentioning. Conventional standard errors are computed under the assumption that the residual is independent across observations. That is, if we take two observations in the dataset, knowing the value of one of the u's tell us nothing about the value of the other one. That assumption is often wrong. For example, if our dataset has multiple students from the same school district, there may be common influences on their wages that are left out of the regression. To give another example, in a dataset that involves data that evolve over time, such as outcomes for the economy as a whole, the influences that are omitted from the model may change slowly from one period to the next. The reason this is important is that in some cases, positive correlation of the residuals of related observations can cause the standard errors that are conventionally reported to be too low—sometimes dramatically so. In short, standard errors should be treated with caution.

Nonlinearities and functional forms can be important. Equation (1) assumes that there is a linear relationship between years of education and the log of the wage. But perhaps the actual relationship is more complicated. Perhaps the percentage increase in the wage from each additional year of schooling declines as the amount of schooling rises. Or perhaps some years of education (such as the last few years of high school or the first few years of college) are especially valuable. Or perhaps beyond some point, further education actually leads to lower wages.

The fact that theory rarely tells us the correct functional form of a relationship reinforces the case for treating your data with care and using graphs to visualize your data. For example, if the relationship between the dependent and the independent variable is highly nonlinear, that will probably be apparent from a simple scatterplot.

Two questions that it often makes sense to think about concerning functional forms are whether to enter data in levels or in logs, and whether (in cases where you have observations over time) to enter them in changes. In many cases, it makes sense to take the log of a variable before using it in a regression. For example, consider our idea of using family income to capture the influences of the economic circumstances individuals grew up in on the amount of schooling they get, and so to improve our estimates of the impact of schooling on wages. Entering income in levels (as we did in equation [7]) makes the difference between an income of $2 million and an income of $1 million 50 times as large as the difference between an income of $40,000 and an income of $20,000. Entering them in logs makes the two differences equally large. Although there is no way to know a priori which approach will do a better job of capturing the influences of family circumstances, there is a good chance that the log approach will do better. And when we have time series data, regressions that do not use changes in the variables often suffer severely from various problems, including the issue of correlation in the residual across related observations. As a result,

although regressions using changes must still grapple with omitted variable bias, they are often a better starting point.

The discussion in this section is intended only to scratch the surface of some of the issues that come up in doing empirical work in practice. As you learn more econometrics and do empirical work of your own, you will quickly discover that there are many more complications than just those we have discussed here.

VIII Summary and Main Messages

1. An OLS regression is a way of choosing coefficient values that provide the best fit to some set of data.

2. Correlation is not causation.

3. Correlation between the effects of variables omitted from the regression and the right-hand side variable causes omitted variable bias.

4. Instrumental variables regression (with a valid instrument) can address the problem of omitted variable bias.

5. Statistical significance is not the same as economic significance. Always focus on point estimates and confidence intervals and their economic interpretation, not on *t*-statistics and statistical significance.

6. Empirical work is complicated.

7. You should take a course in econometrics.

PROBLEMS

1. Supply and demand. Suppose your goal is to estimate the elasticity of demand for blueberries. Your model of blueberry demand is $\ln Q_t = a - b \ln P_t + e_t$, where Q_t is the quantity of blueberries bought and sold in month *t* and P_t is the price of blueberries in month *t*. (The reason for entering the variables in logs is that we are interested in estimating the elasticity of demand rather than the slope of the demand curve. The reason for the minus sign, which is not essential, is that because we think demand depends negatively on price, it makes the equation easier to interpret.) You are considering trying to estimate *b* by an ordinary least squares regression of $\ln Q_t$ on a constant and $\ln P_t$.

a. The condition for a regression to give us a good estimate of the impact of the independent variable on the dependent one is that the residual is not systematically correlated with the independent variable. Is there likely to be systematic correlation between e_t and $\ln P_t$? (Hint: Your answer should involve supply and demand diagrams and discussing the effects of shifts of the curves.)

b. Suppose that in addition to data on *Q* and *P*, you have data on two other variables. The first, *X*, is a variable (such as the weather in blueberry-growing areas) that shifts the supply curve of blueberries but is not systematically correlated with factors that shift demand. The second, *Z*, is a variable (such as income in blueberry-consuming areas) that shifts the demand curve of blueberries but is not systematically correlated with factors that shift supply. Suppose you were going to use either *X* or *Z* as an instrument and estimate the regression by instrumental variables rather than ordinary least squares. Which one would you use, and why?

2. Suppose you estimate equation (1), $\ln W_i = a + bE_i + u_i$, by instrumental variables using an instrument you believe is valid and obtain the following results:

$$\begin{array}{lll} \ln W_i = 1.07 & + & 0.139\, E_i. \\ \quad (0.36) & & (0.041) \end{array}$$

What is the point estimate of *b*? What does the point estimate imply about how much an additional year of schooling raises wages? What is the standard error of the estimate of *b*? The two-standard error confidence interval for *b*? The *t*-statistic for *b*?

3. Suppose your goal is not to estimate the *effect* of education on wages, but to *predict* how much an individual earns. Concretely, suppose that the data on the wage that one individual in our dataset is missing by chance. If you had to guess that individual's wage, would you use the results of the OLS regression, which are $1.87 + 0.104E_i$, or the results of the IV regression, which are $2.44 + 0.070E_i$? Why?

