Mostly Harmless Econometrics

An Empiricist's Companion

Joshua D. Angrist
and
Jörn-Steffen Pischke

PRINCETON UNIVERSITY PRESS ■ PRINCETON AND OXFORD

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
6 Oxford Street, Woodstock, Oxfordshire OX20 1TW

Library of Congress Cataloging-in-Publication Data

Angrist, Joshua David.
Mostly harmless econometrics : an empiricist's companion /
Joshua D. Angrist, Jörn-Steffen Pischke.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-691-12034-8 (hardcover : alk. paper) —
ISBN 978-0-691-12035-5 (pbk. : alk. paper) 1. Econometrics.
2. Regression analysis. I. Pischke, Jörn-Steffen. II. Title.
HB139.A54 2008
330.01'5195—dc22 2008036265

British Library Cataloging-in-Publication Data is available

This book has been composed in Sabon
with Hel. Neue Cond. family display

Illustrations by Karen Norberg

Printed on acid-free paper. ∞

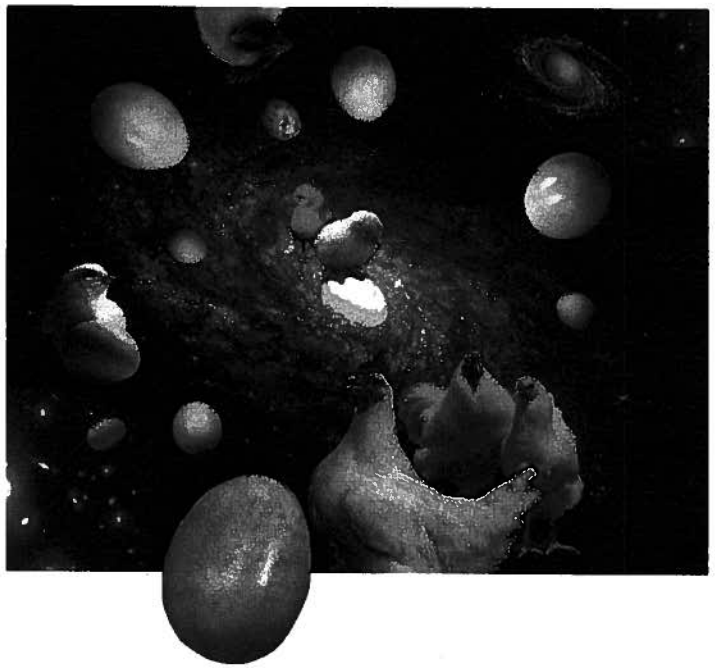
press.princeton.edu

Printed in the United States of America

1 3 5 7 9 10 8 6 4 2

CONTENTS

FIGURES



Chapter 1

Questions about *Questions*

"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

This chapter briefly discusses the basis for a successful research project. Like the biblical story of Exodus, a research agenda can be organized around four questions. We call these frequently asked questions (FAQs), because they should be. The FAQs ask about the relationship of interest, the ideal experiment, the identification strategy, and the mode of inference.

In the beginning, we should ask, *What is the causal relationship of interest?* Although purely descriptive research has an important role to play, we believe that the most interesting research in social science is about questions of cause and effect, such as the effect of class size on children's test scores, discussed in chapters 2 and 6. A causal relationship is useful for making predictions about the consequences of changing circumstances or policies; it tells us what would happen in alternative (or "counterfactual") worlds. For example, as part of a research agenda investigating human productive capacity—what labor economists call human capital—we have both investigated the causal effect of schooling on wages (Card, 1999, surveys research in this area). The causal effect of schooling on wages is the increment to wages an individual would receive if he or she got more schooling. A range of studies suggest the causal effect of a college degree is about 40 percent higher wages on average, quite a payoff. The causal

effect of schooling on wages is useful for predicting the earnings consequences of, say, changing the costs of attending college, or strengthening compulsory attendance laws. This relation is also of theoretical interest since it can be derived from an economic model.

