UNIVERSITY OF CALIFORNIA
DEPARTMENT OF ECONOMICS

Economics 134
Spring 2018
Professor David Romer

LECTURE 5

INTRODUCTION TO REGRESSIONS AND EVIDENCE ABOUT CAUSATION
JANUARY 31, 2018

I. INTRODUCTION

II. ORDINARY LEAST SQUARES REGRESSIONS

A. A simple model of the determinants of wages
B. An OLS regression
C. Discussion in the context of education and wages
D. The general issue: omitted variable bias
E. Message: correlation is not causation

III. EXPERIMENTS

A. Types of experiments: controlled, randomized, and natural
B. One way to interpret the results of experiments: direct comparison of the treated and control groups ("reduced form")

IV. INSTRUMENTAL VARIABLES

A. Requirements for a good instrument
B. The first stage
C. The second stage
D. Discussion
E. Example in the context of education and wages
F. A second way of interpreting the results of experiments: constructing an instrument based on the experiment ("instrumental variables")

V. INTERPRETING REGRESSIONS

A. Point estimates and standard errors
B. Confidence intervals
    1. What a confidence interval is
    2. Rejecting and failing to reject hypotheses
    2. Two common errors in interpreting confidence intervals
C. *t*-statistics
    1. What a *t*-statistic is
    2. Two common errors in interpreting *t*-statistics
D. Message: always focus on point estimates and confidence intervals and their economic interpretation, not on *t*-statistics and statistical significance

Economics 134
Spring 2018

David Romer

LECTURE 5

Introduction to Regressions and Evidence about Causation

January 31, 2018

Announcement

- Problem Set 1 is being distributed.
- It is due at the ***beginning*** of lecture a week from today (Feb. 7).

Problem Set Ground Rules

- You may work together on the problems, but:
  - I strongly recommend working on the problems by yourself first.
  - Your answers must be handwritten and in your own words.
  - You must list other students you worked with at the start of your answers.
- Optional problem set work session: Monday, Feb. 5, 6:45–8:15, in 597 Evans Hall.

0. A Little More on MP vs. LM

MP or LM?

- Where the two models differ is in what they assume about how monetary policy is conducted.
- Thus, in deciding whether to use MP or LM, the key consideration is how monetary policy is conducted in the situation you are looking at.

The Fed under Pure LM

- Between policy meetings: The NY Fed would keep the stock of high-powered money constant.
- At policy meetings: The “default” decision would be to continue to keep the stock of high-powered money constant.

