Applied Econometrics with EViews for Research

A Step-by-Step Practical Guide for Theses, Journal Articles and Data Analysis

Willian Sebastian Flores Sotelo

Lima, Peru; 2026

Applied Econometrics with EViews for Research
A Step-by-Step Practical Guide for Theses, Journal Articles and Data Analysis
Econometrics without Fear series, book 2
© Willian Sebastian Flores Sotelo. All rights reserved.
Publisher: Saircols S.A.C. (RUC 20607441872), Cal. Garcilazo Mz. A Lote 23, Urb. Retablo I, Comas, Lima, Peru
First digital edition: November 2026.
Legal Deposit at the National Library of Peru No. 2026-13844
Adapted from the Spanish edition, Econometría aplicada con EViews para la investigación (2026), ISBN 978-612-03-3326-6.
No part of this book may be reproduced without the author's written permission, except for brief quotations in reviews or teaching.
The book's .prg programs are provided for educational use. The datasets cited belong to the institutions that produce them.
Digital edition. Distributed by Amazon Kindle Direct Publishing and Hotmart.
EViews is a registered trademark of S&P Global Inc. or its affiliates. This book is an independent work: it is not sponsored, authorized or endorsed by S&P Global. Stata, Python, R and Excel are trademarks of their respective owners.

Contents at a Glance

About the author iv
List of tables xii
Preface: How to use this book xiii
Part One. Before you estimate 1
1. Why does your regression lie? Correlation and causation 2
2. Your data ready in EViews in one afternoon: from ENAHO and the World Bank to a clean dataset 27
3. Hypothesis testing and confidence intervals: the statistics your research needs 61
Part Two. Regression in EViews, step by step 88
4. Simple linear regression, step by step 89
5. Multiple linear regression and omitted variable bias 116
6. How to read a regression without panic 137
7. Regression assumptions: heteroskedasticity, multicollinearity and autocorrelation 160
8. Logistic regression: logit and probit models 194
Part Three. Panels and time series 221
9. Panel data: fixed vs. random effects and the Hausman test 222
10. Time series and the ARIMA model, step by step 254
11. Forecasting with VAR models and cointegration 283
Part Four. Effects, not just associations 316
12. Impact evaluation I: difference in differences 317
13. Impact evaluation II: instrumental variables 345
14. Impact evaluation III: regression discontinuity and propensity score matching 374
Part Five. Reporting, and one step further 403
15. How to write up the results of your thesis or journal article 404
16. ARDL models and the bounds test 429
Appendix A. Eviews · stata · python · r equivalences 456
Appendix B. The book's datasets 463

Contents

About the author iv

List of tables xii

Preface: How to use this book xiii

Part One. Before you estimate 1

1. Why does your regression lie? Correlation and causation 2

◎ WHAT YOU WILL ACHIEVE 2

1.1 Correlation does not tell you why 3

1.2 The question that really matters: what would have happened if…? 3

1.3 Draw your problem before you estimate it 4

1.4 The trap, in a single line 4

1.5 The experiment's trick 5

1.6 What if you cannot run an experiment? 5

▶️ HANDS ON 6

1.7 The mistake almost every thesis student makes 19

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 20

★ SUMMARY AND CONCLUSIONS 20

✓ SOLVED EXERCISES 21

✎ FOR YOUR THESIS OR ARTICLE 25

REFERENCES 26

2. Your data ready in EViews in one afternoon: from ENAHO and the World Bank to a clean dataset 27

◎ WHAT YOU WILL ACHIEVE 27

2.1 Two kinds of data, two different questions 28

2.2 ENAHO: modules, keys and a factor that turns the sample into a country 28

2.3 The World Bank: countries, aggregates and two formats 29

▶️ HANDS ON 29

2.4 The mistake that ruins more datasets than you think 53

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 54

★ SUMMARY AND CONCLUSIONS 54

✓ SOLVED EXERCISES 55

✎ FOR YOUR THESIS OR ARTICLE 59

REFERENCES 60

3. Hypothesis testing and confidence intervals: the statistics your research needs 61

◎ WHAT YOU WILL ACHIEVE 61

3.1 The question: does the countryside pay less? 62

3.2 Your sample is one of a hundred possible samples 62

3.3 The confidence interval: a net, not a bullseye 63

3.4 The hypothesis test: the smoke detector 63

3.5 Three things the p-value keeps quiet .. 64

3.6 Two ways to fail: type I and type II errors 64

3.7 A significant result may not matter 64

▶️ HANDS ON 65

3.8 The mistake almost every thesis student makes with the p-value 81

⚠ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 82
★ SUMMARY AND CONCLUSIONS 82
✓ SOLVED EXERCISES 83
✎ FOR YOUR THESIS OR ARTICLE 86
REFERENCES 87
Part Two. Regression in EViews, step by step. 88
4. Simple linear regression, step by step 89
◎ WHAT YOU WILL ACHIEVE 89
4.1 From the dummy to the line 90
4.2 Which line does EViews choose? 90
4.3 Why the logarithm of income? 91
4.4 From the logarithm to a percentage ... 91
4.5 The R²: how much of the cloud the line captures 91
4.6 What the line sees and what it leaves out 91
▶ HANDS ON 92
4.7 The mistake that turns a slope into a promise 109
⚠ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 110
★ SUMMARY AND CONCLUSIONS 110
✓ SOLVED EXERCISES 111
✎ FOR YOUR THESIS OR ARTICLE 114
REFERENCES 115
5. Multiple linear regression and omitted variable bias 116
◎ WHAT YOU WILL ACHIEVE 116
5.1 What changes when you include two variables at once? 117
5.2 The variable that is missing but still counts 117
5.3 Dummies: always compared with someone 118
5.4 Is the R² useful for choosing the model? 118
▶ HANDS ON 119
5.5 The mistake almost every thesis student makes: "my coefficient survived the controls" 129
⚠ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 130
★ SUMMARY AND CONCLUSIONS 130
✓ SOLVED EXERCISES 131
✎ FOR YOUR THESIS OR ARTICLE 135
REFERENCES 136
6. How to read a regression without panic ... 137
◎ WHAT YOU WILL ACHIEVE 137
6.1 One window, three blocks 137
6.2 What does a coefficient say? It depends on the units 138
6.3 The p-value answers a small question 139
6.4 Several coefficients at once: the Wald test 139
6.5 Comparing without cheating: beta, Akaike and Schwarz 139
▶ HANDS ON 140
6.6 The mistake almost every thesis student makes when reading the output: the traffic light of stars 153
⚠ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 154
★ SUMMARY AND CONCLUSIONS 154
✓ SOLVED EXERCISES 155
✎ FOR YOUR THESIS OR ARTICLE 158

REFERENCES 159

7. Regression assumptions: heteroskedasticity, multicollinearity and autocorrelation 160

◎ WHAT YOU WILL ACHIEVE 160

7.1 Some assumptions protect the coefficient; others, its standard error 161

7.2 Heteroskedasticity: the error does not shake the same for everyone 161

7.3 Clustered errors: when neighbors look alike 162

7.4 Multicollinearity: two variables telling the same story 162

7.5 Autocorrelation: this year's error remembers last year's 163

7.6 Normality and specification: the assumption that scares the most and the one that matters most 163

▶️ HANDS ON 164

7.7 The mistake almost every thesis student makes with the assumptions: the four-test appendix 185

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 186

★ SUMMARY AND CONCLUSIONS 186

✓ SOLVED EXERCISES 187

✎ FOR YOUR THESIS OR ARTICLE 191

REFERENCES 192

8. Logistic regression: logit and probit models 194

◎ WHAT YOU WILL ACHIEVE 194

8.1 What if your dependent variable can only be 0 or 1? 195

8.2 Bending the line: logit and probit 195

8.3 Three ways to read a logit 196

8.4 How good is your model? 196

▶️ HANDS ON 197

8.5 The mistake almost every thesis student makes with the logit 214

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 215

★ SUMMARY AND CONCLUSIONS 215

✓ SOLVED EXERCISES 215

✎ FOR YOUR THESIS OR ARTICLE 219

REFERENCES 220

Part Three. Panels and time series 221

9. Panel data: fixed vs. random effects and the Hausman test 222

◎ WHAT YOU WILL ACHIEVE 222

9.1 What does a panel have that a survey does not? 223

9.2 Fixed effects: each country against its own average 223

9.3 Shocks common to the whole region 224

9.4 Random effects: subtract only a part . 224

9.5 The Hausman test and what it really compares 224

▶️ HANDS ON 225

9.6 The mistake almost every thesis student makes when choosing between fixed and random effects 248

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 249

★ SUMMARY AND CONCLUSIONS 249

✓ SOLVED EXERCISES 250

✎ FOR YOUR THESIS OR ARTICLE 252

REFERENCES 253

10. Time series and the ARIMA model, step by step 254
◎ WHAT YOU WILL ACHIEVE 254
10.1 In a series, the order of the data is the data 255
10.2 Trend and seasons: the two enemies of the mean 255
10.3 Do you need to difference? Two tests, two opposite hypotheses 255
10.4 The pieces of ARIMA and its seasonal version 256
10.5 Identify, estimate, diagnose, forecast 256
▶ HANDS ON 257
10.6 The mistake almost every thesis student makes with ARIMA 277
⚠ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 277
★ SUMMARY AND CONCLUSIONS 278
✓ SOLVED EXERCISES 278
✎ FOR YOUR THESIS OR ARTICLE 281
REFERENCES 282
11. Forecasting with VAR models and cointegration 283
◎ WHAT YOU WILL ACHIEVE 283
11.1 A VAR: each variable with its own past and that of the others 283
11.2 Before reading anything: lags, stability and clean residuals 284
11.3 Granger causality: the past that helps to forecast 284
11.4 Impulse response and variance decomposition 285
11.5 Random walks and spurious regression 285
11.6 Cointegration and error correction 285
11.7 Two ways to test for cointegration 286
▶ HANDS ON 286
11.8 The mistake almost every thesis student makes with VAR and cointegration 311
⚠ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 312
★ SUMMARY AND CONCLUSIONS 312
✓ SOLVED EXERCISES 313
✎ FOR YOUR THESIS OR ARTICLE 314
REFERENCES 315
Part Four. Effects, not just associations 316
12. Impact evaluation I: difference in differences 317
◎ WHAT YOU WILL ACHIEVE 317
12.1 Two subtractions that, on their own, lie 318
12.2 The double subtraction, in four cells 318
12.3 The assumption that holds everything up: parallel trends 318
12.4 Standard errors and the problem of a single treated unit 319
12.5 Many periods: two-way fixed effects and event studies 319
12.6 When treatment arrives at different dates 320
▶ HANDS ON 320
12.7 The mistake thesis students make when evaluating a policy with DiD 339