As labor economists, we're most likely to study causal effects in samples of workers, but the unit of observation in causal research need not be an individual human being. Causal questions can be asked about firms or, for that matter, countries. Take, for example, Acemoglu, Johnson, and Robinson's (2001) research on the effect of colonial institutions on economic growth. This study is concerned with whether countries that inherited more democratic institutions from their colonial rulers later enjoyed higher economic growth as a consequence. The answer to this question has implications for our understanding of history and for the consequences of contemporary development policy. Today, we might wonder whether newly forming democratic institutions are important for economic development in Iraq and Afghanistan. The case for democracy is far from clear-cut; at the moment, China is enjoying robust economic growth without the benefit of complete political freedom, while much of Latin America has democratized without a big growth payoff.

The second research FAQ is concerned with *the experiment that could ideally be used to capture the causal effect of interest*. In the case of schooling and wages, for example, we can imagine offering potential dropouts a reward for finishing school, and then studying the consequences. In fact, Angrist and Lavy (2008) have run just such an experiment. Although their study looked at short-term effects such as college enrollment, a longer-term follow-up might well look at wages. In the case of political institutions, we might like to go back in time and randomly assign different government structures in former colonies on their independence day (an experiment that is more likely to be made into a movie than to get funded by the National Science Foundation).

Ideal experiments are most often hypothetical. Still, hypothetical experiments are worth contemplating because they help us pick fruitful research topics. We'll support this claim by

asking you to picture yourself as a researcher with no budget constraint and no Human Subjects Committee policing your inquiry for social correctness: something like a well-funded Stanley Milgram, the psychologist who did pathbreaking work on the response to authority in the 1960s using highly controversial experimental designs that would likely cost him his job today.

Seeking to understand the response to authority, Milgram (1963) showed he could convince experimental subjects to administer painful electric shocks to pitifully protesting victims (the shocks were fake and the victims were actors). This turned out to be controversial as well as clever: some psychologists claimed that the subjects who administered shocks were psychologically harmed by the experiment. Still, Milgram's study illustrates the point that there are many experiments we can think about, even if some are better left on the drawing board.[1] If you can't devise an experiment that answers your question in a world where anything goes, then the odds of generating useful results with a modest budget and nonexperimental survey data seem pretty slim. The description of an ideal experiment also helps you formulate causal questions precisely. The mechanics of an ideal experiment highlight the forces you'd like to manipulate and the factors you'd like to hold constant.

Research questions that cannot be answered by any experiment are FUQs: fundamentally unidentified questions. What exactly does a FUQ look like? At first blush, questions about the causal effect of race or gender seem good candidates because these things are hard to manipulate in isolation ("imagine your chromosomes were switched at birth"). On the other hand, the issue economists care most about in the realm of race and sex, labor market discrimination, turns on whether someone treats you differently because they *believe* you to be black or white, male or female. The notion of a counterfactual world where men are perceived as women or vice versa has a long history and does not require Douglas Adams-style outlandishness to entertain (Rosalind disguised

[1] Milgram was later played by the actor William Shatner in a TV special, an honor that no economist has yet received, though Angrist is still hopeful.

as Ganymede fools everyone in Shakespeare's *As You Like It*). The idea of changing race is similarly near-fetched: in *The Human Stain*, Philip Roth imagines the world of Coleman Silk, a black literature professor who passes as white in professional life. Labor economists imagine this sort of thing all the time. Sometimes we even construct such scenarios for the advancement of science, as in audit studies involving fake job applicants and résumés.[2]

A little imagination goes a long way when it comes to research design, but imagination cannot solve every problem. Suppose that we are interested in whether children do better in school by virtue of having started school a little older. Maybe the 7-year-old brain is better prepared for learning than the 6-year-old brain. This question has a policy angle coming from the fact that, in an effort to boost test scores, some school districts are now imposing older start ages (Deming and Dynarski, 2008). To assess the effects of delayed school entry on learning, we could randomly select some kids to start first grade at age 7, while others start at age 6, as is still typical. We are interested in whether those held back learn more in school, as evidenced by their elementary school test scores. To be concrete, let's look at test scores in first grade.