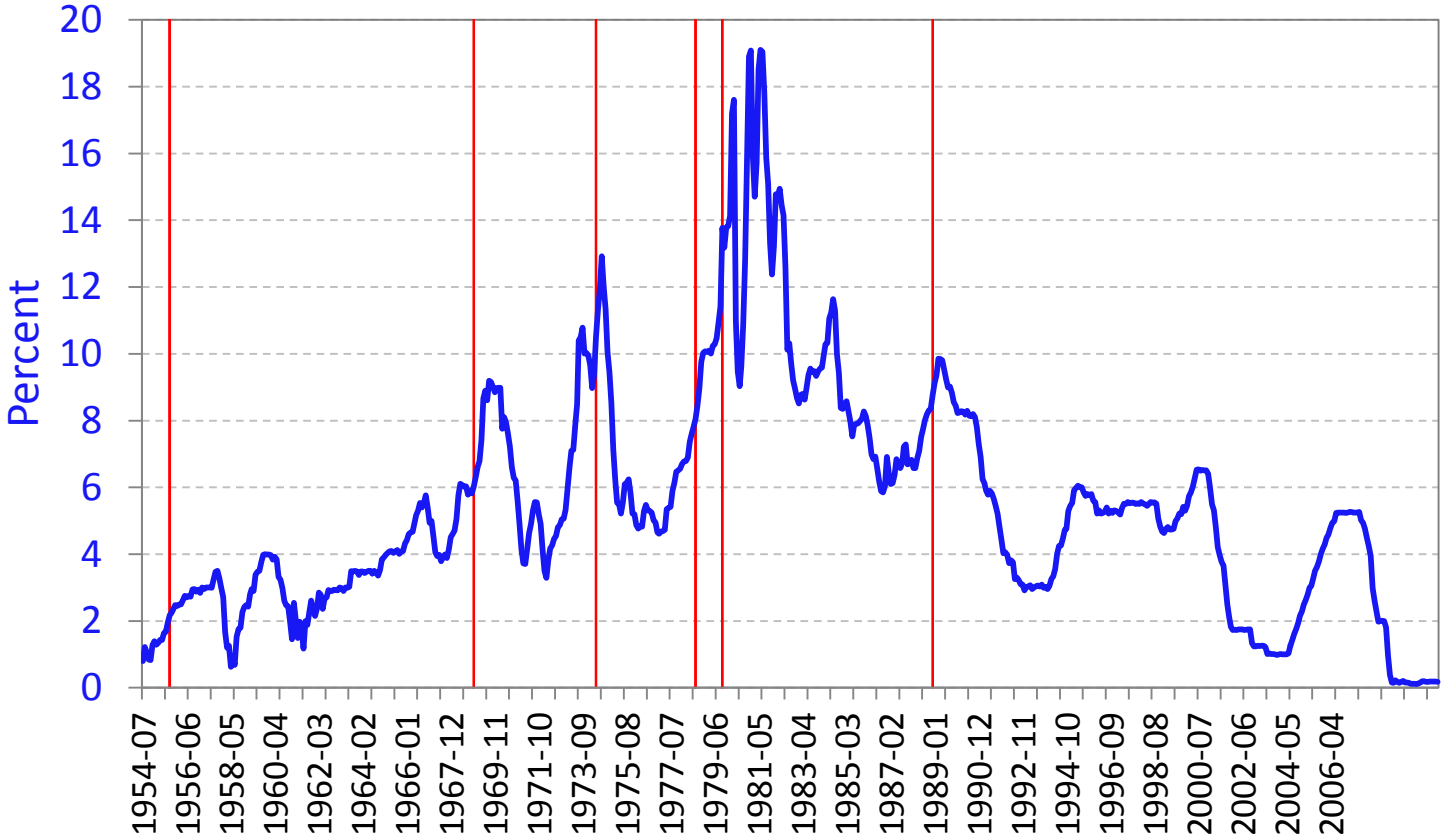
Actual Cases of Pure LM

- Hard to find any—perhaps the Island of Yap.
- Cases of impure MP: The Fed from 1930 to 1932 and from 1979 to 1982; the Bundesbank before the adoption of the euro.

Actual Cases of Pure MP

- The Fed during the Great Moderation (roughly 1985–2005).
- More generally, many modern central banks.
- (Note: This discussion ignores the zero lower bound, which complicates things and which we will discuss later in the course.)

Federal Funds Rate
1954:7-2007:12



Interest rates were very volatile in the period when the Fed was – to some extent – targeting the money supply.

I. Introduction

II. Ordinary Least Squares Regression

The Example We Will Focus on: The Impact of Education on Earnings

- Our question: What is the effect of the number of years of schooling that an individual obtains on their wage?

A Simple Model of the Determinants of Individuals' Wages

$$\ln W_i = a + bE_i + u_i,$$

where *i* indexes individuals, *E* denotes years of education, and *W* denote the wage.

Our goal is to obtain evidence about *b*.

An OLS Regression

- One candidate way to get an estimate of b would be to run an ***ordinary least squares*** (or ***OLS***) regression.

- For our purposes, just think of a regression as choosing values of a and b in

$$\ln W_i = a + bE_i + u_i$$

to provide the "best" fit to the data.