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 340
★ SUMMARY AND CONCLUSIONS 340
✓ SOLVED EXERCISES 341
✒ FOR YOUR THESIS OR ARTICLE 343
REFERENCES 343
13. Impact evaluation II: instrumental variables 345
◎ WHAT YOU WILL ACHIEVE 345
13.1 Three paths through which endogeneity slips in 346
13.2 The instrument: a lever that touches only the cause 346
13.3 The Wald calculation: dividing two effects 347
13.4 Two-stage least squares 347
13.5 Whose effect is it? The compliers 348
13.6 Weak instruments: when the lever barely moves 348
13.7 Two more tests: endogeneity and overidentification 349
▶️ HANDS ON 349
13.8 The mistake thesis students make with instrumental variables 368
⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 368
★ SUMMARY AND CONCLUSIONS 369
✓ SOLVED EXERCISES 369
✒ FOR YOUR THESIS OR ARTICLE 373
REFERENCES 373
14. Impact evaluation III: regression discontinuity and propensity score matching 374
◎ WHAT YOU WILL ACHIEVE 374
14.1 A rule with a cutoff 375
14.2 The only thing that may change at the cutoff 375
14.3 Sharp or fuzzy? 375
14.4 How close, and with what line 375
14.5 When people sort around the cutoff 376
14.6 Hands on: PANES and support for the government 376
14.7 No cutoff, no lottery: looking for pairs 387
14.8 Hands on: matching step by step 388
14.9 Which method should I choose? 396
14.10 The mistake thesis students make with discontinuity and matching 397
⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 397
★ SUMMARY AND CONCLUSIONS 398
✓ SOLVED EXERCISES 398
✒ FOR YOUR THESIS OR ARTICLE 401
REFERENCES 401
Part Five. Reporting, and one step further 403
15. How to write up the results of your thesis or journal article 404
◎ WHAT YOU WILL ACHIEVE 404
15.1 What your reader needs, in order 405
15.2 The methods section, in five paragraphs 405
15.3 A table that stands on its own 406
15.4 How each number is reported 406
15.5 What your numbers do not say 407
▶️ HANDS ON 407

15.6 The mistake almost every thesis student makes: pasting the EViews window 423

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 423

★ SUMMARY AND CONCLUSIONS 424

✓ SOLVED EXERCISES 424

✒️ FOR YOUR THESIS OR ARTICLE 427

REFERENCES 427

16. ARDL models and the bounds test 429

◎ WHAT YOU WILL ACHIEVE 429

16.1 Why another cointegration model? 429

16.2 An ARDL is a regression with memory 430

16.3 The error correction form 430

16.4 The bounds test: two bounds instead of one 431

16.5 Before you believe it: assumptions, residuals and stability 432

▶️ HANDS ON 432

16.6 The mistake almost every thesis student makes with ARDL 450

⚠️ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES 451

★ SUMMARY AND CONCLUSIONS 451

✓ SOLVED EXERCISES 452

✒️ FOR YOUR THESIS OR ARTICLE 454

REFERENCES 454

Appendix A. Eviews · stata · python · r equivalences 456

Appendix B. The book's datasets 463

1. Why does your regression lie? Correlation and causation

A few decades ago, epidemiologists had a promising lead against lung cancer. People who ate plenty of fruits and vegetables rich in carotenoids, and people with high levels of beta-carotene in their blood, got sick less often. Study after study said the same thing.

What else was needed? If beta-carotene protects, you only have to give it in a pill to those who need it most: smokers.

In southwestern Finland, someone decided to test it the right way. The study enrolled 29,133 male smokers aged 50 to 69. The men were not allowed to choose what to take. A lottery decided: some received vitamin E, others beta-carotene, others both and others a pill with nothing active in it. Neither the men nor the staff caring for them knew who had received what. They were followed for five to eight years.

The results came out in 1994. Among the men who received beta-carotene there was **more** lung cancer, not less: 18% more than among those who did not receive it. Their total mortality was also higher, by 8%. The authors themselves called the finding unexpected (Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group, 1994).

Years of studies said "it protects." The experiment said "it does not protect, and it may harm." How could so much evidence be wrong at the same time?

Think about who has high beta-carotene in their blood. Usually, someone who eats well. And people who eat well tend to differ in many other ways: how much they smoke, how much they move, how much they earn, how often they see a doctor. The observational studies compared different people and credited the whole difference to a single molecule. The lottery, by contrast, compared men who were alike in everything except the pill.

Now think about your own research. You have a survey, a variable you care about and software that runs a regression in two seconds. You get a positive coefficient with three stars. How do you know the same thing is not happening to you?

That is what this chapter is about. And, deep down, that is what this whole book is about.

◎ WHAT YOU WILL ACHIEVE

By the end of this chapter you will be able to:

- Tell when a number describes a relationship and when it measures an effect.
- Explain, with a single equation, why a simple comparison can fool you.
- Simulate in EViews, step by step, one case where the regression lies and another where it tells the truth.
- Recognize in your own thesis the question your committee will ask first.

1.1 Correlation does not tell you why

Let us start with what you already know. Two variables are correlated when they move together: when one goes up, the other tends to go up (or down). Correlation is useful, because it tells you that something is there. It does not tell you **what** is there.

Franz Messerli, a cardiologist in New York, made the point with an example that went around the world. He gathered data from 23 countries and compared chocolate consumption per person with the number of Nobel prizes each country had won per person. The relationship came out strong and highly significant, with Switzerland at the top of both lists.

Messerli published the paper in one of the most widely read medical journals in the world. His purpose was to warn how easy it is to take seriously a correlation that makes no sense (Messerli, 2012).

Why does this happen? When two variables move together, there are four possible explanations. Keep them close at hand, because your committee will.

1. **Causation.** X causes Y. For example, more fertilizer raises the yield of a plot, and a frost shrinks the potato harvest.

2. **Reverse causation.** Y causes X. Imagine you find that the firms that spend the most on advertising are the ones that sell the most. Does advertising generate the sales? It may, but many firms set their advertising budget as a share of what they sell. So selling more makes them spend more on advertising, and not only the other way around.

3. **A third variable.** Something else moves both. In the chocolate case, an obvious candidate is the country's wealth: a rich country consumes more chocolate and also funds more universities and laboratories. In the beta-carotene story, eating well went hand in hand with many other healthy habits. Econometricians call that third variable a **confounder** or **confounding variable**.

4. **Chance.** For example, quinoa exports and the rainfall in a distant city can move together for a few years without having anything to do with each other. With enough variables, some will be correlated by pure coincidence. If you cross a hundred pairs of unrelated variables, about five will come out "significant" at the 5% level. That is how the threshold works.

Notice something uncomfortable. All four explanations produce **the same table** in your statistics software. A coefficient, a standard error, a few stars. The software does not know which of the four is true. You have to.

1.2 The question that really matters: what would have happened if...?

Let us move to a closer case. Suppose Diego is writing his thesis on school performance in Trujillo, a city on Peru's northern coast. He has data on 5,000 students in their final year of high school. He knows their final grades, on the 0-to-20 scale used in Peruvian schools, and whether or not they paid for private tutoring. Those who paid for tutoring scored, on average, three points higher. Does tutoring raise grades by three points?

The question Diego wants to answer is not "who got higher grades?" It is a harder one: **what grade would each student have earned without tutoring?**

That unseen grade has a name. It is called a **potential outcome**. Each student has two:

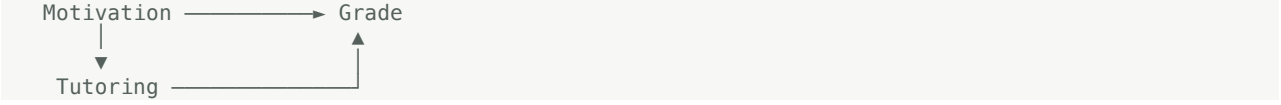
- The grade they would earn **with** private tutoring.
- The grade they would earn **without** private tutoring.

The effect of tutoring for that student is the difference between the two. Simple. The problem is that life never lets us see both. Each student either took tutoring or did not. One of their two potential outcomes is left forever in the realm of "what would have happened if…?" Textbooks call this the **fundamental problem of causal inference** (Cunningham, 2021).

So nothing can be measured? It can, though not student by student: it is measured on average, and with care. And the care begins with understanding where the trap is.

1.3 Draw your problem before you estimate it

Before you open EViews, take a sheet of paper and draw arrows. It looks like child's play, but it is one of the most serious tools of modern causal inference. Each arrow means "this causes that." Diego's problem, drawn out, looks like this:



The bottom arrow, from Tutoring to Grade, is the one Diego wants to measure. The other two are the problem: motivation pushes students toward tutoring and, along a different path, also pushes their grades up. That alternative path, running through motivation, is called a **backdoor path**. As long as it stays open, part of the relationship between tutoring and grades travels along it, and your regression mistakes it for the effect of tutoring (Cunningham, 2021).

Drawing forces you to state what you believe about the world. If someone tells you "you forgot to account for family income," the two of you are no longer arguing about stars and p-values: you are arguing about an arrow. Does family income change the chance of paying for tutoring? Surely. Does it move grades along another path? Probably, because a household with more income also buys books, internet access and peace of mind. Then it is one more backdoor you have to close.

There are three ways to close a backdoor, and this book covers all of them. You can measure the confounder and control for it, use data in which the confounder does not change, or find a source of variation that does not run through it. A lottery, for instance. For now, just draw. You will see that half of your committee's questions can be answered by looking at that drawing.

1.4 The trap, in a single line

When Diego compares the average grade of those who took tutoring with that of those who did not, he is computing this:

> **Observed difference = causal effect + selection bias**

This equation is the most important one in the book. Angrist and Pischke put it at the center of their textbook, and rightly so (Angrist & Pischke, 2014). Read it slowly.

The **causal effect** is what Diego wants: how much tutoring raises the grades of those who took it.

The **selection bias** is the difference that already existed between the two groups before any tutoring. Who pays for private tutoring? The most motivated students. Those whose parents keep an eye on their grades. Those with some extra money at home. Those students would have earned higher grades **even without a single tutoring session**.

If selection bias is large, the observed difference can be three times the true effect. Or it can have the opposite sign, as in the beta-carotene story. And no p-value will warn you.

That is why econometricians always ask the same uncomfortable question: **were the two groups comparable before the treatment?** If the answer is no, your difference mixes two things and you do not know how much each one weighs.

1.5 The experiment's trick

How do you get rid of selection bias? With a coin.

If Diego could hold a lottery to decide who gets private tutoring, motivation, money and attentive parents would end up spread more or less equally across the two groups. Nobody chose. Chance chose. Selection bias drops to zero and the observed difference measures the causal effect. That is what the Finnish study did, and that is why it found what years of observational studies had missed.

Latin America has its own famous example.

Colombia had a program, PACES, that handed out vouchers to pay part of the tuition at a private secondary school. The vouchers were renewed only if the student kept up satisfactory academic progress. And the vouchers were assigned **by lottery** (Angrist et al., 2002).

Because chance decided who won, winners and losers were comparable, and the difference between them could be read as the effect of the voucher. Three years after the lotteries, winners were about 10 percentage points more likely to have finished 8th grade, mainly because they repeated fewer grades. They also scored 0.2 standard deviations higher on achievement tests (Angrist et al., 2002).

It all started with a lottery.

1.6 What if you cannot run an experiment?

Here comes the bad news: you almost never will. You cannot randomly assign who earns a master's degree, who lives in Lima, the capital, or who gets a loan. Your research probably uses a survey that already exists, such as Peru's National Household Survey (ENAHO), run by the National Institute of Statistics and Informatics (INEI). Or it uses company data, or central bank time series.

The good news is that econometrics is precisely the art of getting close to the experiment when you do not have one. This book covers the tools in order of difficulty:

- **Control for what you do observe** (chapters 4 to 8). If the confounder is in your data, you put it in the regression.
- **Follow the same people or firms over time** (chapter 9). Panel data remove whatever does not change, even if you never measure it.
- **Look for something that resembles chance** (chapters 12 to 14). Instrumental variables, difference-in-differences, regression discontinuity and matching.

Every tool rests on an assumption. Your job, and this book's, is to know which one it is and to defend why it holds in your case.

Enough theory. Let us watch the trap with our own eyes.

▶ HANDS ON

Open EViews. At the top you will see the **Command** window: you type an instruction there, press Enter and EViews runs it. Copy the blocks in order, line by line. If you prefer to have them all together, open the file `ch01.prg` from the book's repository with *File › Open › Programs* and click the *Run* button. The preface explains how to download it. You do not need to install anything extra: everything we use ships with EViews.