The problem with this question—the effects of start age on first grade test scores—is that the group that started school at age 7 is ... older. And older kids tend to do better on tests, a pure maturation effect. Now, it might seem we can fix this by holding age constant instead of grade. Suppose we wait to test those who started at age 6 until second grade and test those who started at age 7 in first grade, so that everybody is tested at age 7. But the first group has spent more time in school, a fact that raises achievement if school is worth anything. There is no way to disentangle the effect of start age on learning from maturation and time-in-school effects as long as kids are still in school. The problem here is that for students, start age

[2]A recent example is Bertrand and Mullainathan (2004), who compared employers' reponses to résumés with blacker-sounding and whiter-sounding first names, such as Lakisha and Emily (though Fryer and Levitt, 2004, note that names may carry information about socioeconomic status as well as race.)

equals current age minus time in school. This deterministic link disappears in a sample of adults, so we can investigate pure start-age effects on adult outcomes, such as earnings or highest grade completed (as in Black, Devereux, and Salvanes, 2008). But the effect of start age on elementary school test scores is impossible to interpret even in a randomized trial, and therefore, in a word, FUQed.

The third and fourth research FAQs are concerned with the nuts-and-bolts elements that produce a specific study. Question number 3 asks, *What is your identification strategy?* Angrist and Krueger (1999) used the term *identification strategy* to describe the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment. Returning to the schooling example, Angrist and Krueger (1991) used the interaction between compulsory attendance laws in American states and students' season of birth as a natural experiment to estimate the causal effects of finishing high school on wages (season of birth affects the degree to which high school students are constrained by laws allowing them to drop out after their 16th birthday). Chapters 3–6 are primarily concerned with conceptual frameworks for identification strategies.

Although a focus on credible identification strategies is emblematic of modern empirical work, the juxtaposition of ideal and natural experiments has a long history in econometrics. Here is our econometrics forefather, Trygve Haavelmo (1944, p. 14), appealing for more explicit discussion of both kinds of experimental designs:

> A design of experiments (a prescription of what the physicists call a "crucial experiment") is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although—unfortunately—most economists do not describe their design of experiments explicitly. If they did, they would see that the experiments they have in mind may be grouped into two different classes, namely, (1) experiments that *we should like to make* to see if certain real economic phenomena—when artificially isolated from "other influences"—would verify certain

> hypotheses, and (2) the stream of experiments that Nature is steadily turning out from her own enormous laboratory, and which we merely watch as passive observers. In both cases the aim of the theory is the same, to become master of the happenings of real life.

The fourth research FAQ borrows language from Rubin (1991): *What is your mode of statistical inference?* The answer to this question describes the population to be studied, the sample to be used, and the assumptions made when constructing standard errors. Sometimes inference is straightforward, as when you use census microdata samples to study the American population. Often inference is more complex, however, especially with data that are clustered or grouped. The last chapter covers practical problems that arise once you've answered question number 4. Although inference issues are rarely very exciting, and often quite technical, the ultimate success of even a well-conceived and conceptually exciting project turns on the details of statistical inference. This sometimes dispiriting fact inspired the following econometrics haiku, penned by Keisuke Hirano after completing his thesis:

T-stat looks too good
Try clustered standard errors—
Significance gone

As should be clear from the above discussion, the four research FAQs are part of a process of project development. The following chapters are concerned mostly with the econometric questions that come up after you've answered the research FAQs—in other words, issues that arise once your research agenda has been set. Before turning to the nuts and bolts of empirical work, however, we begin with a more detailed explanation of why randomized trials give us our benchmark.

Chapter 2

The Experimental Ideal

It is an important and popular fact that things are not always what they seem. For instance, on the planet Earth, man had always assumed that he was more intelligent than dolphins because he had achieved so much—the wheel, New York, wars and so on—while all the dolphins had ever done was muck about in the water having a good time. But conversely, the dolphins had always believed that they were far more intelligent than man—for precisely the same reasons. In fact there was only one species on the planet more intelligent than dolphins, and they spent a lot of their time in behavioral research laboratories running round inside wheels and conducting frighteningly elegant and subtle experiments on man. The fact that once again man completely misinterpreted this relationship was entirely according to these creatures' plans.