An Example of a Regression

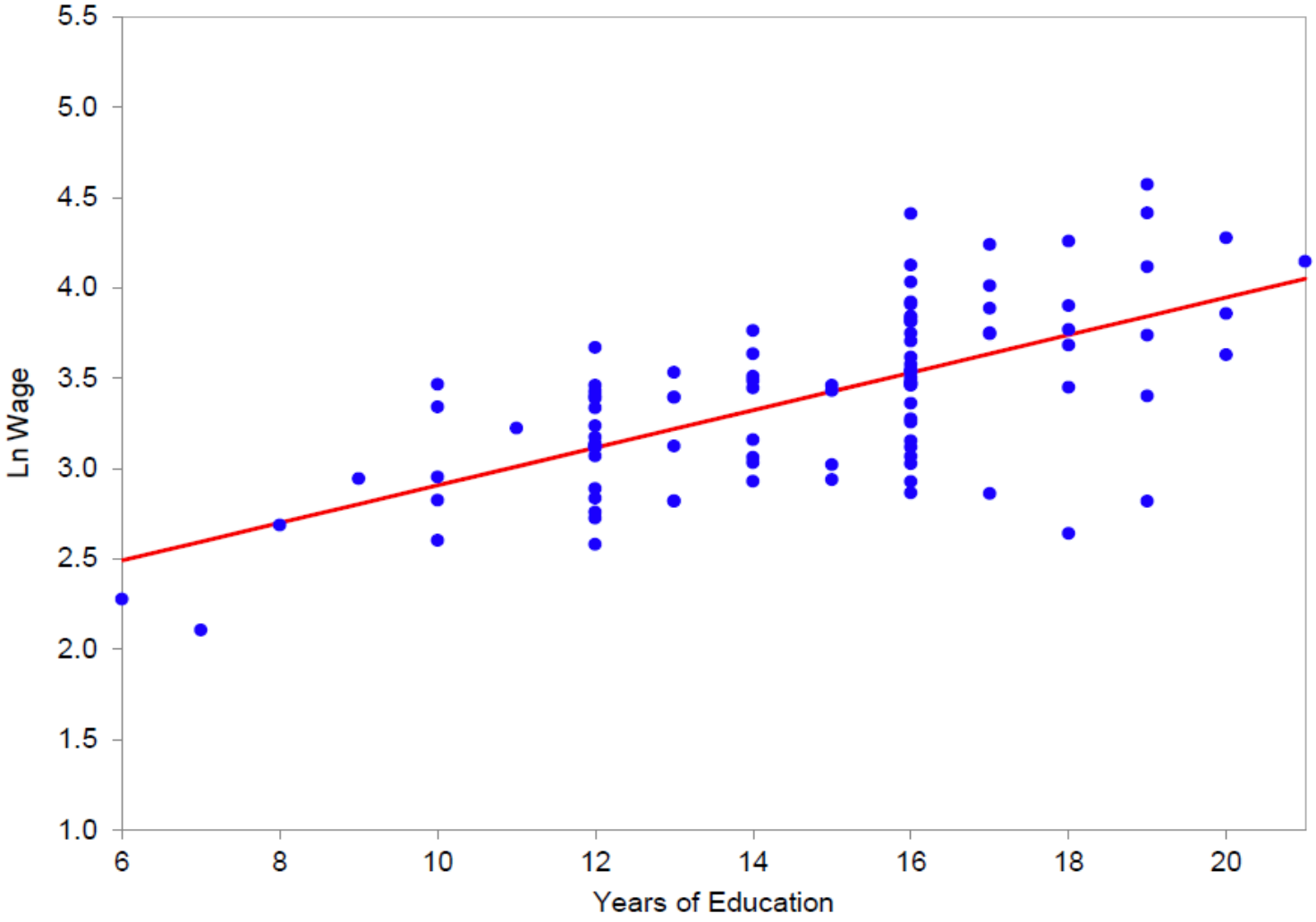


Figure 2. Regression Line and Scatterplot of Log Wages against Years of Education

The Results of the Regression

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

- The numbers in the top line (1.87 and 0.104) are the coefficient estimates, $\hat{a}$ and $\hat{b}$, from the regression.
- The numbers in parentheses below the coefficient estimates are the "standard errors" of $\hat{a}$ and $\hat{b}$. (We'll discuss these later.)
- "*N*" is the number of observations in our sample.

Discussion

- Is this a good way to estimate b?
- 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 suggest that one more year of education ***causes*** the wage to be on average 10 percent higher?
- NO!