We are going to do something that is impossible in real life: **create a world where we know the truth**. We will simulate Diego's 5,000 students. We will decide that private tutoring raises the grade by exactly **1 point**. Then we will see what each method tells us. The data in this section are simulated; they do not come from any survey.

The key is a variable that no survey measures: motivation. In our world, motivation does two things. It raises the grade. And it increases the chance of paying for tutoring.

Step 1. Create 5,000 students with a hidden motivation

```
wfcreate(wf=ch01) u 5000
rndseed 2026
series motivation = nrnd
show motivation.hist
```

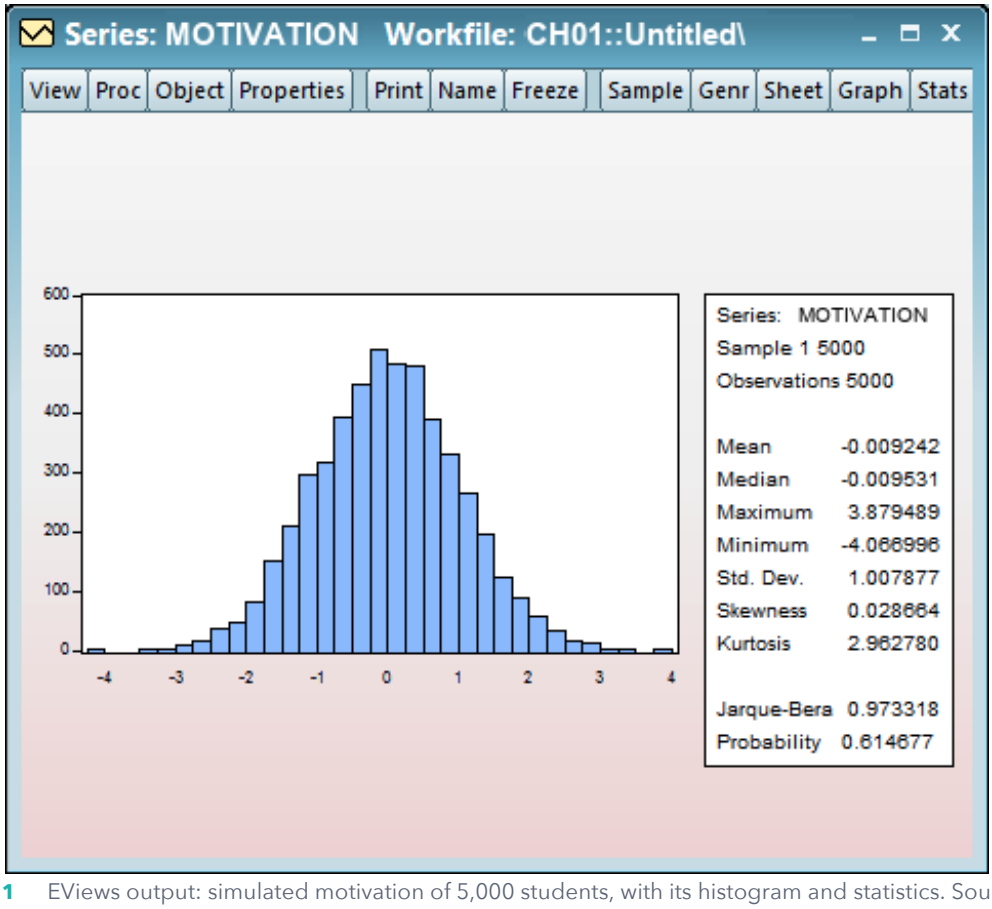


FIGURE 1.1 EViews output: simulated motivation of 5,000 students, with its histogram and statistics. Source: author's own work with EViews 14 and simulated data (ch01.prg).

What did we do? `wfcreate` creates a **workfile**, EViews' working file. `wf=ch01` gives it a name, `u` says the data have no dates (they are people, not months) and `5000` creates 5,000 rows, one per student.

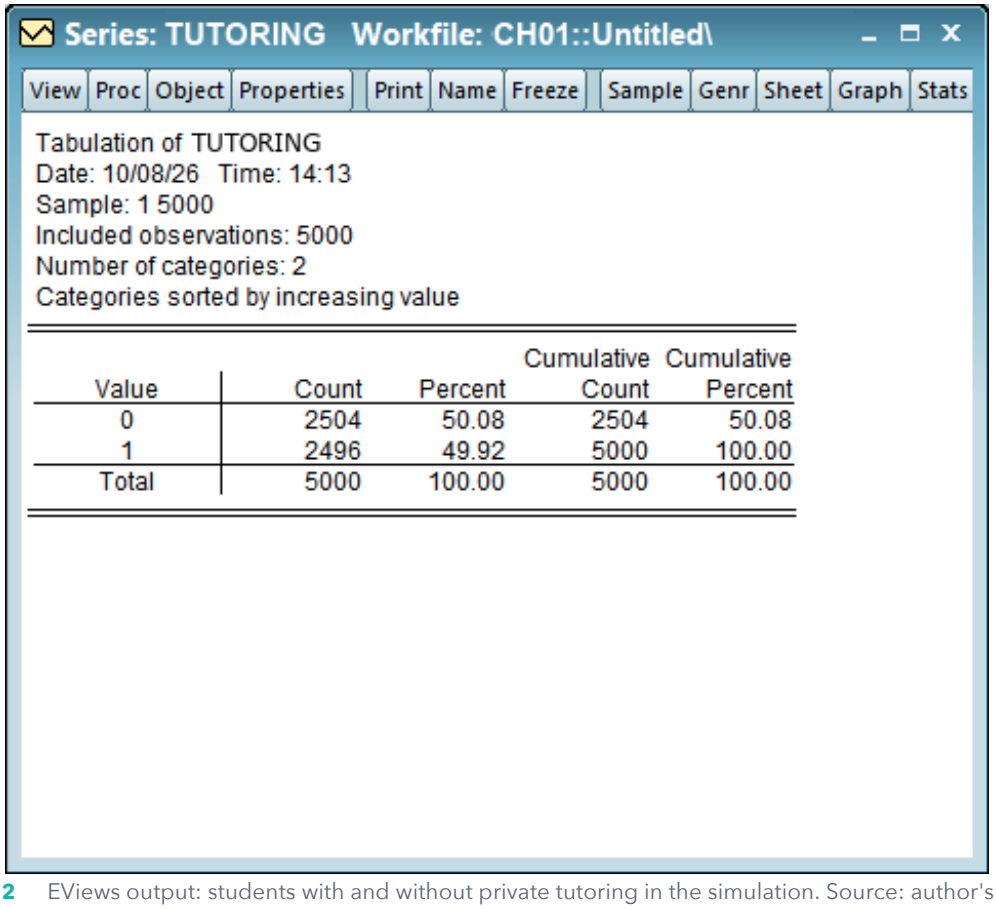
`rndseed 2026` fixes the seed of the random number generator: thanks to it, you will get exactly the same numbers that appear in the book. The `nrnd` function gives each student a random motivation, with mean 0 and standard deviation 1. Finally, `show motivation.hist` opens the series in its histogram view.

The window confirms what we asked for: a mean of −0.009 and a standard deviation of 1.008. No survey measures this motivation, but in our world it exists.

A practical tip: run the blocks once, and in order. If you repeat a block that draws random numbers, the sequence changes and your figures will no longer match the book's. If that happens, start again from step 1.

Step 2. Make the most motivated students pay for tutoring

```
series tutoring = @rbinom(1, @logit(1.5*motivation))
show tutoring.freq
```



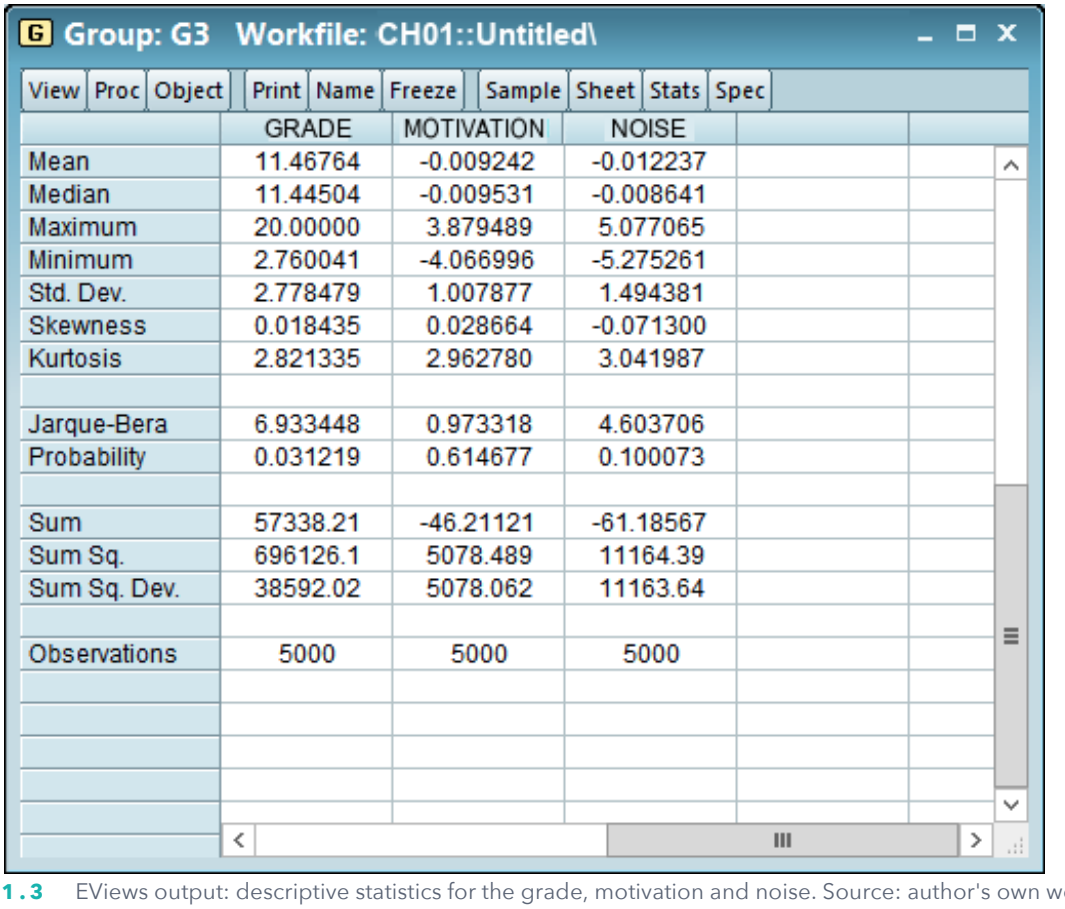
Value	Count	Percent	Cumulative Count	Cumulative Percent
0	2504	50.08	2504	50.08
1	2496	49.92	5000	100.00
Total	5000	100.00	5000	100.00

FIGURE 1.2 EViews output: students with and without private tutoring in the simulation. Source: author's own work with EViews 14 and simulated data (ch01.prg).

This line is the heart of the problem. The `@logit()` function turns motivation into a probability between 0 and 1. For example, a student with motivation 0 has a 50% chance of paying for tutoring, and a highly motivated one gets close to 100%. Then `@rbinom(1, p)` flips a coin loaded with that probability. It records 1 if the student paid for tutoring and 0 if not. The `freq` view counts each type: 2,496 students (49.92%) took tutoring and 2,504 did not.