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

The most credible and influential research designs use random assignment. A case in point is the Perry preschool project, a 1962 randomized experiment designed to assess the effects of an early intervention program involving 123 black preschoolers in Ypsilanti, Michigan. The Perry treatment group was randomly assigned to an intensive intervention that included preschool education and home visits. It's hard to exaggerate the impact of the small but well-designed Perry experiment, which generated follow-up data through 1993 on the participants at age 27. Dozens of academic studies cite or use the Perry findings (see, e.g., Barnett, 1992). Most important, the Perry project provided the intellectual basis for the massive Head Start preschool program, begun in 1964,

which ultimately served (and continues to serve) millions of American children.[1]

2.1 The Selection Problem

We take a brief time-out for a more formal discussion of the role experiments play in uncovering causal effects. Suppose you are interested in a causal if-then question. To be concrete, let us consider a simple example: Do hospitals make people healthier? For our purposes, this question is allegorical, but it is surprisingly close to the sort of causal question health economists care about. To make this question more realistic, let's imagine we're studying a poor elderly population that uses hospital emergency rooms for primary care. Some of these patients are admitted to the hospital. This sort of care is expensive, crowds hospital facilities, and is, perhaps, not very effective (see, e.g., Grumbach, Keane, and Bindman, 1993). In fact, exposure to other sick patients by those who are themselves vulnerable might have a net negative impact on their health.

Since those admitted to the hospital get many valuable services, the answer to the hospital effectiveness question still seems likely to be yes. But will the data back this up? The natural approach for an empirically minded person is to compare the health status of those who have been to the hospital with the health of those who have not. The National Health Interview Survey (NHIS) contains the information needed to make this comparison. Specifically, it includes a question, "During the past 12 months, was the respondent a patient in a hospital overnight?" which we can use to identify recent hospital visitors. The NHIS also asks, "Would you say your health in general is excellent, very good, good, fair, poor?"

[1]The Perry data continue to get attention, particularly as policy interest has returned to early education. A recent reanalysis by Michael Anderson (2008) confirmed many of the findings from the original Perry study, though Anderson also shows that the overall positive effects of the Perry project are driven entirely by the impact on girls. The Perry intervention seems to have done nothing for boys.

The following table displays the mean health status (assigning a 1 to poor health and a 5 to excellent health) among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean Health Status	Std. Error
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

The difference in means is 0.72, a large and highly significant contrast in favor of the nonhospitalized, with a t-statistic of 58.9.

Taken at face value, this result suggests that going to the hospital makes people sicker. It's not impossible this is the right answer since hospitals are full of other sick people who might infect us and dangerous machines and chemicals that might hurt us. Still, it's easy to see why this comparison should not be taken at face value: people who go to the hospital are probably less healthy to begin with. Moreover, even after hospitalization people who have sought medical care are not as healthy, on average, as those who were never hospitalized in the first place, though they may well be better off than they otherwise would have been.

To describe this problem more precisely, we can think about hospital treatment as described by a binary random variable, $D_i = \{0, 1\}$. The outcome of interest, a measure of health status, is denoted by Y_i. The question is whether Y_i is *affected* by hospital care. To address this question, we assume we can imagine what might have happened to someone who went to the hospital if that person had not gone, and vice versa. Hence, for any individual there are two potential health variables:

$$\textit{Potential outcome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}.$$

In other words, Y_{0i} is the health status of an individual had he not gone to the hospital, irrespective of whether he actually went, while Y_{1i} is the individual's health status if he goes. We would like to know the difference between Y_{1i} and Y_{0i}, which can be said to be the causal effect of going to the hospital for

individual i. This is what we would measure if we could go back in time and change a person's treatment status.[2]

The observed outcome, Y_i, can be written in terms of potential outcomes as

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$
$$= Y_{0i} + (Y_{1i} - Y_{0i})D_i. \qquad (2.1.1)$$

This notation is useful because $Y_{1i} - Y_{0i}$ is the causal effect of hospitalization for an individual. In general, there is likely to be a distribution of both Y_{1i} and Y_{0i} in the population, so the treatment effect can be different for different people. But because we never see both potential outcomes for any one person, we must learn about the effects of hospitalization by comparing the average health of those who were and were not hospitalized.