Some Possible Reasons that $\hat{b}$ (the Regression Coefficient) Might Not Be a Good Estimate of b (the Effect of Schooling on Wages)

- Individuals who come from more advantaged backgrounds may get more education, but also have other advantages in the labor market.
- Some personality traits (such as self-discipline) may increase the amount of education individuals get, but also make them more productive in other ways.
- Being healthy may help individuals get more education, but also cause them to earn more for a given amount of education.
- …

The General Issue: Omitted Variable Bias

- Recall: $\ln W_i = a + bE_i + u_i$.
- If there is systematic correlation between the factors left out of the regression (here, u_i) and the variables in the regression (here, E_i), the coefficient estimate from the regression, $\hat{b}$, will be a biased estimate of b (the effect of schooling on wages).
- If the correlation is positive, $\hat{b}$ will tend to overstate b; if it is negative, $\hat{b}$ will tend to understate b.
- The name for this problem is ***omitted variable bias***.

The Message Boiled Down to Its Essence

- Correlation is not causation.

III. EXPERIMENTS

Controlled Experiments

- ***Only*** difference between treated and control groups is in the variable of interest.
- Rarely feasible in economics!

Randomized Experiments

- Random assignment to treatment and control groups.
- Eliminates possibility of omitted variable bias.

Natural Experiments

- Situations where ***as if*** random assignment.

One Way to Interpret the Results of Experiments: Direct Comparison of the Treated and Control Groups

- With a controlled experiment, two observations could be enough.
- With a randomized experiment, we have to worry about random differences between the two groups.
- With a natural experiment, we also have to worry about whether the assignment really is as if random.

Advantages and Disadvantages of Direct Comparison

- Advantage: Simple, straightforward.
- Disadvantage: Hard to interpret the size of any difference between the groups.

IV. Instrumental Variables

The Solution to Omitted Variable Bias Is Instrumental Variables ("IV")—with a Good Instrument

The two requirements for a good instrument:

- Correlated with the right-hand side variable (in our case, E).
- Not systematically correlated with the residual—that is, with u.

Using an Instrument – Step 1

- Run a regression of the right-hand side variable (in our case, E) on the instrument and get the "fitted values."

- That is, run the regression $E_i = c + dZ_i + v_i$ (where Z is the instrument) by OLS, and compute $\hat{E}_i = \hat{c} + \hat{d}Z_i$ for each observation.

Using an Instrument – Step 2

- Run a regression of the left-hand side variable (in our case, ln W) on the fitted values you constructed in the first step.
- That is, run the regression $\ln W_i = a + b\hat{E}_i + u_i$ by OLS.
- Claim: The estimate of b that comes out of this regression, $\hat{b}$, is a good estimate of the causal effect of schooling on wages, b.

Why an IV Regression (with a Good Instrument) Does Not Suffer from Omitted Variable Bias

- Recall our model: $\ln W_i = a + bE_i + u_i$.
- We can write E_i as $\hat{E}_i + \hat{v}_i$ (where $\hat{E}_i$ is the fitted value from our first step and $\hat{v}_i$ is defined as $E_i - \hat{E}_i$).
- Thus, $\ln W_i = a + b(\hat{E}_i + \hat{v}_i) + u_i$

$$= a + b\hat{E}_i + (b\hat{v}_i + u_i)$$

$$\equiv a + b\hat{E}_i + \eta_i,$$

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

Why an IV Regression Does Not Suffer from Omitted Variable Bias (continued)

- $\ln W_i = a + b\hat{E}_i + \eta_i$.
- $\hat{E}$ is a linear function of Z. So whether $\hat{E}$ is correlated with η is determined by whether Z is correlated with η.
- η has two pieces, $b\hat{v}$ and u.
- $\hat{v}$ is the residual from the OLS regression of E on Z. By construction, Z and $\hat{v}$ are therefore uncorrelated.
- And one of our two assumptions about Z is that it is not systematically correlated with u.
- Thus, there is no systematic correlation between $\hat{E}$ and η, and so the regression gives us a valid estimate of b—that is, of the true effect of schooling on wages.

Discussion

Example in the Context of Education and Wages

- Suppose some high-school students who are on the margin of going to community college are pushed randomly or quasi-randomly to go, and some are pushed to not go.
- The instrument is a variable equal to +1 for those for are pushed to go, −1 for those pushed to not go, and 0 for everyone else.