Step 3. Build the grade with a true effect of 1 point

```
series noise = 1.5*nrnd
series grade = 11 + 1*tutoring + 2*motivation + noise
grade = @recode(grade<0, 0, @recode(grade>20, 20, grade))
group g3 grade motivation noise
show g3.stats
```



Group: G3 Workfile: CH01::Untitled\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

	GRADE	MOTIVATION	NOISE
Mean	11.46764	-0.009242	-0.012237
Median	11.44504	-0.009531	-0.008641
Maximum	20.00000	3.879489	5.077065
Minimum	2.760041	-4.066996	-5.275261
Std. Dev.	2.778479	1.007877	1.494381
Skewness	0.018435	0.028664	-0.071300
Kurtosis	2.821335	2.962780	3.041987
Jarque-Bera	6.933448	0.973318	4.603706
Probability	0.031219	0.614677	0.100073
Sum	57338.21	-46.21121	-61.18567
Sum Sq.	696126.1	5078.489	11164.39
Sum Sq. Dev.	38592.02	5078.062	11163.64
Observations	5000	5000	5000

FIGURE 1.3 EViews output: descriptive statistics for the grade, motivation and noise. Source: author's own work with EViews 14 and simulated data (ch01.prg).

Read the second line slowly, because it holds the truth we will later try to uncover. The grade starts at 11 points. It adds **1 point** if the student took tutoring, 2 points for each unit of motivation and a little noise, which stands for luck on exam day. The third line trims the grade so it never falls below 0 or rises above 20. `@recode(condition, a, b)` returns `a` if the condition holds and `b` if it does not.

The last two lines introduce a new object. A **group** gathers several series to view them side by side, and its `stats` view computes the statistics of all of them at once. The summary shows an average grade of 11.47 and an average motivation of practically zero (−0.009).

Step 4. First attempt: compare means

This is what most people would do. Split into two groups and compare.

```
show grade.statby tutoring
```

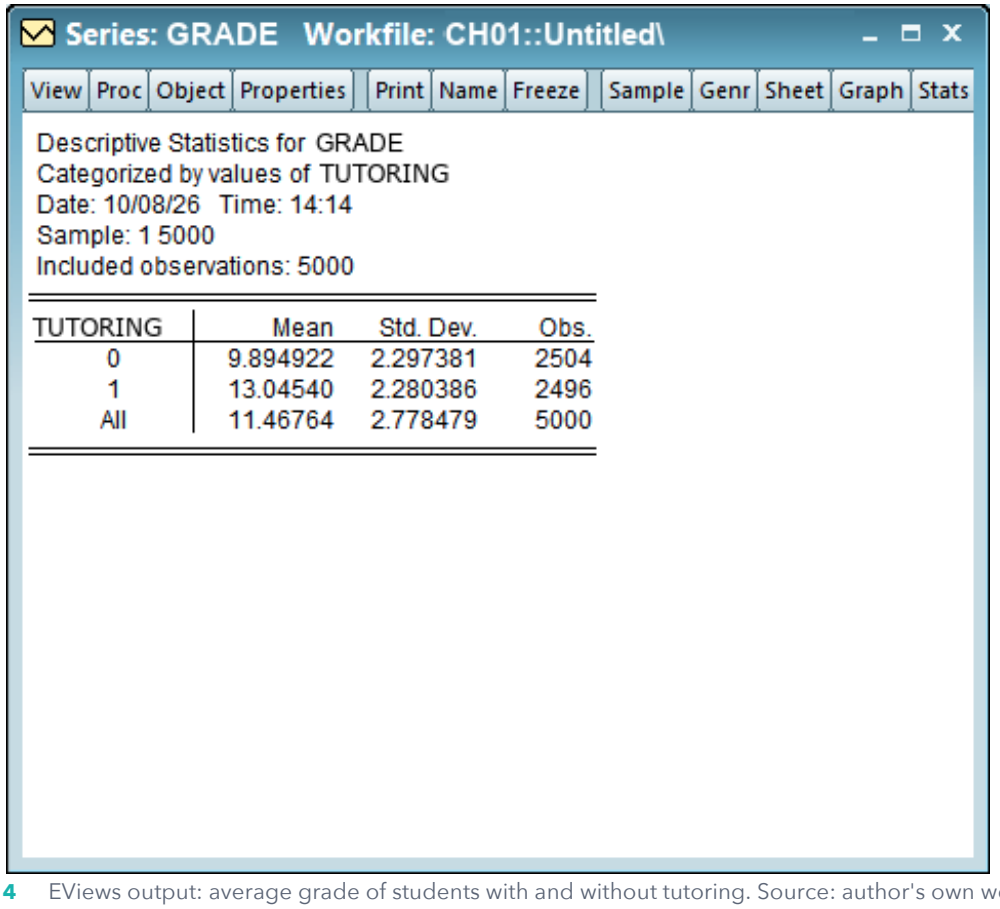


FIGURE 1.4 EViews output: average grade of students with and without tutoring. Source: author's own work with EViews 14 and simulated data (ch01.prg).

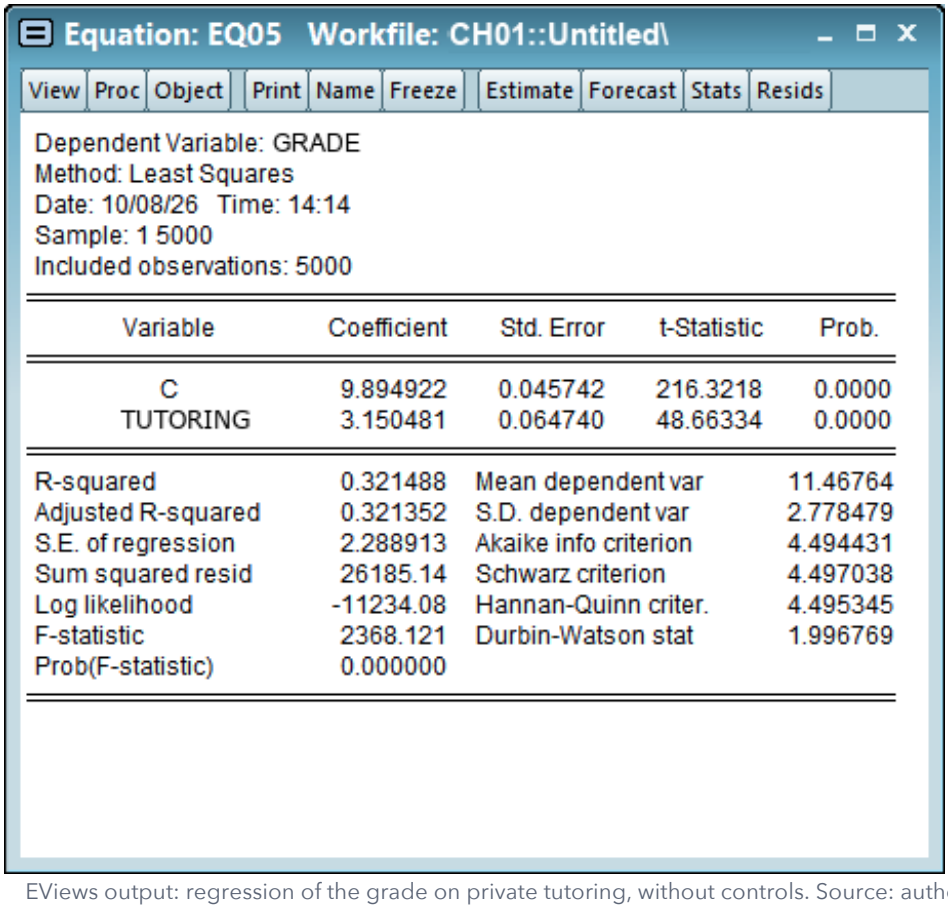
The `statby` view computes the statistics of one series for each value of another. Students without tutoring average 9.89; those with tutoring, 13.05. The difference is 3.15 points. We know the truth is one. The simple comparison tripled the effect.

A reading tip: the screen shows five decimals and the text rounds to two, so 13.04540 on the screen is 13.05 in the prose.

Step 5. Second attempt: the regression

"Fine," you will say, "but I don't compare means, I run a regression." Let us see.

```
equation eq05.ls grade c tutoring
show eq05
```



Equation: EQ05 Workfile: CH01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GRADE
Method: Least Squares
Date: 10/08/26 Time: 14:14
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.894922	0.045742	216.3218	0.0000
TUTORING	3.150481	0.064740	48.66334	0.0000

R-squared	0.321488	Mean dependent var	11.46764
Adjusted R-squared	0.321352	S.D. dependent var	2.778479
S.E. of regression	2.288913	Akaike info criterion	4.494431
Sum squared resid	26185.14	Schwarz criterion	4.497038
Log likelihood	-11234.08	Hannan-Quinn criter.	4.495345
F-statistic	2368.121	Durbin-Watson stat	1.996769
Prob(F-statistic)	0.000000		

FIGURE 1.5 EViews output: regression of the grade on private tutoring, without controls. Source: author's own work with EViews 14 and simulated data (ch01.prg).

The first line creates an **equation** object called `eq05` and estimates it by least squares (`ls`). After the dot comes the dependent variable and then the explanatory ones; `c` is the constant. EViews does not add it on its own: if you forget it, you estimate a line through the origin.

The coefficient on `TUTORING` is 3.15. The p-value (column `Prob.`) is 0.0000. And the confidence interval? EViews does not show it in this table, but it is one command away.

```
eq05.cinterval 0.95
```

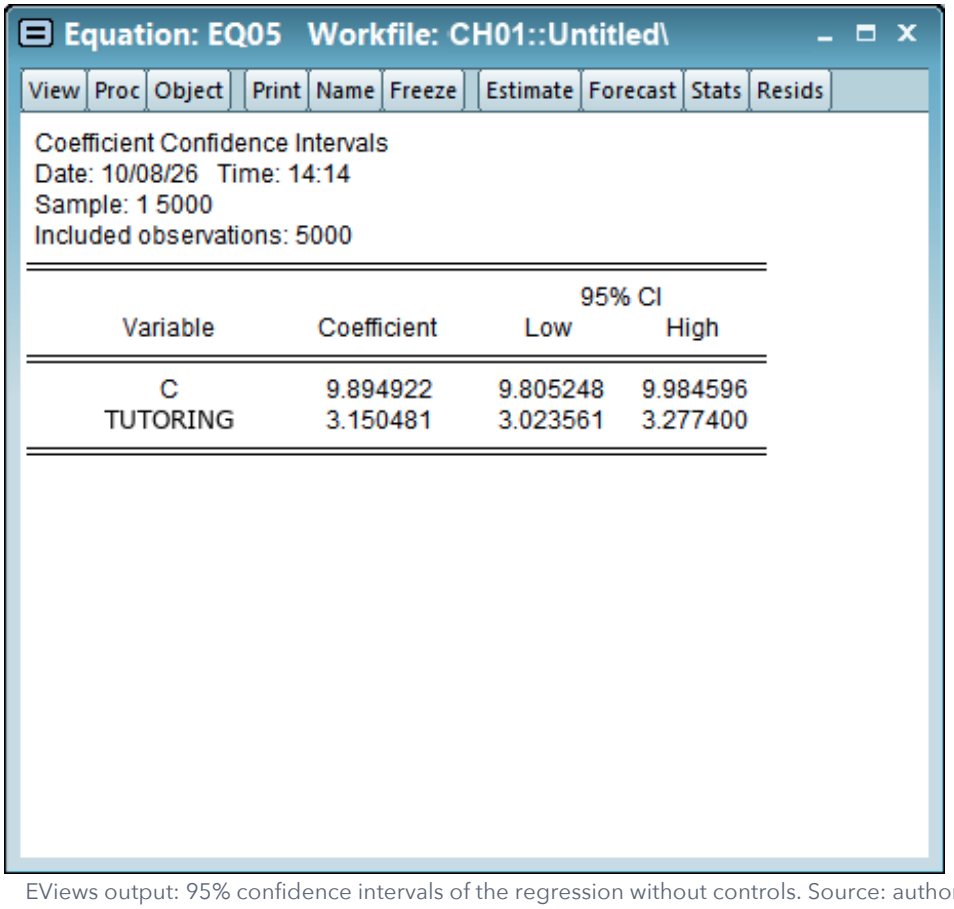


FIGURE 1.6 EViews output: 95% confidence intervals of the regression without controls. Source: author's own work with EViews 14 and simulated data (ch01.prg).