A naive comparison of averages by hospitalization status tells us something about potential outcomes, though not necessarily what we want to know. The comparison of average health conditional on hospitalization status is formally linked to the average causal effect by the equation:

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{Observed difference in average health}} = \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Average treatment effect on the treated}}$$
$$+ \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection bias}}.$$

The term

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1]$$

is the *average causal effect of hospitalization on those who were hospitalized.* This term captures the averages difference between the health of the hospitalized, $E[Y_{1i}|D_i = 1]$, and what would have happened to them had they not been hospitalized,

[2]The potential outcomes idea is a fundamental building block in modern research on causal effects. Important references developing this idea are Rubin (1974, 1977) and Holland (1986), who refers to a causal framework involving potential outcomes as the Rubin causal model.

$E[Y_{0i}|D_i = 1]$. The observed difference in health status, however, adds to this causal effect a term called *selection bias.* This term is the difference in average Y_{0i} between those who were and those who were not hospitalized. Because the sick are more likely than the healthy to seek treatment, those who were hospitalized have worse values of Y_{0i}, making selection bias negative in this example. The selection bias may be so large (in absolute value) that it completely masks a positive treatment effect. The goal of most empirical economic research is to overcome selection bias, and therefore to say something about the causal effect of a variable like D_i.[3]

2.2 Random Assignment Solves the Selection Problem

Random assignment of D_i solves the selection problem because random assignment makes D_i independent of potential outcomes. To see this, note that

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$
$$= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1],$$

where the independence of Y_{0i} and D_i allows us to swap $E[Y_{0i}|D_i = 1]$ for $E[Y_{0i}|D_i = 0]$ in the second line. In fact, given random assignment, this simplifies further to

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1]$$
$$= E[Y_{1i} - Y_{0i}].$$

The effect of randomly assigned hospitalization on the hospitalized is the same as the effect of hospitalization on a randomly chosen patient. The main thing, however, is that random assignment of D_i eliminates selection bias. This does not mean that randomized trials are problem-free, but in principle they solve the most important problem that arises in empirical research.

[3]This section marks our first use of the conditional expectation operator (e.g., $E[Y_i|D_i = 1]$ and $E[Y_i|D_i = 0]$). We use this to denote the population (or infinitely large sample) average of one random variable with the value of another held fixed. A more formal and detailed definition appears in Chapter 3.

How relevant is our hospitalization allegory? Experiments often reveal things that are not what they seem on the basis of naive comparisons alone. A recent example from medicine is the evaluation of hormone replacement therapy (HRT). This is a medical intervention that was recommended for middle-aged women to reduce menopause symptoms. Evidence from the Nurses Health Study, a large and influential nonexperimental survey of nurses, showed better health among HRT users. In contrast, the results of a recently completed randomized trial showed few benefits of HRT. Worse, the randomized trial revealed serious side effects that were not apparent in the nonexperimental data (see, e.g., Women's Health Initiative [WHI], Hsia et al., 2006).

An iconic example from our own field of labor economics is the evaluation of government-subsidized training programs. These are programs that provide a combination of classroom instruction and on-the-job training for groups of disadvantaged workers such as the long-term unemployed, drug addicts, and ex-offenders. The idea is to increase employment and earnings. Paradoxically, studies based on nonexperimental comparisons of participants and nonparticipants often show that after training, the trainees earn less than plausible comparison groups (see, e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Lalonde 1995). Here, too, selection bias is a natural concern, since subsidized training programs are meant to serve men and women with low earnings potential. Not surprisingly, therefore, simple comparisons of program participants with nonparticipants often show lower earnings for the participants. In contrast, evidence from randomized evaluations of training programs generate mostly positive effects (see, e.g., Lalonde, 1986; Orr et al., 1996).