The First Stage

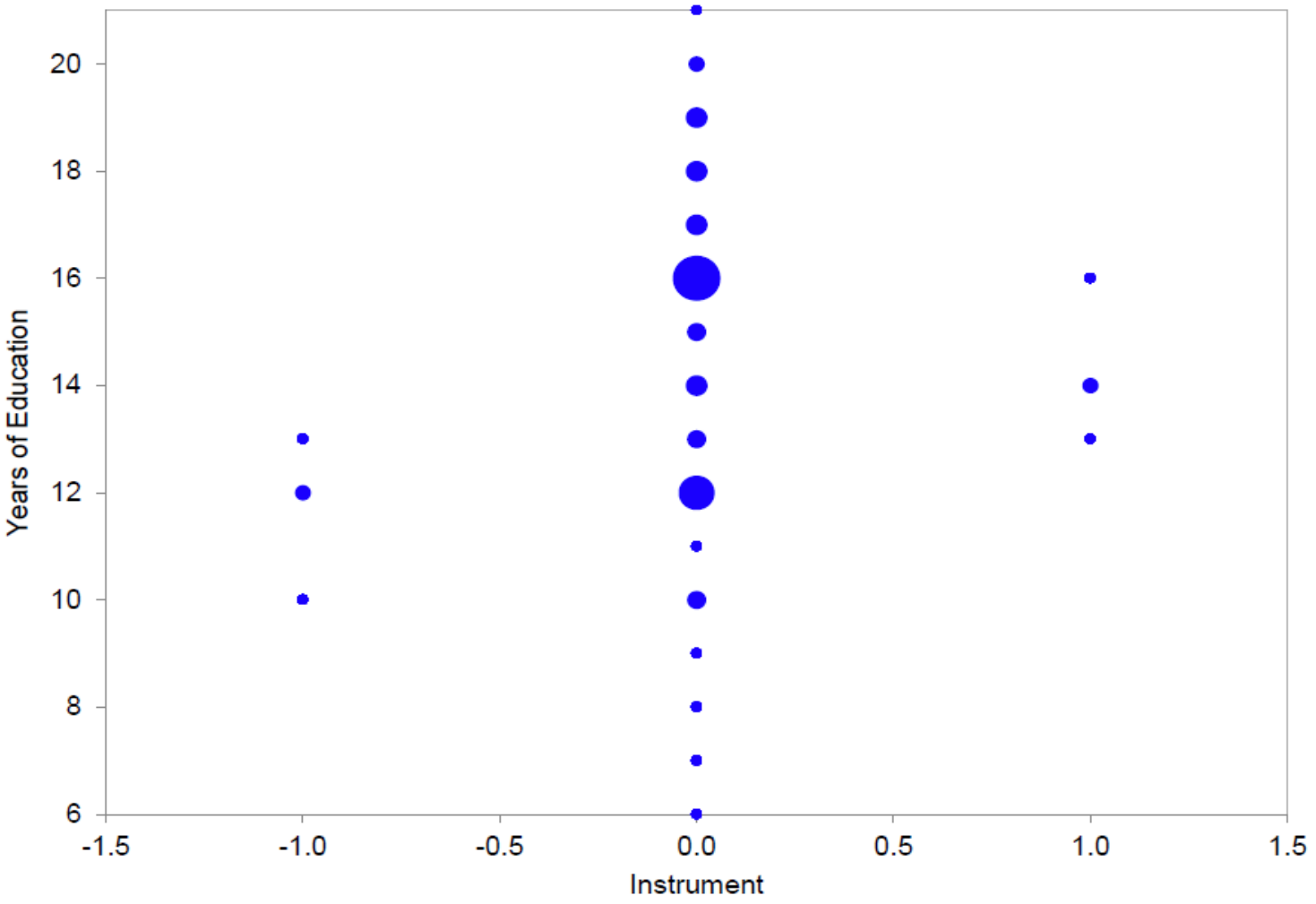