The 95% confidence interval runs from 3.02 to 3.28: it does not even come close to the true 1. It is the most convincing table you will ever see, and it is wrong.

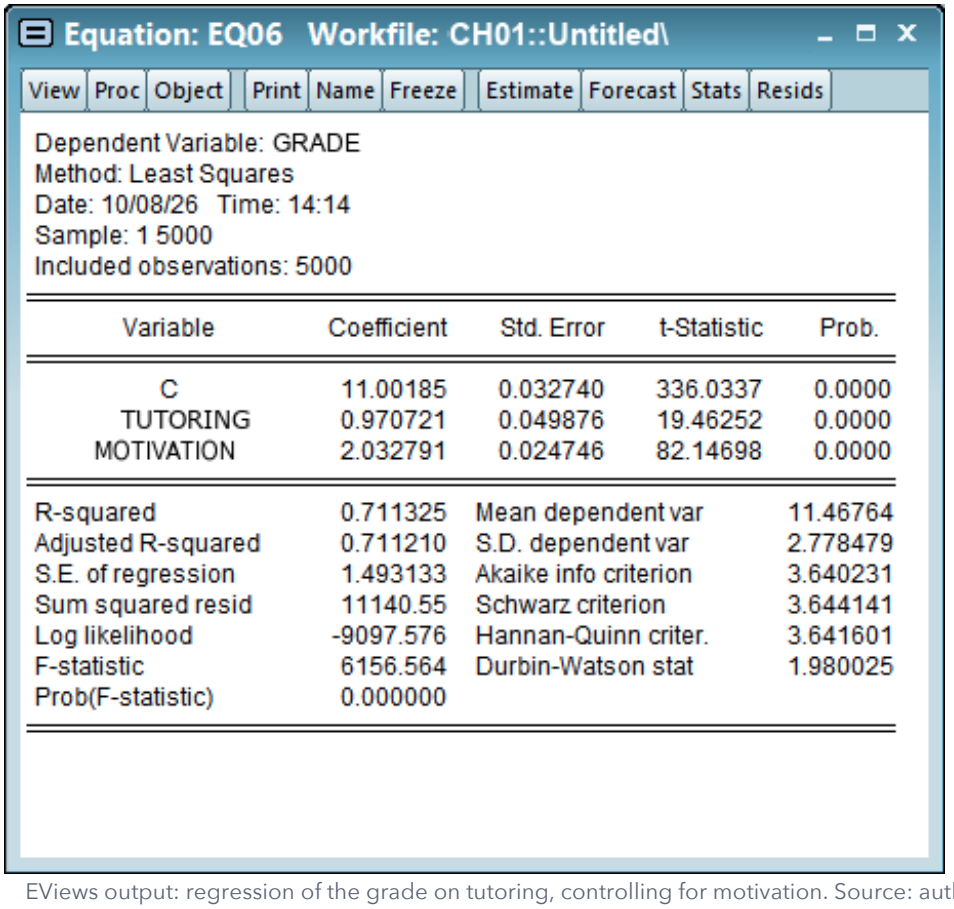
Look also at the constant, `C`: 9.89. It is exactly the average of the students without tutoring in step 4. And 9.89 plus 3.15 gives 13.05: the average of those with tutoring.

Stop here for a moment, because this is the lesson of the chapter. **A regression with a single variable is exactly the same comparison of means**, just with more decimals. Neither the p-value nor the sample size corrects selection bias. With 50,000 students, the error would be just as large; you would only be more certain of it.

Step 6. Third attempt: if you could measure motivation

In our simulated world we have something Diego does not: each student's motivation. Let us put it in the regression.

```
equation eq06.ls grade c tutoring motivation
show eq06
```



Equation: EQ06 Workfile: CH01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GRADE
Method: Least Squares
Date: 10/08/26 Time: 14:14
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.00185	0.032740	336.0337	0.0000
TUTORING	0.970721	0.049876	19.46252	0.0000
MOTIVATION	2.032791	0.024746	82.14698	0.0000

R-squared	0.711325	Mean dependent var	11.46764
Adjusted R-squared	0.711210	S.D. dependent var	2.778479
S.E. of regression	1.493133	Akaike info criterion	3.640231
Sum squared resid	11140.55	Schwarz criterion	3.644141
Log likelihood	-9097.576	Hannan-Quinn criter.	3.641601
F-statistic	6156.564	Durbin-Watson stat	1.980025
Prob(F-statistic)	0.000000		

FIGURE 1.7 EViews output: regression of the grade on tutoring, controlling for motivation. Source: author's own work with EViews 14 and simulated data (ch01.prg).

Now the coefficient on `TUTORING` is 0.97, with a standard error of 0.05: the true 1 lies less than one standard error away. The regression stopped lying. What changed? We gave it the confounder. By comparing students with **the same** motivation, the selection bias disappeared. That is what "controlling" for a variable means, and you will study it in depth in chapter 5.

Look also at the coefficient on `MOTIVATION`: 2.03. It is almost exactly what we put in when we created the data.

Step 7. Fourth attempt: the experiment

What if, instead of letting each student choose, we held a lottery for tutoring? We will use the same students, with the same motivation and the same luck on exam day. The only thing that changes is who decides: now a coin decides.

```
series tutoring_rand = @rbinom(1, 0.5)
series grade_rand = 11 + 1*tutoring_rand + 2*motivation + noise
grade_rand = @recode(grade_rand<0, 0, @recode(grade_rand>20, 20, grade_rand))
equation bal07.ls motivation c tutoring_rand
show bal07
```

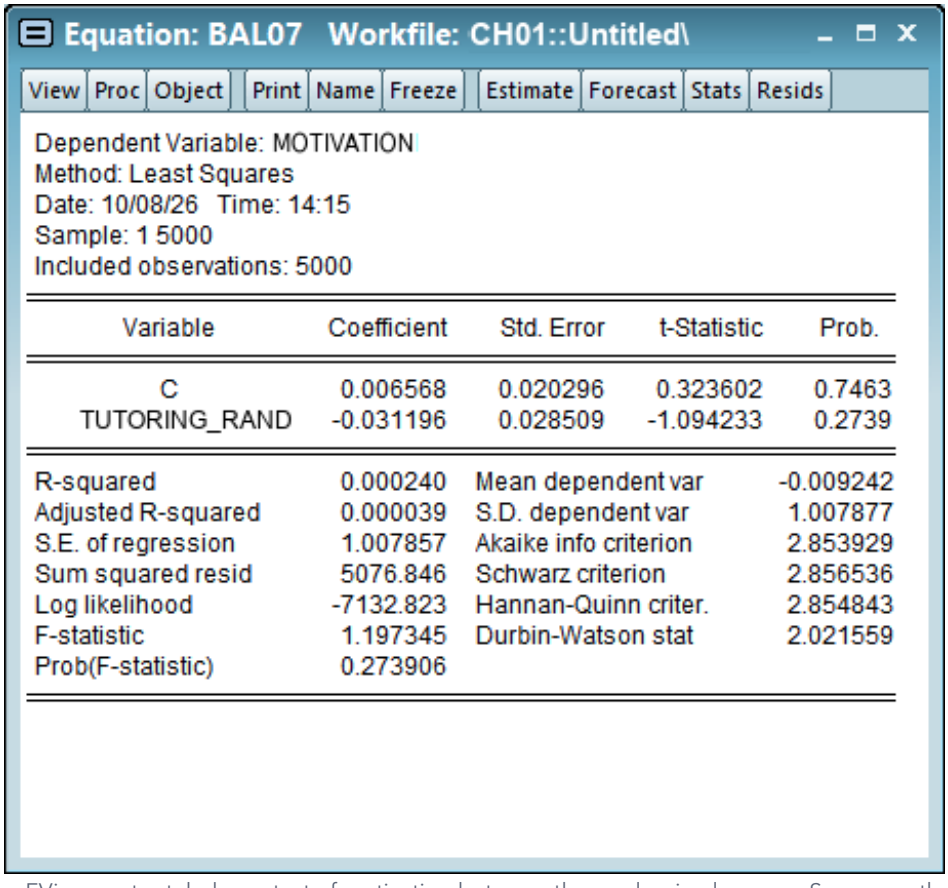


FIGURE 1.8 EViews output: balance test of motivation between the randomized groups. Source: author's own work with EViews 14 and simulated data (ch01.prg).

Before estimating the grade, we check **balance**: do the two randomized groups have the same motivation? To find out, we use the trick from step 5 in reverse. A regression of motivation on the lottery variable compares the means of the two groups. The constant is the mean of the group without tutoring, and the coefficient is the difference.

The difference is a mere −0.03, and its p-value is 0.27. There is no difference worth the name. If you prefer the classic t test, the view `motivation.testby tutoring_rand` gives you the same t statistic, with the sign flipped.

The lottery spread motivation evenly **without our having to measure it**. Now, the simple regression.

```
equation eq08.ls grade_rand c tutoring_rand
show eq08
```

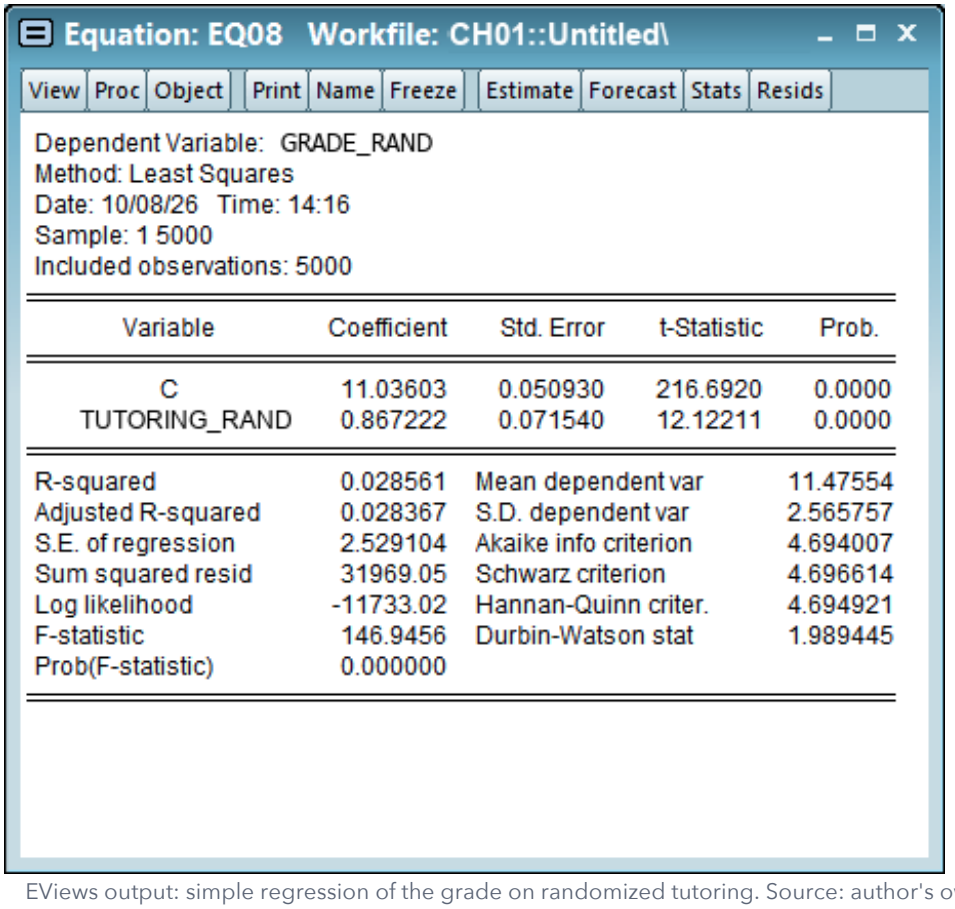


FIGURE 1.9 EViews output: simple regression of the grade on randomized tutoring. Source: author's own work with EViews 14 and simulated data (ch01.prg).