Randomized trials are not yet as common in social science as in medicine, but they are becoming more prevalent. One area where the importance of random assignment is growing rapidly is education research (Angrist, 2004). The 2002 Education Sciences Reform Act passed by the U.S. Congress mandates the use of rigorous experimental or quasi-experimental research designs for all federally funded education studies. We can therefore expect to see many more randomized trials in education research in the years to come. A pioneering randomized study from the field of education is the Tennessee STAR experiment, designed to estimate the effects of smaller classes in primary school.

Labor economists and others have a long tradition of trying to establish causal links between features of the classroom environment and children's learning, an area of investigation that we call "education production." This terminology reflects the fact that we think of features of the school environment as inputs that cost money, while the output that schools produce is student learning. A key question in research on education production is which inputs produce the most learning given their costs. One of the most expensive inputs is class size, since smaller classes can only be achieved by hiring more teachers. It is therefore important to know whether the expense of smaller classes has a payoff in terms of higher student achievement. The STAR experiment was meant to answer this question.

Many studies of education production using nonexperimental data suggest there is little or no link between class size and student learning. So perhaps school systems can save money by hiring fewer teachers, with no consequent reduction in achievement. The observed relation between class size and student achievement should not be taken at face value, however, since weaker students are often deliberately grouped into smaller classes. A randomized trial overcomes this problem by ensuring that we are comparing apples to apples, that is, that the students assigned to classes of different sizes are otherwise comparable. Results from the Tennessee STAR experiment point to a strong and lasting payoff to smaller classes (see Finn and Achilles, 1990, for the original study, and Krueger, 1999, for an econometric analysis of the STAR data).

The STAR experiment was unusually ambitious and influential, and therefore worth describing in some detail. It cost about $12 million and was implemented for a cohort of kindergartners in 1985–86. The study ran for four years, until the original cohort of kindergartners was in third grade, and involved about 11,600 children. The average class size in regular Tennessee classes in 1985–86 was about 22.3. The experiment assigned students to one of three treatments: small

classes with 13–17 children, regular classes with 22–25 children and a part-time teacher's aide (the usual arrangement), or regular classes with a full-time teacher's aide. Schools with at least three classes in each grade could choose to participate in the experiment.

The first question to ask about a randomized experiment is whether the randomization successfully balanced subjects' characteristics across the different treatment groups. To assess this, it's common to compare pretreatment outcomes or other covariates across groups. Unfortunately, the STAR data fail to include any pretreatment test scores, though it is possible to look at characteristics of children such as race and age. Table 2.2.1, reproduced from Krueger (1999), compares the means of these variables. The student characteristics in the table are a free lunch variable, student race, and student age. Free lunch status is a good measure of family income, since only poor children qualify for a free school lunch. Differences in these characteristics across the three class types are small, and none is significantly different from zero, as indicated by the p-values in the last column. This suggests the random assignment worked as intended.

Table 2.2.1 also presents information on average class size, the attrition rate, and test scores, measured here on a percentile scale. The attrition rate (proportion of students lost to follow-up) was lower in small kindergarten classrooms. This is potentially a problem, at least in principle.[4] Class sizes are significantly lower in the assigned-to-be-small classrooms, which means that the experiment succeeded in creating the desired variation. If many of the parents of children assigned to regular classes had successfully lobbied teachers and principals to get their children assigned to small classes, the gap in class size across groups would be much smaller.

Because randomization eliminates selection bias, the difference in outcomes across treatment groups captures the average causal effect of class size (relative to regular classes with a part-time aide). In practice, the difference in means between treatment and control groups can be obtained from a regression of test scores on dummies for each treatment group, a point we expand on below. Regression estimates of treatment-control differences for kindergartners, reported in table 2.2.2 (derived from Krueger, 1999, table V), show a small-class effect of about five percentile points (other rows in the table show coefficients on control variables in the regressions). The effect size is about $.2\sigma$, where σ is the standard deviation of the percentile score in kindergarten. The small-class effect is significantly different from zero, while the regular/aide effect is small and insignificant.