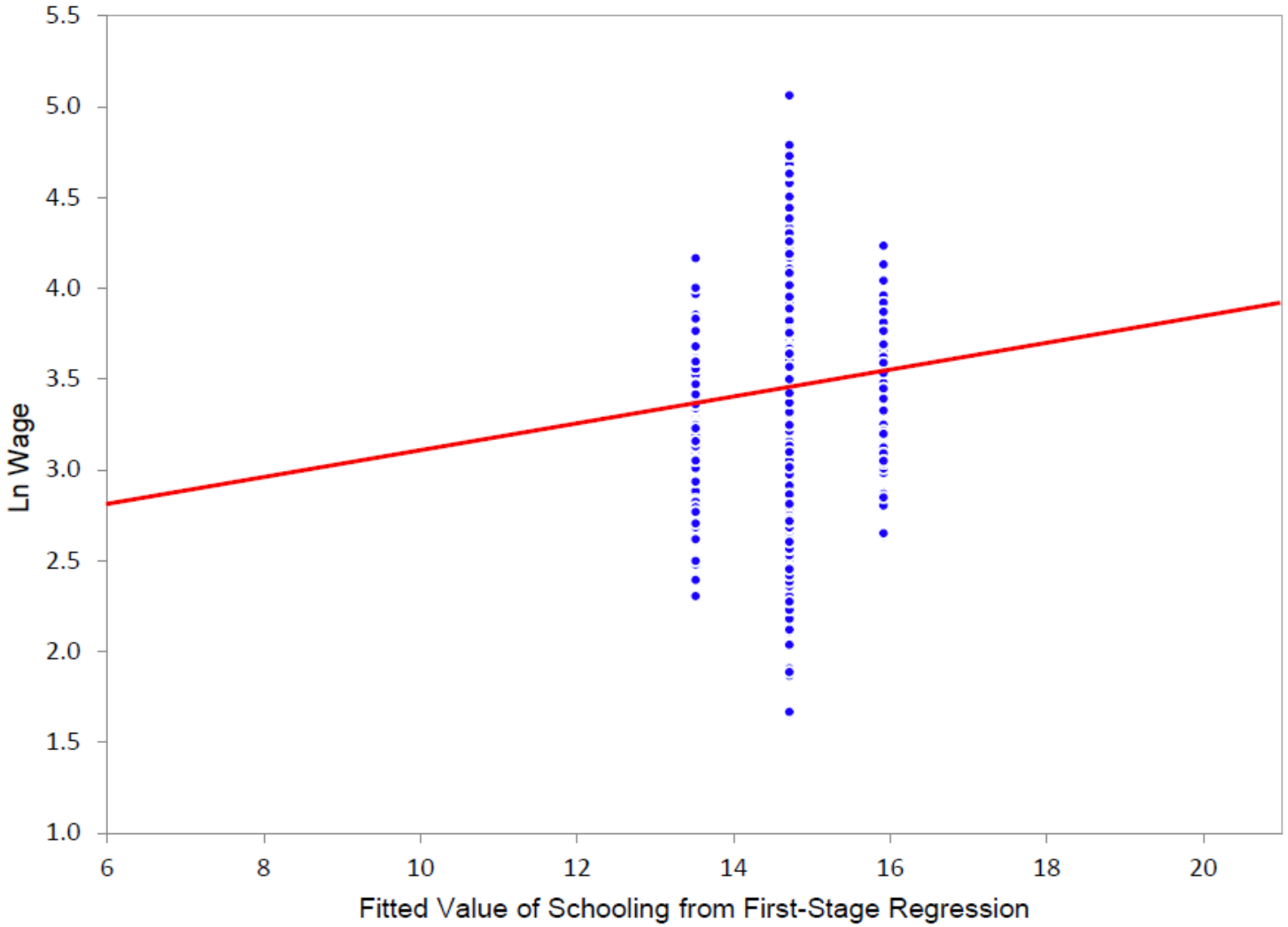