Without any control, the coefficient is 0.87, with a standard error of 0.07. It lies less than two standard errors from the true 1: the 95% confidence interval contains it. Chance did the work that the control did in the third attempt. Why is it not exactly 1.00? Because a single lottery never spreads everything perfectly. With another seed it would come out a little above or a little below.

Look at the R-squared: only 0.03. The regression explains little of the grade, because motivation and luck weigh heavily. Even so, the effect is well measured. A low R-squared does not mean the coefficient is wrong. Imagine a program that raises every student's grade by one point: it would be a great program, even though it explains little about why some students score 8 and others 18.

Step 8. The smoking gun

Let us go back to the original data, where everyone chose. Let us compare the motivation of the two groups with the same balance test.

```
equation bal09.ls motivation c tutoring
show bal09
```

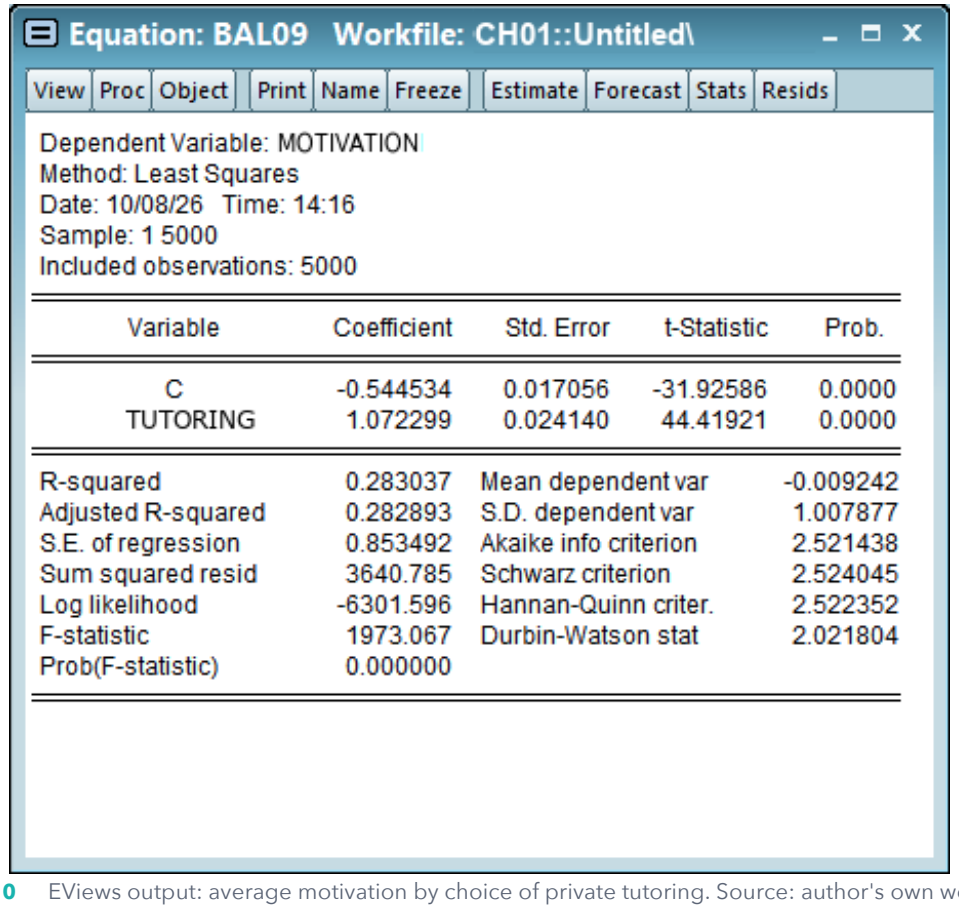


FIGURE 1.10 EViews output: average motivation by choice of private tutoring. Source: author's own work with EViews 14 and simulated data (ch01.prg).

There is the selection bias, in plain sight. Students who did not take tutoring had an average motivation of −0.54 (the constant). Those who did had 1.07 more: about 0.53. The t statistic is 44.42, and the p-value is 0.0000.

Since each unit of motivation is worth 2 grade points, that difference alone explains about 2.14 of the 3.15 points we saw in the first attempt. Do the math: 1 point of true effect plus 2.14 of bias. The equation of the trap, in numbers.

Step 9. Controlling is not magic: the noisy proxy

The third attempt worked because we had each student's exact motivation. In your research you will not. No survey asks "how motivated are you, on a perfect scale?" What surveys usually have is something similar, such as the hours of study the student reports. That stand-in variable is called a **proxy**.

Is it enough to control for the proxy? Let us try. We will create study hours that depend on motivation, but with a lot of noise, as with any survey answer.

```
series hours = 6 + 2*motivation + 3*nrnd
hours = @recode(hours<0, 0, hours)
group g10 motivation hours
show g10.cor
```

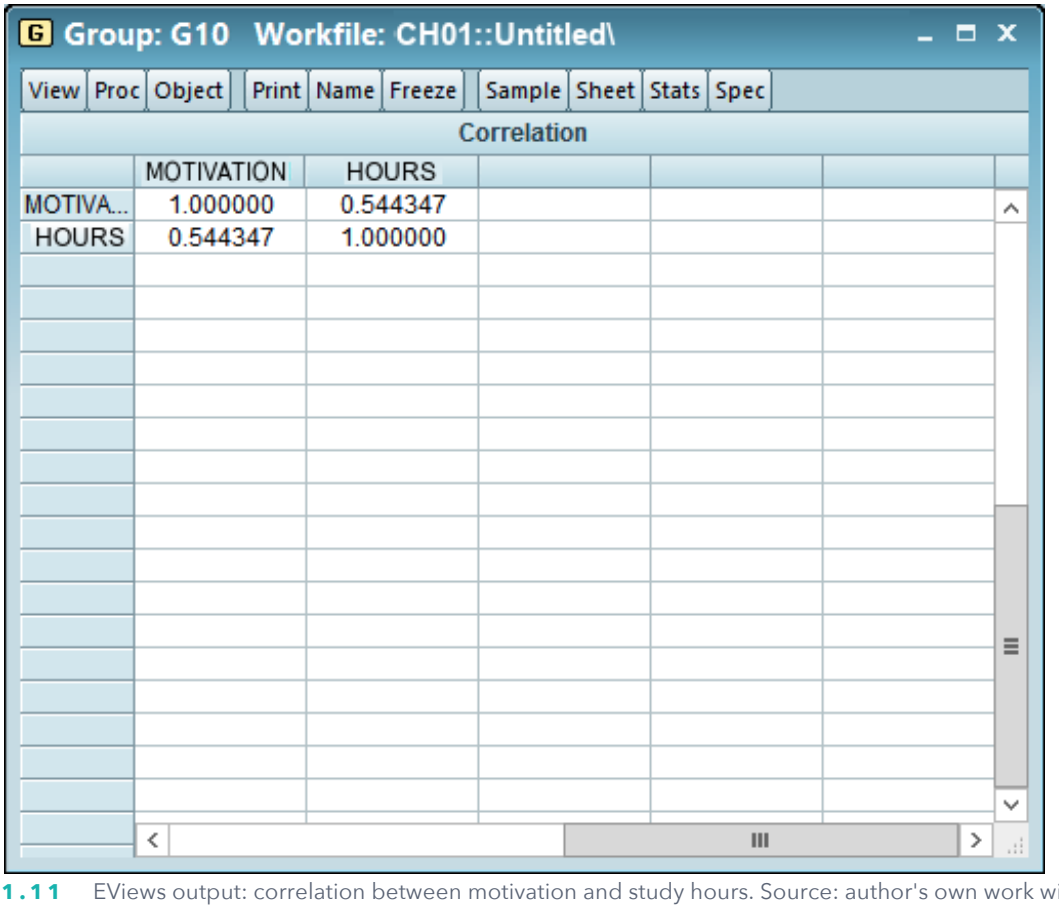


FIGURE 1.11 EViews output: correlation between motivation and study hours. Source: author's own work with EViews 14 and simulated data (ch01.prg).

Study hours resemble motivation: the correlation is 0.54. Will that be enough?

```
equation eq11.ls grade c tutoring hours
show eq11
```

Equation: EQ11 Workfile: CH01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GRADE
Method: Least Squares
Date: 10/08/26 Time: 14:17
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.594337	0.062735	136.9932	0.0000
TUTORING	2.637508	0.062838	41.97306	0.0000
HOURS	0.256301	0.009095	28.18054	0.0000

R-squared	0.414533	Mean dependent var	11.46764
Adjusted R-squared	0.414299	S.D. dependent var	2.778479
S.E. of regression	2.126402	Akaike info criterion	4.347339
Sum squared resid	22594.36	Schwarz criterion	4.351250
Log likelihood	-10865.35	Hannan-Quinn criter.	4.348710
F-statistic	1769.033	Durbin-Watson stat	1.967668
Prob(F-statistic)	0.000000		

FIGURE 1.12 EViews output: regression of the grade on tutoring, with study hours as a proxy for motivation. Source: author's own work with EViews 14 and simulated data (ch01.prg).

It is not. The coefficient on tutoring only fell from 3.15 to 2.64. We are still far from the true 1, and the table still looks flawless, with its p-value of 0.0000.

The lesson is uncomfortable, but it is worth gold: **controlling for an imperfect measure of the confounder corrects only part of the bias**. The worse the proxy measures the confounder, the more bias remains. That is why a good econometrician is never satisfied with "I controlled for everything I had." They ask how well each control measures what really worries them. When the answer is "sort of," they look for a different strategy. Those strategies fill the last part of the book.

Let us put all the attempts side by side.

TABLE 1.1 Five ways to estimate the effect of private tutoring (true effect: 1 point)

Method	Estimated effect	Does it get it right?
Compare means	3.15	No: it triples the effect
Simple regression	3.15	No: same thing with more decimals
Regression with the confounder	0.97	Yes, if you measure the confounder
Regression with a proxy for the confounder	2.64	No: it corrects only part
Experiment (lottery)	0.87	Yes, within its margin of error, without measuring anything else

Source: author's own work with EViews 14 and simulated data (ch01.prg); figures from Figures 1.4, 1.5, 1.7, 1.12 and 1.9.

1.7 The mistake almost every thesis student makes

Open any university's thesis repository and search for titles that begin with "The influence of…" or "The impact of…". You will find plenty. Many rely on a survey applied once, to a single group, with no lottery and no strategy to deal with selection bias. And they conclude that X "significantly influences" Y because the p-value came out below 0.05.