The STAR study, an exemplary randomized trial in the annals of social science, also highlights the logistical difficulty, long duration, and potentially high cost of randomized trials.

[4]Krueger (1999) devotes considerable attention to the attrition problem. Differences in attrition rates across groups may result in a sample of students in higher grades that is not randomly distributed across class types. The kindergarten results, which were unaffected by attrition, are therefore the most reliable.

TABLE 2.2.1
Comparison of treatment and control characteristics in the Tennessee STAR experiment

Variable	Class Size			P-value for equality across groups
	Small	Regular	Regular/Aide	
Free lunch	.47	.48	.50	.09
White/Asian	.68	.67	.66	.26
Age in 1985	5.44	5.43	5.42	.32
Attrition rate	.49	.52	.53	.02
Class size in kindergarten	15.10	22.40	22.80	.00
Percentile score in kindergarten	54.70	48.90	50.00	.00

Notes: Adapted from Krueger (1999), table I. The table shows means of variables by treatment status for the sample of students who entered STAR in kindergarten. The P-value in the last column is for the F-test of equality of variable means across all three groups. The free lunch variable is the fraction receiving a free lunch. The percentile score is the average percentile score on three Stanford Achievement Tests. The attrition rate is the proportion lost to follow-up before completing third grade.

TABLE 2.2.2
Experimental estimates of the effect of class size on test scores

Explanatory Variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian	—	—	8.35	8.44
			(1.35)	(1.36)
Girl	—	—	4.48	4.39
			(.63)	(.63)
Free lunch	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Teacher Master's degree	—	—	—	−0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Notes: Adapted from Krueger (1999), table V. The dependent variable is the Stanford Achievement Test percentile score. Robust standard errors allowing for correlated residuals within classes are shown in parentheses. The sample size is 5,681.

In many cases, such trials are impractical.[5] In other cases, we would like an answer sooner rather than later. Much of

[5]Randomized trials are never perfect, and STAR is no exception. Pupils who repeated or skipped a grade left the experiment. Students who entered an experimental school one grade later were added to the experiment and randomly assigned to one of the classes. One unfortunate aspect of the experiment is that students in the regular and regular/aide classes were reassigned after the kindergarten year, possibly because of protests by the parents with children in the regular classrooms. There was also some switching of children after the kindergarten year. But Krueger's (1999) analysis suggests that none of these implementation problems affected the main conclusions of the study.

the research we do, therefore, attempts to exploit cheaper and more readily available sources of variation. We hope to find natural or quasi-experiments that mimic a randomized trial by changing the variable of interest while other factors are kept balanced. Can we always find a convincing natural experiment? Of course not. Nevertheless, we take the position that a notional randomized trial is our benchmark. Not all researchers share this view, but many do. We heard it first from our teacher and thesis advisor, Orley Ashenfelter, a pioneering proponent of experiments and quasi-experimental research designs in social science. Here is Ashenfelter (1991) assessing the credibility of the observational studies linking schooling and income:

> How convincing is the evidence linking education and income? Here is my answer: Pretty convincing. If I had to bet on what an ideal experiment would indicate, I bet that it would show that better educated workers earn more.

The quasi-experimental study of class size by Angrist and Lavy (1999) illustrates the manner in which nonexperimental data can be analyzed in an experimental spirit. The Angrist and Lavy study relied on the fact that in Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in a fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions, such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."

The Angrist-Lavy study compared students in grades with enrollments above and below bureaucratic class size cutoffs to construct well-controlled estimates of the effects of a sharp change in class size without the benefit of a real experiment. As in the Tennessee STAR study, the Angrist and Lavy (1999) results pointed to a strong link between class size and achievement. This was in marked contrast to naive analyses, also reported by Angrist and Lavy, based on simple comparisons between those enrolled in larger and smaller classes. These comparisons showed students in smaller classes doing worse

on standardized tests. The hospital allegory of selection bias would therefore seem to apply to the class size question as well.[6]