Figure 3. Scatterplot of Years of Education against Instrument

The Second Stage



The Results of the IV Regression

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

- The numbers in the top line are the coefficient estimates, $\hat{a}$ and $\hat{b}$, from the IV regression.
- 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.

A Tiny Bit about a Second Way of Interpreting the Results of Experiments

- As the previous example shows, you can use an experiment to construct an instrument, and use that in a regression.
- Advantage: Gives us an estimate of a parameter of interest (in our case, b).
- Disadvantage: More complicated than the simple comparison of the treatment and control groups.
- (We won't emphasize this way of interpreting experiments.)

V. Interpreting Regressions

Regression Results

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

- How should we interpret this?
- Note: The discussion that follows assumes that the instrument is a good one, and thus that we do not have to worry about omitted variable bias.

Point Estimates

$$\ln W_i = 2.44 + 0.070\, E_i, \quad N = 2500.$$
$$(0.28) \quad (0.019)$$

- The numbers in the top line (2.44 and 0.070) are called the ***point estimates***.
- They are our best estimates of the values of the parameters a and b in our ***model*** of schooling and wages, $\ln W_i = a + bE_i + u_i$.

Standard Errors

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

- The numbers in parentheses in the second line (0.28 and 0.019) are called the ***standard errors***.
- They are estimates of the likely size of the difference between the point estimates and the true parameter values that would arise just by chance.
- Concretely, under certain assumptions, the difference between the point estimate, $\hat{b}$, and the true parameter value, b, will be approximately normally distributed with a mean of 0 and a standard deviation of 0.019.

Two-Standard Error Confidence Intervals

- The chances that $\hat{b} - b$ will be greater than twice the standard error or less than (that is, more negative than) twice the standard error is small—about 5%.
- Researchers therefore often focus on the ***two-standard error confidence interval*** for b: from the point estimate minus twice the standard error to the point estimate plus twice the standard error.
- It shows is the range of values of *b* for which it would not be surprising to obtain the estimate of $\hat{b}$ that we did.

Two-Standard Error Confidence Intervals

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

- In our case, the two-standard error confidence interval is from 0.070 – 2•0.019 to 0.070 + 2•0.019.

- Thus, it is (0.032,0.108).

Rejecting Hypotheses

- The data provide strong evidence against any proposed value of *b* that is outside the confidence interval (in the sense that we would be quite unlikely to get a value of $\hat{b}$ that far from the true value of *b*).
- For a proposed value of b that is outside the confidence interval, we say that the data ***reject the hypothesis*** ("at the 5 percent level") that b is equal to that value.

Failing to Reject Hypotheses

- The data do not provide strong evidence against any proposed value of *b* that is inside the confidence interval (in the sense that it would not be particularly surprising to get a value of $\hat{b}$ that far from the true value of *b*).
- For a proposed value of b that is inside the confidence interval, we say that the data ***fail to reject the hypothesis*** ("at the 5 percent level") that b is equal to that value.

Two Common Errors in Interpreting Confidence Intervals

- **#1:** Going from the (true) statement that we fail to reject a hypothesis to the (false) statement that we accept a hypothesis.
- (Even if some value of *b* that we are especially interested in lies in the 95% conf. interval, so do many other values. Thus regression results can never point to one specific value of the parameter as being the correct one.)

Two Common Errors in Interpreting Confidence Intervals (cont.)

- **#2:** Going from the (true) statement that 95% of the time $\hat{b}$ will lie within two standard errors of the true value of *b* to the (perhaps false) statement that there is a 95% chance that the true value of *b* is in the 95% confidence interval.
- (The statement that Hypothesis H implies that there is less than a 5% chance will observe Event E is not equivalent to the statement that if we observe Event E, there is less than a 5% chance that Hypothesis H is correct.)

t-statistics

- The ***t-statistic*** is the ratio of the point estimate to the standard error.
- In our example, the t-statistic for b is 0.070/0.019, or 3.68.
- If the true value of b is 0 (and, again, we have a regression that does not suffer from omitted variable bias), the t-statistic will be distributed approximately normally with a mean of 0 and a standard deviation of 1.
- If the t-statistic is greater than 2 or less than −2, we say the coefficient is ***statistically significant***.
- If the t-statistic is between −2 and 2, we say the coefficient is ***statistically insignificant***.

Two Common Errors in Interpreting t-statistics

- **#1:** Going from the (true) statement that a coefficient is statistically insignificant to the (false) statement that the data indicate that the coefficient is zero.
- **#2:** Going from the (true) statement that a coefficient is ***statistically*** significant to the (perhaps false) statement that we have found something that is ***economically*** important.

Example: Which of the Following Would Provide the Best Evidence that Schooling Is Not Important to Wages?