The mistake is not in the calculation. It is in the word. "Significant" means the relationship is probably not a product of chance: it rules out the fourth explanation on our list. It says nothing about the other three. A coefficient can be significant and still be contaminated by reverse causation or by a confounder, as you just saw with the 3.15.

How do you avoid it? With two habits.

First, choose the verb according to your design. If your strategy does not remove selection bias, write "is associated with," "is related to," "we observe a difference of." Save "effect," "impact" and "influences" for when you have an experiment or one of the methods in chapters 12 to 14. Your committee will notice. And if they don't, whoever reads your thesis later will.

Second, before running any regression, write down your list of confounders. Ask yourself: what could move both my explanatory variable and my outcome at the same time? If it is in your data, control for it. If it is not, say so in your limitations and explain in which direction it would bias your result. That single page does more for your research than ten normality tests.

And its twin: controlling for everything

There is an opposite mistake, just as common. The thesis student who has learned to control puts every variable in the workfile into the regression, "just in case." It sounds prudent, but it is not.

Some variables should never be controlled for, because they are consequences of the treatment you are studying. Suppose you want to measure how much a loan raises a farmer's income, and you control for how many hectares the farmer planted that season. It seems reasonable, until you ask what the loan was for: largely, to plant more.

By comparing only farmers who planted the same area, you erase precisely one of the paths through which the loan raises income, and you also compare groups that are no longer comparable. Angrist and Pischke call these variables **bad controls** (Angrist & Pischke, 2009, pp. 64-68).

The practical rule fits in one line: **control for what happened before the treatment, not for what the treatment caused**. Go back to your arrow diagram. If a variable receives an arrow from your treatment, leave it out.

⚠ THE MISTAKE RESEARCHERS MAKE WHEN WRITING ARTICLES

The typical mistake shows up in the abstract: "The results show that education has a positive impact on income." Three pages later, in the limitations, the author admits the data are observational. The reviewer reads the abstract first and then looks for the design, and when it is not there, the paper comes back with the word "causality" underlined.

First, the verbs change halfway through. The methods section says "association," the results say "effect" and the conclusions recommend a policy as if the effect had been proven. An article must use the same verb from the title to the last line.

Second, a limitation is not a magic spell. Writing "causality cannot be claimed" at the end does not fix an abstract that claims it. If your design does not remove selection bias, say so in the introduction, where readers decide whether to keep reading.

How to write it well: "Students who took private tutoring score 3.15 points higher, but that difference falls to 0.97 when we compare students with the same motivation. Because the data are observational, we interpret the relationship as an association, and we discuss which confounders could explain it."

★ SUMMARY AND CONCLUSIONS

1. A correlation can come from causation, reverse causation, a confounder or chance. Software cannot tell the four apart.
2. A causal effect compares two potential outcomes of the same person, and one of them is never observed.
3. **Observed difference = causal effect + selection bias.** If the groups were not comparable, the difference misleads you.
4. A simple regression is a comparison of means. A small p-value does not correct the bias.
5. Controlling helps if you measure the confounder well; with an imperfect proxy, bias remains, and controlling for consequences of the treatment makes it worse.
6. A lottery removes selection bias. When you cannot randomize, econometrics gives you tools to get close, each with its own assumption.

In chapter 2 we leave simulation behind and work with real data: you will download the ENAHO and the World Bank's indicators, and have them ready in EViews in a single afternoon.

✓ SOLVED EXERCISES

All three are at the end of `ch01.prg`, under the heading `SOLVED EXERCISES`. Run them after steps 1 to 9 and in this order, because they use the workfile you already have open. Try to predict the result before running each one.

Exercise 1. What if motivation did not decide who takes tutoring?

Suppose students still choose, but now what decides is the distance between their home and the tutoring center, from 0 to 10 kilometers. The farther away they live, the less likely they are to pay for tutoring. Distance has nothing to do with grades or motivation. What will the simple comparison of means give now?

```
series distance = 10*rnd
series tutoring1 = @rbinom(1, @logit(2 - 0.4*distance))
series grade1 = 11 + 1*tutoring1 + 2*motivation + noise
grade1 = @recode(grade1<0, 0, @recode(grade1>20, 20, grade1))
show grade1.statby tutoring1
```

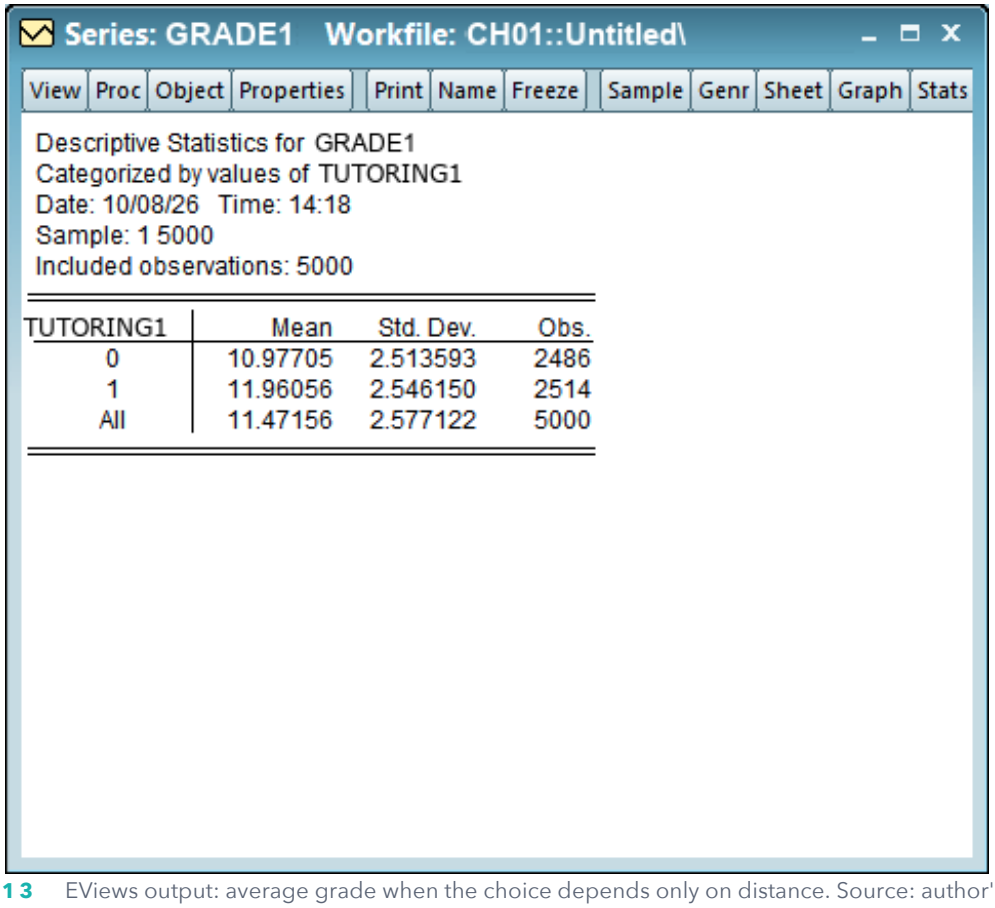


FIGURE 1.13 EViews output: average grade when the choice depends only on distance. Source: author's own work with EViews 14 and simulated data (ch01.prg).

```
show motivation.statby tutoring1
```

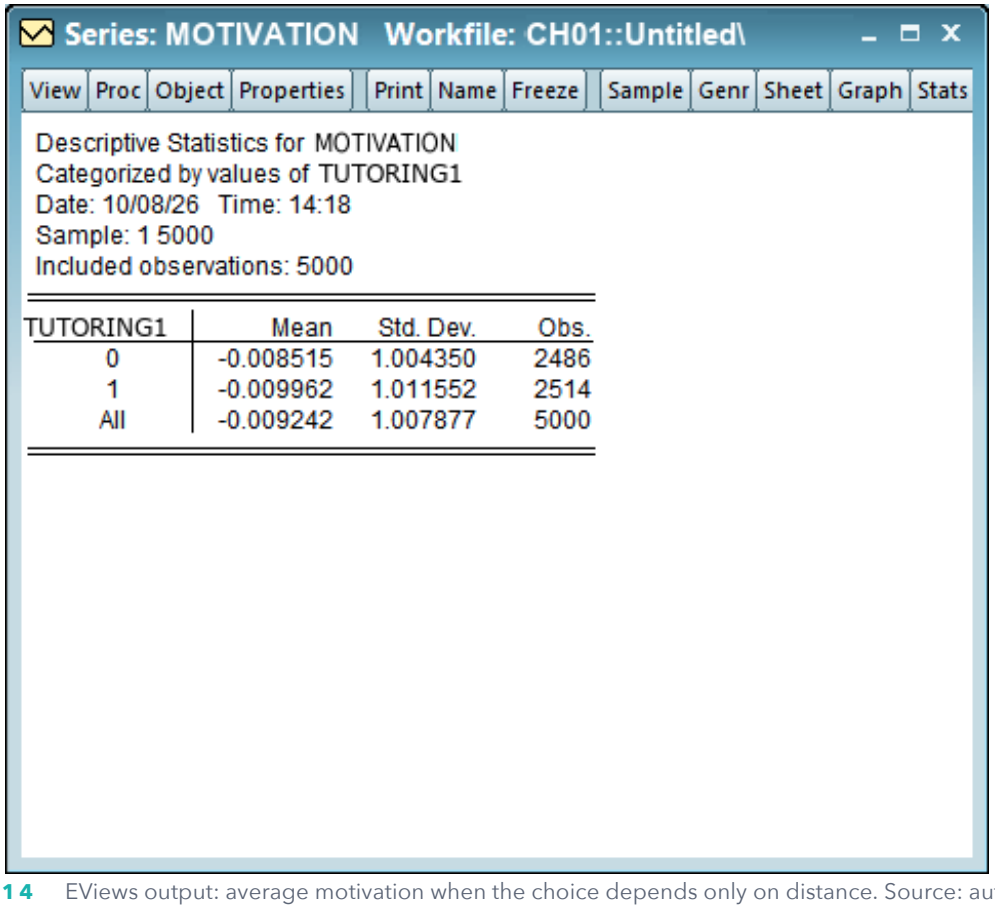


FIGURE 1.14 EViews output: average motivation when the choice depends only on distance. Source: author's own work with EViews 14 and simulated data (ch01.prg).

Solution. The `rnd` function draws a number between 0 and 1, so `10*rnd` spreads distance at random between 0 and 10 kilometers. Students with tutoring average 11.96 and those without, 10.98: a difference of 0.98, almost the true 1. And that is even though motivation still raises the grade by 2 points per unit. The second table explains why: average motivation is −0.010 in one group and −0.009 in the other, practically the same.

Nobody flipped a coin and, even so, the comparison gets it right. A confounder only biases your comparison if it moves **both things**: the treatment and the outcome. If whatever drives the choice does not touch the grade, it stops being a problem. Careful, though: in your research you would have to defend that assumption with arguments, because the data cannot prove it. Think about real distance: don't families in higher-income neighborhoods tend to live closer to tutoring centers? If so, distance would no longer be innocent.

Exercise 2. Put a number on selection bias.