4. Explain what is wrong with the following argument: "I had a dataset giving years of schooling and wages for a sample of workers. I estimated the OLS regression $\ln W_i = a + bE_i + u_i$ for that sample. After running the regression, I computed the correlation between years of schooling (E_i) and the regression residual ($\hat{u}_i$, defined as $\ln W_i - [\hat{a} + \hat{b}E_i]$, where $\hat{a}$ and $\hat{b}$ are

the coefficient estimates from the regression). It was exactly zero. Thus, since there is no correlation between the right-hand side variable and the residual, I know that my regression does not suffer from omitted variable bias."

5. (Classical measurement error.) Actual data are often not completely accurate. To see the possible effects of measurement error, consider the following example. Let E_i^* denote the true value of E_i. Wages are determined by $\ln W_i = a + bE_i^* + u_i$. For simplicity, assume that u_i is not systematically correlated with E_i^*, so we do not have to worry about omitted variable bias. Unfortunately, however, there is measurement error: what we observe is not E_i^* but $\hat{E}_i$, where $\hat{E}_i = E_i^* + v_i$. *v* is not systematically correlated with either *E** or *u*.

Our goal is to figure out whether running a regression of (log) wages on *measured* schooling will give a good estimate of the impact of schooling on wages, *b*.

a. If we write $\ln W_i = a + b\hat{E}_i + \delta_i$, what is δ_i in terms of u_i, v_i, and b? (Hint: Use the facts that we are defining δ_i by $\delta_i \equiv \ln W_i - (a + b\hat{E}_i)$, that $\ln W_i = a + bE_i^* + u_i$, and that $\hat{E}_i = E_i^* + v_i$.)

b. Assume $b > 0$. In the equation $\ln W_i = a + b\hat{E}_i + \delta_i$, is there systematic correlation between the residual, δ_i, and the independent variable, $\hat{E}_i$? If so, is it positive or negative? (Hint: Is v_i positively or negatively correlated with itself?) What will be the relationship between the true value of *b* and the estimate of *b* we tend to get from the regression?

c. What happens if the true value of *b* is negative? If it is zero?

6. (Another type of measurement error.) Suppose, as in Problem 5, that wages are determined by $\ln W_i = a + bE_i + u_i$. Unfortunately, there are two components of education, E^A and E^B, so that $E_i = E_i^A + E_i^B$, and we only have data on one of the components, E^A. (For example, E^A might be years of traditional schooling and E^B years of home schooling.) Both E^A and E^B are not systematically correlated with *u*. Our goal is to figure out whether running a regression of (log) wages on E^A will give a good estimate of the impact of schooling on wages, *b*.

a. If we write $\ln W_i = a + bE_i^A + \delta_i$, what is δ_i in terms of u_i, E_i^B, and b?

b. We know that an OLS regression is a good way to estimate a causal effect if the residual is not systemically correlated with the right-hand side variable. In the equation $\ln W_i = a + bE_i^A + \delta_i$, is there systematic correlation between the residual, δ_i, and the independent variable, E_i^A, if:

i. E^A and E^B are not systematically correlated?

ii. E^A and E^B are systemically negatively correlated?

7. (Measurement error in the dependent variable.) Suppose that wages are determined by $\ln W_i^* = a + bE_i + u_i$. Assume that u_i is not systematically correlated with E_i, so we do not

have to worry about omitted variable bias. Unfortunately, however, now there is measurement error in wages: what we observe is not $\ln W_i^*$ but Z_i, where $Z_i = \ln W_i^* + v_i$. v is not systematically correlated with *E*.

As usual, our goal is to figure out whether running a regression of measured (log) wages on schooling will give a good estimate of the impact of schooling on wages, b. Use the approach you took in Problem 5 (or 6) to answer this question.

8. Consider Problem 1. Suppose you were to going to use ordinary least squares, but you were going to include either *X* or *Z* as a control variable. That is, your decision is whether to run the regression $\ln Q_t = a - b \ln P_t + cX_t + e_t$ or the regression $\ln Q_t = a - b \ln P_t + cZ_t + e_t$. Your goal continues to be to estimate the elasticity of demand for blueberries. Which variable would you include as the control, and why?

9. (An example of *sample selection bias*.)[13] You are interested in whether countries that were poor a century ago have grown faster than countries that were relatively rich. There are four types of countries: countries started either rich or poor, and then grew either slowly or rapidly. Unfortunately, countries that started poor and grew slowly have not constructed income data that go back a century.

In the *sample of countries with data that are available*, are the countries that started poor more likely to have grown rapidly than the countries that started rich? Is the frequency of rapid growth in your dataset in the countries that started rich a good estimate of the fraction of all countries that started rich that grew rapidly? Is the frequency of rapid growth in your dataset in the countries that started poor a good estimate of the fraction of all countries that started poor that grew rapidly? Explain your answers.

[13] This problem is based on J. Bradford DeLong, ‘‘Productivity Growth, Convergence, and Welfare: Comment,’’ *American Economic Review* 78 (December 1988): 1138–1154.