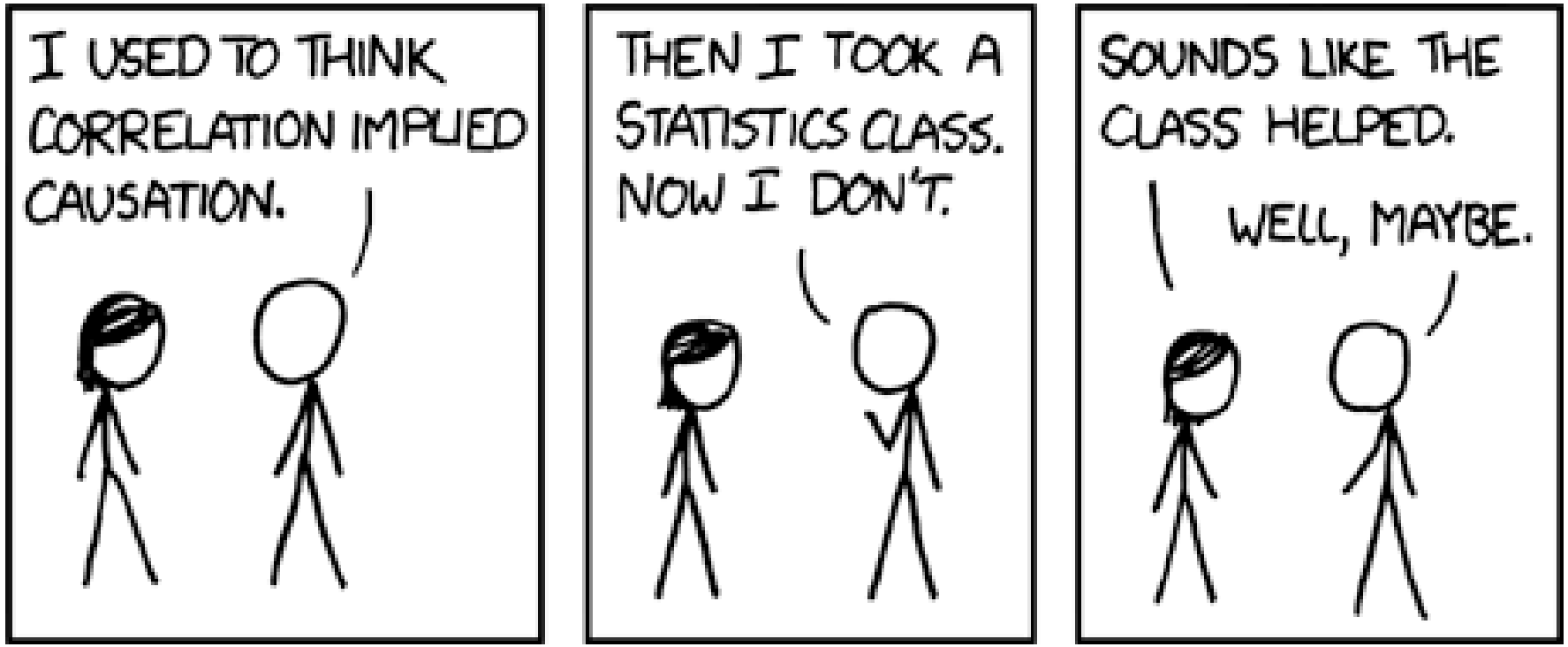
- (a) $\hat{b}$ = 0.20, s.e. = 0.08 (so t = 2.5).
- (b) $\hat{b}$ = 0.0020, s.e. = 0.0004 (so t = 5.0).
- (c) $\hat{b}$ = 0.0010, s.e. = 0.50 (so t = 0.002).
- (d) $\hat{b}$ = −0.05, s.e. = 0.20 (so t = −0.25).

The two-standard error confidence intervals: (0.04,0.36), (0.0012,0.0028), (−0.999,1.001), (−0.45,0.35).

So the correct answer is (b): It provides strong evidence that schooling has some effect on wages—but that the effect is small.

Interpreting Regressions, Boiled Down to Its Essence

- Always focus on point estimates and confidence intervals and their economic interpretation, not on *t*-statistics and statistical significance.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.