With the original data, compute selection bias in two ways. First, as the observed difference minus the true effect. Second, as the difference in motivation between groups multiplied by its effect on the grade (2 points).

```
equation e1.ls grade c tutoring
scalar bias_a = e1.@coefs(2) - 1
equation e2.ls motivation c tutoring
scalar bias_b = 2*e2.@coefs(2)
table(3,2) tbias
tbias.setwidth(1) 24
tbias.setwidth(2) 12
tbias(1,1) = "How it is computed"
tbias(1,2) = "Bias"
tbias(2,1) = "Observed - true"
tbias(2,2) = bias_a
tbias(3,1) = "Motivation x 2"
tbias(3,2) = bias_b
show tbias
```

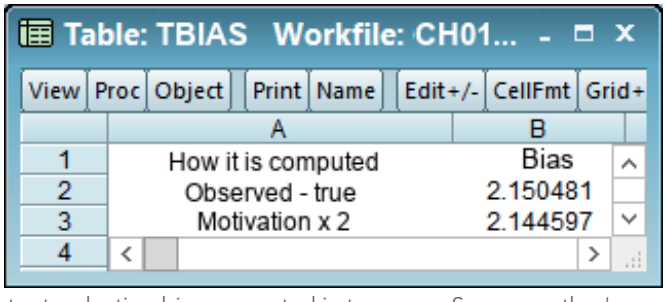


FIGURE 1.15 EViews output: selection bias computed in two ways. Source: author's own work with EViews 14 and simulated data (ch01.prg).

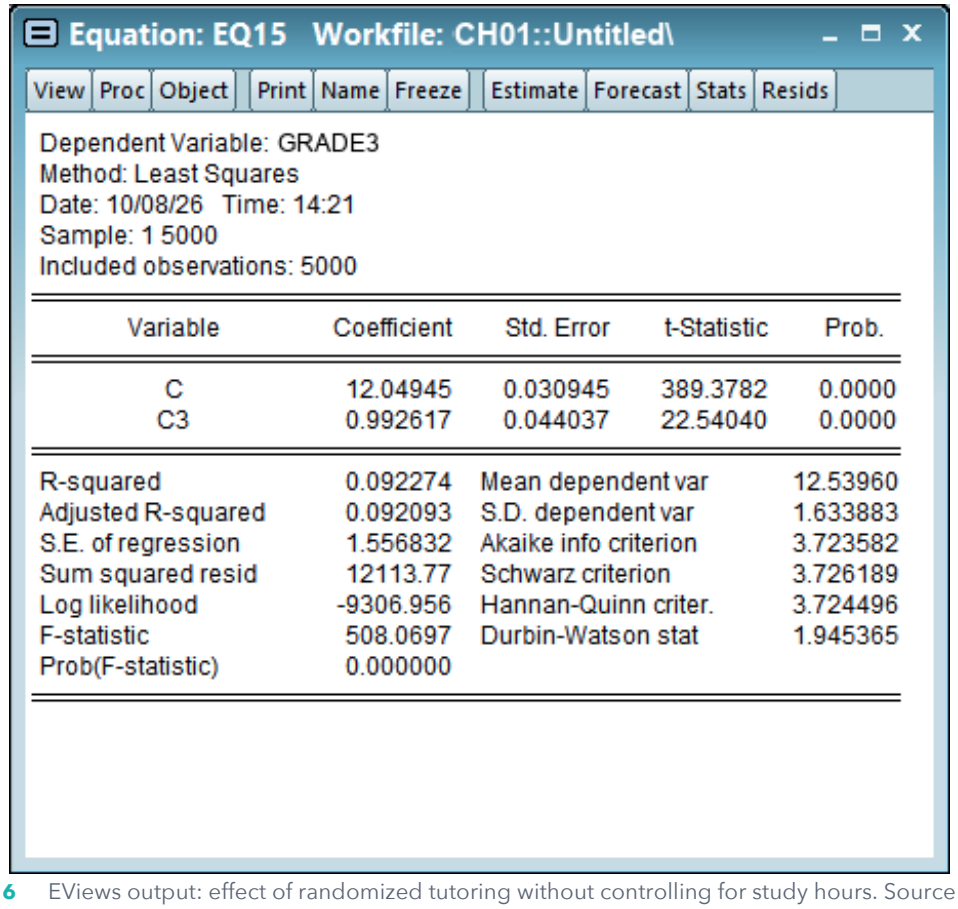
Solution. Three new tools appear here. An equation estimated with `equation` and no `show` is computed silently: it is saved in the workfile, but no window opens. `e1.@coefs(2)` is the second coefficient of that equation (the first is the constant). And `scalar` stores it in a named number. To see them together, we build a **table** with 3 rows and 2 columns. We set each column's width with `setwidth` and fill each cell with `tbias(row, column)`.

The two calculations nearly match: 2.150 and 2.145. It is the equation of the trap, checked with numbers: 3.15 observed = 1 of effect + about 2.15 of bias. The small gap between the two comes from the noise and from trimming the grade between 0 and 20. With your real data you can never do this calculation, because you do not know the true effect. That is why it pays to train your intuition here, where you do.

Exercise 3. A bad control in action.

Now tutoring is assigned by lottery, and it also makes students study 3 more hours per week. Each hour raises the grade by 0.2 points, and tutoring adds 0.4 points directly. What happens if you "control" for hours?

```
series c3 = @rbinom(1, 0.5)
series hours3 = 5 + 3*c3 + 2*nrnd
series grade3 = 11 + 0.4*c3 + 0.2*hours3 + 1.5*nrnd
grade3 = @recode(grade3<0, 0, @recode(grade3>20, 20, grade3))
equation eq15.ls grade3 c c3
show eq15
```



Equation: EQ15 Workfile: CH01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

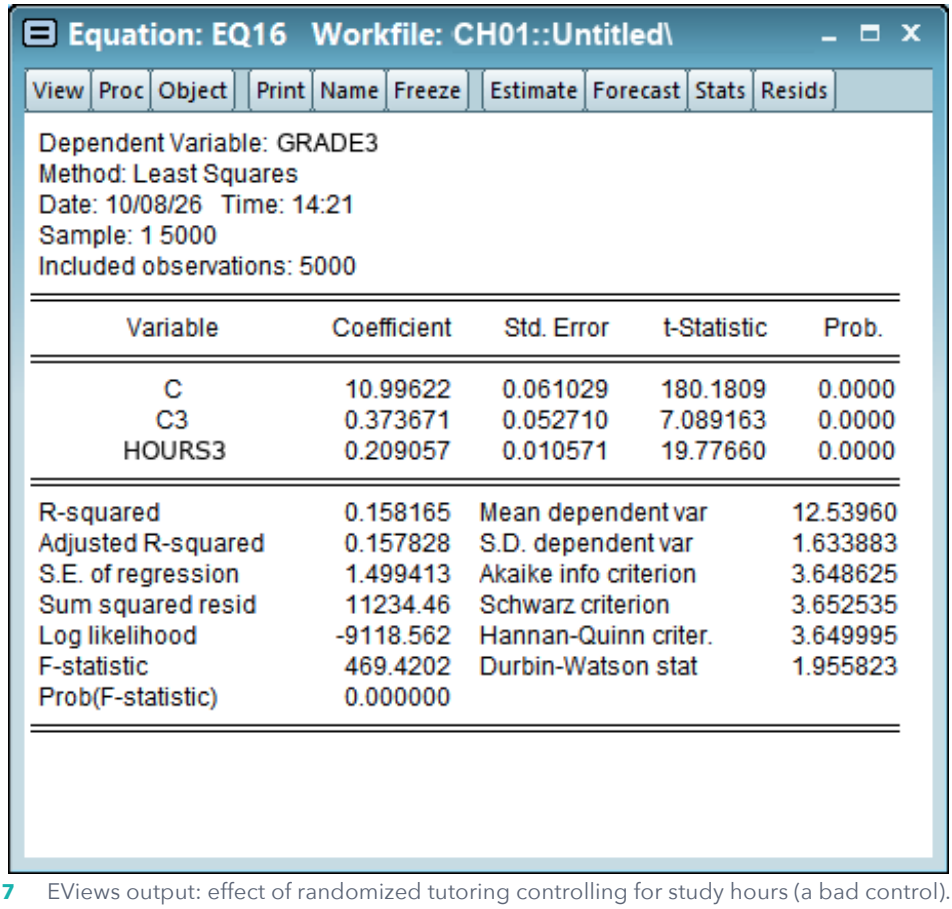
Dependent Variable: GRADE3
Method: Least Squares
Date: 10/08/26 Time: 14:21
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12.04945	0.030945	389.3782	0.0000
C3	0.992617	0.044037	22.54040	0.0000

R-squared	0.092274	Mean dependent var	12.53960
Adjusted R-squared	0.092093	S.D. dependent var	1.633883
S.E. of regression	1.556832	Akaike info criterion	3.723582
Sum squared resid	12113.77	Schwarz criterion	3.726189
Log likelihood	-9306.956	Hannan-Quinn criter.	3.724496
F-statistic	508.0697	Durbin-Watson stat	1.945365
Prob(F-statistic)	0.000000		

FIGURE 1.16 EViews output: effect of randomized tutoring without controlling for study hours. Source: author's own work with EViews 14 and simulated data (ch01.prg).

```
equation eq16.ls grade3 c c3 hours3
show eq16
```



Equation: EQ16 Workfile: CH01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GRADE3
Method: Least Squares
Date: 10/08/26 Time: 14:21
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.99622	0.061029	180.1809	0.0000
C3	0.373671	0.052710	7.089163	0.0000
HOURS3	0.209057	0.010571	19.77660	0.0000

R-squared	0.158165	Mean dependent var	12.53960
Adjusted R-squared	0.157828	S.D. dependent var	1.633883
S.E. of regression	1.499413	Akaike info criterion	3.648625
Sum squared resid	11234.46	Schwarz criterion	3.652535
Log likelihood	-9118.562	Hannan-Quinn criter.	3.649995
F-statistic	469.4202	Durbin-Watson stat	1.955823
Prob(F-statistic)	0.000000		

FIGURE 1.17 EViews output: effect of randomized tutoring controlling for study hours (a bad control). Source: author's own work with EViews 14 and simulated data (ch01.prg).

Solution. The total effect of tutoring is 1 point: 0.4 directly plus 0.6 that arrives because the student studies more (3 hours × 0.2). The regression without controls estimates it well, at 0.99, with a standard error of 0.04. Once you control for hours, the coefficient drops to 0.37. You took away from tutoring the very path through which it produces a large part of its effect. Hours received an arrow from the treatment, so they were a bad control.

✎ FOR YOUR THESIS OR ARTICLE

1. **Write your question as a "what would have happened if…?"** For example: "How much would this farmer have earned without the loan?" If you cannot phrase it that way, your research may be descriptive. That is fine too: just change the verbs.

2. **Make your list of three confounders.** For each one, note whether it is in your data and in which direction it would move your result if you left it out.

3. **Check the title of your thesis or article.** If it says "influence," "effect" or "impact," ask yourself which part of your design removes selection bias.

What your committee will ask: "How do you know this relationship is causal and not a correlation?" After this chapter you know how to answer: name the confounder that worries you most and explain what you did about it.

If it is an article: At a journal, the committee's question is asked by the reviewer, and it is asked earlier. If the title or the abstract says "effect" or "impact," the reviewer will look in the first paragraphs of the introduction for the design that removes selection bias. If your design does not remove it, change the verbs to "is associated with" starting with the title: it is cheaper than a rejection.

REFERENCES

Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *The New England Journal of Medicine, 330*(15), 1029-1035. https://doi.org/10.1056/NEJM199404143301501

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review, 92*(5), 1535-1558. https://doi.org/10.1257/000282802762024629

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691120355/mostly-harmless-econometrics

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691152844/mastering-metrics

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. https://mixtape.scunning.com/

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel laureates. *The New England Journal of Medicine, 367*(16), 1562-1564. https://doi.org/10.1056/NEJMon1211064