2.3 Regression Analysis of Experiments

Regression is a useful tool for the study of causal questions, including the analysis of data from experiments. Suppose (for now) that the treatment effect is the same for everyone, say $Y_{1i} - Y_{0i} = \rho$, a constant. With constant treatment effects, we can rewrite (2.1.1) in the form

$$\begin{array}{ccccccc} Y_i = & \alpha & + & \rho & D_i & + & \eta_i, \\ & \| & & \| & & & \| \\ & E(Y_{0i}) & & (Y_{1i} - Y_{0i}) & & & Y_{0i} - E(Y_{0i}), \end{array} \qquad (2.3.1)$$

where η_i is the random part of Y_{0i}. Evaluating the conditional expectation of this equation with treatment status switched off and on gives

$$E[Y_i|D_i = 1] = \alpha + \rho + E[\eta_i|D_i = 1]$$
$$E[Y_i|D_i = 0] = \alpha + E[\eta_i|D_i = 0],$$

so that

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \underbrace{\rho}_{\text{Treatment effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{Selection bias}}.$$

Thus, selection bias amounts to correlation between the regression error term, η_i, and the regressor, D_i. Since

$$E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0] = E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0],$$

this correlation reflects the difference in (no-treatment) potential outcomes between those who get treated and those who don't. In the hospital allegory, those who were treated had poorer health outcomes in the no-treatment state, while in the Angrist and Lavy (1999) study, students in smaller classes tended to have intrinsically lower test scores.

[6]The Angrist-Lavy (1999) results turn up again in chapter 6, as an illustration of the quasi-experimental regression-discontinuity research design.

In the STAR experiment, where D_i is randomly assigned, the selection bias term disappears, and a regression of Y_i on D_i estimates the causal effect of interest, ρ. Table 2.2.2 shows different regression specifications, some of which include covariates other than the random assignment indicator, D_i. Covariates play two roles in regression analyses of experimental data. First, the STAR experimental design used conditional random assignment. In particular, assignment to classes of different sizes was random within schools but not across schools. Students attending schools of different types (say, urban versus rural) were a bit more or less likely to be assigned to a small class. The comparison in column 1 of table 2.2.2, which makes no adjustment for this, might therefore be contaminated by differences in achievement in schools of different types. To adjust for this, some of Krueger's regression models include school fixed effects, that is, a separate intercept for each school in the STAR data. In practice, the consequences of adjusting for school fixed effects is rather minor, but we wouldn't know this without taking a look. We have more to say about regression models with fixed effects in chapter 5.

The other controls in Krueger's table describe student characteristics such as race, age, and free lunch status. We saw before that these individual characteristics are balanced across class types, that is, they are not systematically related to the class size assignment of the student. If these controls, call them X_i, are uncorrelated with the treatment D_i, then they will not affect the estimate of ρ. In other words, estimates of ρ in the long regression,

$$Y_i = \alpha + \rho D_i + X_i'\gamma + \eta_i, \qquad (2.3.2)$$

will be close to estimates of ρ in the short regression, (2.3.1). This is a point we expand on in chapter 3.

Inclusion of the variables X_i, although not necessary in this case, may generate more precise estimates of the causal effect

of interest. Notice that the standard error of the estimated treatment effects in column 3 is smaller than the corresponding standard error in column 2. Although the control variables, X_i, are uncorrelated with D_i, they have substantial explanatory power for Y_i. Including these control variables therefore reduces the residual variance, which in turn lowers the standard error of the regression estimates. Similarly, the standard errors of the estimates of ρ are reduced by the inclusion of school fixed effects because these too explain an important part of the variance in student performance. The last column adds teacher characteristics. Because teachers were randomly assigned to classes, and teacher characteristics have little to do with student achievement in these data, both the estimated effect of small classes and its standard error are unchanged by the addition of teacher variables.

Regression plays an exceptionally important role in empirical economic research. As we've seen in this chapter, regression is well-suited to the analysis of experimental data. In some cases, regression can also be used to approximate experiments in the absence of random assignment. But before we get into the important question of when a regression is likely to have a causal interpretation, it is useful to review a number of fundamental regression facts and properties. These facts and properties are reliably true for any regression, regardless of the motivation for running it.

Part II

The Core